TABLE OF CONTENTS

Invited Talks and Biography

Paper Listings

Foreword by Dean of the School of Computer Sciences

Associate Professor
Dr. Rosni Abdullah

First of all, welcome to the Second International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'10). This event is co-organized by Laboratoire d'Informatique de Grenoble (LIG) from France and International Research Center for Multimedia, Information, Communication and Applications (MICA) from Vietnam. We would like to thank both of the organizations for helping us to realize this workshop. This year USM has the honour to host this important event especially in the domain of spoken language technologies.

The SLTU meeting is a technical workshop focusing on spoken language processing for under-resourced languages. This event is especially meaningful to us, as Malaysia is home to many languages, and a lot of them are rather rare. Even though Malay is the national language, communities in Malaysia are free to practise other languages. Even Malay exists in different flavour or dialects at different regions of Malaysia. The Chinese community in Malaysia on the other hand speak different Chinese languages such as Hokkien, Cantonese, Hakka, Teochew etc. Many of the languages have evolved and deviated from their origin. For example the Hokkien spoken in Penang has many vocabularies borrowed from Malay. At the east of Malaysia- Sabah is home to over 50 indigenous languages. These languages have very little language resource and not well studied. Most of them do not have a writing system. Some are at the brink of extinction. An example is the language spoken by Orang Kanaq, which is one of 19 indigenous groups in peninsular Malaysia. The Orang Kanaq language has only less than a hundred speakers. Therefore, the work in this workshop is very much relevant to us in Malaysia, and also aligns with the goal of the university which is to create a sustainable future.

In School of Computer Sciences, Universiti Sains Malaysia, there are 3 research clusters: namely the service, architecture and knowledge clusters. We have a language engineering group which is working in the area language technology such as automatic speech recognition, machine translation, speech synthesis and others. In particular, we are interested in the area of Malay automatic speech recognition and Malay speech synthesis as there is a big demand for these applications. At the national level, the development in this area will encourage the usage and development of Malay. Other areas of interest are in speech search technology, which is dubbed as the next frontier for search engine, where we make use of the power of automatic speech recognition and grid to allow user to query about multimedia files. We also hope to explore those local speech phenomena such as code switching and have a deeper study on dialects. We hope this workshop will open up more opportunities of collaboration in the area of speech technologies and other related areas. Last but not least, we wish everyone has a fruitful workshop and enjoyable stay in Malaysia.

Foreword by Local Workshop Chair of SLTU'10

Associate Professor
Dr. Chan Huah Yong

After the first SLTU workshop was held in Hanoi, Vietnam in 2008, we have the honor to organize the second Spoken Languages Technologies for Under-resourced Languages workshop with the help and support from Laboratoire d'Informatique de Grenoble (LIG), France and Multimedia, Information, Communication and Applications (MICA) Research Center, Vietnam.

It took us around a year to prepare for this event, starting from the proposal to the realizing of it. However it seems like only yesterday, as I recalled the time when Tan Tien Ping asked me to be the local chair of this event. I accepted the offer because I felt that SLTU has a very good reason d'être. Today a lot of minorities and aborigines are losing their native mother tongue. They do not have the financial and political resources to promote their languages in schools or via the TV stations. It is a shame to lose out this human heritage. One of the ways to support the existence of these languages is through the advancement of today's ICT. This kind of effort is very much needed as it enables the preservation of human heritage.

We have successfully gathered experts from different countries to share their knowledge and expertise on Natural Language Processing and Computational Linguistics especially in spoken languages. I would like to welcome you to our beautiful university, USM and certainly to Penang which boasts of its beautiful sandy beaches and rich cultural heritage embedded within the island. Penang was recently declared a UNESCO World Heritage Site. It is also known to be the 'food paradise" of Malaysia. Do discover Penang that has a rich multicultural history, and our country, Malaysia that is truly Asia.

I would like to express my sincere thanks and appreciation to the dedicated work and cooperation given by the local committee, as well as the technical and professional support from LIG, MICA, the scientific committee and the financial support from our sponsors. Without this joint effort, SLTU 2010 workshop would not have materialized today.

Thank you. “Terima kasih”.

Foreword by SLTU’10 Chair

Dr. Laurent Besacier

After the 1st workshop held in Vietnam in 2008, we are very pleased to welcome participants to this 2nd international workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU), in Penang, Malaysia.

For 2010, we managed to gather researchers working on ASR, synthesis and translation for more than 20 under-resourced languages (portability issues, multilingual spoken language processing, fast resources acquisition, etc.). We are delighted to find so much interest in this SLTU 2010 workshop.

Overall, 22 papers have been selected for oral communications. The authors and co-authors originated from 17 countries (Bangladesh, Burma, Cambodia, Ethiopia, France, Germany, Hungary, India, Korea, Malaysia, Pakistan, Russia, Singapore, Spain, South-Africa, Vietnam, USA). We observed an increase in the quality of the papers compared to 2008 and the scientific committee has compiled a two-day scientific programme that covers areas of automatic translation, speech recognition and synthesis, as well as collection of language resources.

In addition, three prestigious speakers have been invited to give keynotes: Haizou Li, Ruhi Sarikaya and Alex Waibel. We wish to give special thanks to them. We would also like to thank all the members of the scientific committee for their help in making this event a scientifically recognized workshop. Special thanks also to the Universiti Sains Malaysia which accepted the responsibility to host the workshop and to all the local organizers.

Finally, we would like to extend our thanks to ISCA, AFCP, CNRS and Grenoble-INP for their strong support (notably, the support of ISCA and AFCP which allowed us to give grants to fund the venue of five PhD students).

Thanks and have an excellent workshop!

SLTU’10 Organising Committees

Workshop Chair
Laurent Besacier (LIG)

Workshop Co-chair
Eric Castelli (MICA)

Local Workshop Chair
Chan Huah Yong (USM)

Scientific Committee
Tanja Schultz, CMU, USA
Ruhi Sarikaya, IBM, USA
Haizhou Li, A-star, Singapore
Lori Lamel, LIMSI, France
Jean-Paul Haton, LORIA, France
Pascal Nocera, LIA, France
Bali Ranaivo, Multimedia University, Malaysia
Alvin W. Yeo, UNIMAS, Malaysia
Ch'ng Eng Siong, NTU, Singapore
Tan Tien Ping, USM, Malaysia
Eric Castelli, MICA, Vietnam
Geneviève Caelen-Haumont, MICA, Vietnam
Vincent Berment, LIG, France
Laurent Besacier, LIG, France

Local Organising Committee
Dennis Wong Chin Phang, USM, Malaysia
Faten Damanhoori, USM, Malaysia
Norlia Mustaffa, USM, Malaysia
Norliza Hani Md. Ghazali, USM, Malaysia
Nour Azimah Zulkapli, USM, Malaysia
Sabrina Tiun Abdullah, USM, Malaysia
Siti Khaotijah Mohammad, USM, Malaysia
Tan Ewe Hoe, USM, Malaysia
Tan Tien Ping, USM, Malaysia
Marilyn Lim Chien Hui, USM, Malaysia
Ruslan Ahmad, USM, Malaysia
Badriyah Che May, USM, Malaysia
Ramli Yahaya, USM, Malaysia
Syed Mohamad Syed Sahil, USM, Malaysia

Invited Talk

Invited Talk 1:

BISTRA: Malay-English Bidirectional Speech Translation
Dr. Haizhou Li
Institute for Infocomm Research, Singapore

Abstract

In this talk, I will describe the development of a Malay-English bidirectional speech translation system in the Institute for Infocomm Research, Singapore, as part of the Asian Speech Translation Advanced Research Consortium. I will introduce the basic components and the linguistic resources, in particular, large vocabulary continuous speech recognition, speech synthesis, and machine translation concerning Malay language. I will also discuss the network-based system architecture that supports the real-time speech translation service.

Biography

Dr Haizhou Li is the Principal Scientist of Human Language Technology at the Institute for Infocomm Research in Singapore. His research interests include automatic speech recognition and machine translation. Dr Li taught in the University of Hong Kong and South China University of Technology (1988-1994). He was a Visiting Professor at CRIN/INRIA in France (1994-1995). He was a Research Manager in Apple-ISS Research Centre (1996-1998), a Research Director in Lernout & Hauspie Asia Pacific (1999-2001), and a Vice President in InfoTalk Corp. Ltd (2001-2003), responsible for Asian language products. In 2009, he was named one of the two Nokia Professors by Nokia Fundation.

Dr Li now serves as an Associate Editor of IEEE Transactions on Audio, Speech and Language Processing. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009-2013), an Executive Board Member of the Asian Federation of Natural Language Processing (AFNLP, 2006-2010).

Invited Talk

Invited Talk 2:

Towards Building Effective Language Translation Systems

Dr. Ruhi Sarikaya
IBM Watson USA

Abstract

Automatic Language Translation - widely known as Machine Translation (MT) - has been one of the long-standing elusive goals in natural language processing and artificial intelligence. With the effect of increasing globalization at the individual and enterprise level, and wide-spread use of social networking sites the necessity to exchange knowledge between people who do not share a common language put MT into the spotlight. Now, having access to vast amounts of translation data and powerful computers, we are closer than ever to achieving that goal. In this talk we focus on building usable machine translation systems. We will highlight the practical and fundamental challenges for building MT systems and present our solutions and approaches on both fronts. In particular, we first give an overview of MT research, then focus on parallel data construction for MT, language and MT modeling in continuous space. We also demonstrate working MT systems for various applications between English and several major languages.

Biography

Dr. Ruhi Sarikaya is a research staff member and team lead in the Human Language Technologies Group at IBM T.J. Watson Research Center. He received the B.S. degree from Bilkent University, Turkey in 1995, M.S. degree from Clemson University, SC in 1997 and the Ph.D. degree from Duke University, NC in 2001 all in electrical and computer engineering. He has published over 50 technical papers in refereed journal and conference proceedings and, is the lead inventor of eight patents in the area of speech and natural language processing. At IBM he has received several prestigious awards for his work including two Outstanding Technical Achievement Awards (2005 and

2008) and two Research Division Awards (2005 and 2007). Prior to joining IBM in 2001 he was a researcher at the Center for Spoken Language Research (CSLR) at the University of Colorado at Boulder for two years. He also spent the summer of 1999 at the Panasonic Speech Technology Laboratory, Santa Barbara, CA. He has served as the publicity chair of IEEE ASRU'05 and gave a tutorial on "Processing Morphologically Rich Languages" at Interspeech'07. Dr. Sarikaya is currently serving as associate editors of IEEE Transactions on Audio Speech and Language Processing and IEEE Signal Processing Letters. He also served as the lead guest editor of the special issue on "Processing Morphologically-Rich Languages" for IEEE Trans. on Audio Speech & Language Processing.
His past and present research interests span all aspects of speech and language processing including speech recognition, natural language processing, machine translation, machine learning, speech-to-speech translation, speaker identification/verification, digital signal processing and statistical modeling.

Dr. Sarikaya is a member of IEEE (senior member), ACL and ISCA.

Invited Talk

Invited Talk 3:

Speech Translators for Humanitarian Projects
Dr. Alexander Waibel
Universitat Karlsruhe, Germany

Abstract

This talk will describe Jibbigo and our speech translators designed and experimented in the context of humanitarian exercises in Thailand, Honduras and Indonesia.

Biography

Dr. Alexander Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the Karlsruhe Institute of Technology, Germany. He is the director of the International Center for Advanced Communication Technologies (interACT), a center between seven international research institutions worldwide. At Carnegie Mellon, he also serves as Associate Director of the Language Technologies Institute. Dr. Waibel was one of the founders of C-STAR, the international consortium for speech translation research and served as its chairman from 1998-2000. His team has developed the JANUS speech translation system, the first American and European Speech Translation system, and more recently the first real-time simultaneous speech translation system for lectures. His lab has also developed a number of multimodal systems including perceptual Meeting Rooms, Meeting recognizers, Meeting Browser and multimodal dialog systems for humanoid robots. He directed the CHIL program (FP-6 Integrated Project on multimodality) in Europe and the NSF-ITR project STR-DUST (the first domain independent speech translation project) in the US. He is part of the French-German project Quaero. In the areas of speech, speech translation, and multimodal interfaces Dr. Waibel holds several patents and has founded and co-founded several successful commercial ventures.

Dr. Waibel received the B.S. in Electrical Engineering from the Massachusetts Institute of Technology in 1979, and his M.S. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 1980 and 1986. His work on the Time Delay Neural Networks was awarded the IEEE best paper award in 1990. His contributions to multilingual and speech translation systems was awarded the "Alcatel SEL Research Prize for Technical Communication" in 1994, the "Allen Newell Award for Research Excellence" from CMU in 2002, and the Speech Communication Best Paper Award in 2002.

EXPLOITING MORPHOLOGY IN SPEECH TRANSLATION WITH PHRASE-BASED FINITE-STATE TRANSDUCERS

Alicia Pérez, M. Inés Torres[*]

Department of Electricity and Electronics
University of the Basque Country
manes.torres@ehu.es

Francisco Casacuberta

Instituto Tecnológico de Informática
Technical University of Valencia
fcn@iti.upv.es

ABSTRACT

This work implements a novel formulation for phrase-based translation models making use of morpheme-based translation units under a stochastic finite-state framework. This approach has an additional interest for speech translation tasks since it leads to the integration of the acoustic and translation models.

As a further contribution, this is the first paper addressing a Basque-to-Spanish speech translation task. For this purpose a morpheme based finite-state recognition system is combined with a finite-state transducer that translates phrases of morphemes in the source language into usual sequences of words in the target language.

The proposed models were assessed under a limited-domain application task. Good performances were obtained for the proposed phrase-based finite-state translation model using morphemes as translation units, and also notable improvements are obtained in decoding time.

***Index Terms*—** Speech Translation, Stochastic Finite-State Transducers, Morphology

1. INTRODUCTION

The use of morphological knowledge in machine translation (MT) is relatively recent and has been mainly sustained in tasks where morphologically rich languages were involved. In both transfer-based and example-based MT approaches morphological analysis has been used in the source language to extract lemmas and split words into their compounds so as to predict word-forms in the target language [1, 2]. In [3] it was *Moses* [4], the state-of-the art statistical MT system, that was used to train phrase-based models at morpheme level.

With respect to MT under finite-state framework, in [5] a text-to-text translation paradigm was proposed by combining a phrase-based model dealing with running words and finite-state models including morphological knowledge. Specifically, the finite-state machine consisted of a composition of a word-to-stem statistical analyser in source word, a stem-to-stem translation model from source to target language and a stem-to-word statistical generation module in target language all the constituents being implemented with ATT-tools. No other morphemes except stems were used.

The contribution of this work is twofold: first, the formulation of speech translation based on morphemes under the finite-state framework, and second, its application on Basque to Spanish speech translation. We take advantage of all the compounds of a word, and not only of lemmas. We promote the use of finite-state models due to their decoding speed.

Spanish and Basque languages entail many challenges for current machine translation systems. Due to the fact that both languages are official in the Basque Country, there is a real demand of several documents to be bilingual. In spite of the fact that both languages coexist in the same area, they differ enormously. To begin with, it is precise to note that they have different origin: while Spanish belongs to the set of Romance languages, Basque is a pre-Indoeuropean language. There are notable differences in both morphology and syntax. In contrast to Spanish, Basque is an extremely inflected language, with more than 17 declension cases that can be recursively combined. Inflection makes the size of the vocabulary (in terms of word-forms) grow. Hence, the number of occurrences of word n-grams within the data is much smaller than in the case of Spanish, and this leads to poor or even unreliable statistic estimates. By applying to morpheme based models we aim at tackling sparsity of data and consequently getting improved statistical distributions.

2. MORPHEME-BASED SPEECH TRANSLATION

The goal of statistical speech translation is to find the most likely translation, $\hat{\bar{t}}$, given the acoustic representation, X, of a speech signal from the source language:

$$\hat{\bar{t}} = \arg\max_{\bar{t}} P(\bar{t}|X) \tag{1}$$

The transcription of speech in the source language into a sequence of morphemes, $\bar{m}$, can be introduced as a hidden vari-

[*] This work has been partially supported by the University of the Basque Country under grants 9/UPV 00224.310-15900/2004 and GIU07/57, by the Spanish CICYT under grant TIN2005-08660-C04-03, and by the Spanish program Consolider-Ingenio 2010 under grant CSD2007-00018.

able.

$$\hat{\bar{t}} = \arg\max_{\bar{t}} \sum_{\bar{m}} P(\bar{t}, \bar{m}|X) \quad (2)$$

Applying the Bayes' decision rule:

$$\hat{\bar{t}} = \arg\max_{\bar{t}} \sum_{\bar{m}} \frac{P(\bar{t}, \bar{m})P(X|\bar{t}, \bar{m})}{P(X)} \quad (3)$$

Let us assume that the probability of an utterance does not depend on the transcription in other language. Hence, the denominator would be independent of the variable over which the optimisation is being done, and thus, the decoding would be carried out as follows:

$$\hat{\bar{t}} = \arg\max_{\bar{t}} \sum_{\bar{m}} P(\bar{t}, \bar{m})P(X|\bar{m}) \quad (4)$$

It is the contribution of two terms that drives the search problem: 1) the acoustic model, $P(X|\bar{m})$, connecting a text string in terms of morphemes to its acoustic utterance; 2) the joint translation model, $P(\bar{t}, \bar{m})$, connecting source and target languages. Joint probability translation models are good candidates to be approached by stochastic finite-state transducers (SFSTs).

Some effort has been recently made in order to efficiently take advantage of both acoustic and translation knowledge sources [6] by exploring different architectures. We have implemented the morpheme-based speech translation models under two different architectures described in [7]: a) *integrated architecture* implementing eq. (4) analogously as in an automatic speech recognition (ASR) system where the LM was replaced by a joint probability model. Thanks to the nature of the finite state models a tight integration is allowed, making a difference with respect to other kind of integration; b) *decoupled architecture* where two stages are involved, that is, first, an ASR system copes with transcription of the speech utterance, and later, a text-to-text translation system translates the given transcription.

Finally, there is an important issue to be noted, and it is the fact that this formulation for speech translation makes use of morphemes only in the source language, while using word-forms in the target language. The underlying motivation is simply that a speech translation from a morphologically rich language into another that does not present inflection in nouns is being taken into consideration. This is, in fact, our case when translating from Basque to Spanish.

2.1. Phrase-based stochastic finite-state transducers

An SFST is a finite-state machine that analyses strings in a source language and accordingly produces strings in a target language along with the joint probability of both strings to be translation each other (for a formal definition turn to [6]). The characteristics defining the SFST are the topology and the probability distributions over the transitions and the states. These distinctive features can be automatically learnt from bilingual samples by efficient algorithms such as GIATI (Grammar Inference and Alignments for Transducers Inference) [7], which is applied in this work. As it is well known, an outstanding aspect of the finite-state models is the fact that they count on efficient standard decoding algorithms [8]. Indeed, it is the speed of the decoding stage that makes these models so attractive for speech translation.

In this work we deal with SFSTs based on phrases of morphemes. Previously, in [9], in phrase-based SFSTs were presented based on word-forms (we will refer to this approach as PW-SFST). In such a models the transitions occur consuming a sequence of words. Here we propose the use of sequences of morphemes PM-SFST instead. As for what the standard baseline SFST is concerned (referred to as W-SFST), the difference lies on the fact that the transitions consume isolated word-forms instead of sequences of either words or morphemes. In all the cases, the transitions of SFSTs produce a sequence of zero or more words in the target language and have a probability associated.

2.2. Morphological analysis

In this work we deal with a morphologically rich language: Basque. In Basque there is no freely available linguistic tool that splits the words into proper morphemes. For this reason, morpheme-like units were obtained by means of Morfessor [10], a data-driven approach based on unsupervised learning of morphological word segmentation. For both ASR and SMT it is convenient to keep a low morpheme to word ratio, in order to get better language modelling, acoustic separability and word generation amongst others. Consequently, in a previous work [11], an approach based of decomposing the words into two morpheme-like units, a *root* and an *ending* was presented. By default, Morfessor decomposed the words using 3 types of morphemes: prefixes, stems and suffixes. To convert the decompositions into the desired root-ending form, all the suffixes at the end of the word were joined to form the ending, and the root was built joining all the remaining prefixes, stems and possible suffixes between stems. This procedures led to a vocabulary of 946 morphemes set of [11].

3. EXPERIMENTAL RESULTS

Basque is a minority but official language in the Basque Country (Spain). It counts on scarce linguistic resources and database, in addition, it is a highly inflected language. As a result, exploiting the morphology seems a good choice to improve the reliance on statistics.

The models were assessed under METEUS corpus, consisting of a text and speech of weather forecast reports picked from those published in the Internet. As shown in Table 1, the corpus is divided into a training set and a training-independent test set consisting of 500 sentences. Each sentence of the test was uttered by at least 3 speakers, resulting in a speech evaluation data of 1,800 utterances from 36 speakers. Note that the size of the Basque vocabulary is 38% bigger than the Spanish one due to its inflected nature.

		Basque	**Spanish**
Training (Text)	Pair of sentences	14,615	
	Different pairs	8,220	
	Running words	154,778	168,722
	Vocabulary	1,097	675
	Average length	10.6	11.5
Test (Speech)	Utterances	1,800	
	Length (hours)	3.5	3.0

Table 1. Main features of METEUS corpus.

The phrase-based SFST using morphemes proposed here, PM-SFST, was compared with the other two models, previously mentioned, namely PW-SFST and W-SFST. The three models were trained from the corpus described in Table 1 making use of the so-called GIATI algorithm [7]. Speech translation was carried out using both the integrated and decoupled architectures. Besides, in order to explore the influence on the translation model of errors derived from the recognition process, a verbatim translation was also carried out. In this case, the input of the text-to-text translation system is the transcription of the speech free from errors (as if the recognition process had been flawless).

3.1. Computational cost and performance

The memory required for a model to be allocated in memory along with the invested decoding time are two key parameters to bear in mind when it comes to evaluating a speech translation system. Table 2 shows the spatial cost (in terms of number of transitions and branching factor) of each of the three SFST models studied along with the relative decoding time consumed. Regarding the time units, they are relative to the baseline W-SFST model, that is, given that the test was translated in 1 time unit by W-SFST, the time units required by the PW-SFST and PM-SFST was picked up.

	Transitions	**BF**	**<Time>**
W-SFST	114,531	3.27	1.00
PW-SFST	121,265	3.25	0.76
PM-SFST	127,312	3.21	0.71

Table 2. Spatial cost, in terms of number of transitions and branching factor (BF), and the relative amount of time required by each model for text-input translation (dimensionless magnitude).

Doubtless, it is the performance, measured in terms of translation accuracy or error rate what counts for the evaluation of both speech and text translation. Translation results were assessed under the commonly used automatic evaluation metrics: bilingual evaluation under study (BLEU [12]) and word error rate (WER). Table 3 shows speech translation results with the three approaches mentioned above and the different architectures. The recognition WER for decoupled architecture was obtained trough previous ASR experiments reported in [11] with the same set of moprhemes. We would like to emphasize that speech translation with integrated architecture gives both the transcription and the translation of speech in the same decoding step, as a result, and thus, each model gives its own recognition-word-error-rate.

		Recognition	**Translation**	
		WER	**WER**	**BLEU**
Integrated	W-SFST	6.26	47.5	47.6
	PW-SFST	6.12	48.4	48.0
	PM-SFST	6.06	47.8	48.6
Decoupled	W-SFST	4.93	46.9	47.3
	PW-SFST	4.93	48.5	49.0
	PM-SFST	4.93	47.8	49.3
Verbatim	W-SFST	0	45.6	48.6
	PW-SFST	0	46.5	50.4
	PM-SFST	0	46.7	50.7

Table 3. Speech translation results provided by different translation models (W-SFST, PW-SFST, PM-SFST) under either integrated or decoupled architectures. The verbatim translation is also shown as a baseline.

3.2. Discussion

Both PM-SFST and PW-SFST models outperform the baseline W-SFST with 95% confidence under 1,000 bootstrap samples following the statistical significance test described in [13] with the BLEU evaluation measure. Nevertheless, the differences between PM-SFST and PW-SFST are marginal.

Comparing the two architectures considered, the translation results are similar. Furthermore, taking into account that the LM used for speech transcription in ASR with decoupled architecture and the SFST used to both recognize and translate speech counted on the same amount of data, one could expect that the parameters of the latter would not be as well considered, and accordingly, the performance of the integrated architecture would be worse for recognition purposes.

The differences in translation performance between speech translation with the decoupled architecture and the verbatim translation are small. There are two factors that have influence on this fact: on the one hand, the input of the speech translation was not very degraded; on the other hand, the transducer shows certain capacity to deal with input errors by mechanisms such as smoothing.

With respect to the size and time-efficiency of the models (summarized in Table 2), as it is obvious, the phrase-based models (both PM-SFST and PW-SFST) are bigger than W-SFST. Nevertheless, the branching factor is smaller, which indicates that the phrase-based models are more restrictive than the word-based in that, on average, they allow for a smaller number of transitions per state. Note that in the smoothed W-SFST all the strings have non-zero probability while in the phrase-based approaches only those strings built up in terms of the existing phrases have a non-zero probability. Regarding decoding time (in Table 2) there is a correlation with the branching factor. The higher the branching factor, the higher the required time, and thus, the PM-SFST model shows significant time reductions.

4. CONCLUDING REMARKS AND FUTURE WORK

For natural language processing applications when the language under study is morphologically rich, it might be useful to make use of morphology. By using morpheme-like units, statistics collected over a given database could be improved, and accordingly, the parameters describing statistical models. As far as speech translation is concerned, there is a further interest on the use of morphemes as lexical unit, and it is precisely that the way in which the morphemes were extracted kept a low morpheme to word ratio avoiding so acoustic confusion.

In this work we have dealt with Basque to Spanish speech translation. Morpheme-based speech translation has been proposed in terms of morphemes and within the finite-state framework. The models have been assessed under a limited-domain task giving as a result improvements in both translation accuracy and decoding time.

As far as future work is concerned, the generation of target words from morphemes given a source out of vocabulary word is still an open problem that might, as well, be explored from the statistical approach. That is, instead of doing analysing, as in our case, generation might be tackled.

5. REFERENCES

[1] G. Labaka, N. Stroppa, A. Way, and K. Sarasola, “Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation,” in *Proc. Machine Translation Summit XI*, 2007.

[2] E. Minkov, K. Toutanova, and H. Suzuki, “Generating complex morphology for machine translation,” in *Proc. 45st Annual Meeting of the Asocciation for Computational Linguistics*, 2007, pp. 128–135.

[3] S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi, “Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner,” in *Proc. Machine Translation Summit XI*, 2007, pp. 491–498.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., “Moses: Open source toolkit for statistical machine translation,” *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*, pp. 177–180, 2007.

[5] P. Karageorgakis, A. Potamianos, and I. Klasinas, “Towards incorporating language morphology into statistical machine translation systems,” in *Proc. Automatic Speech Recogn. and Underst. Workshop (ASRU)*, 2005.

[6] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, “Recent efforts in spoken language translation,” *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, 2008.

[7] F. Casacuberta and E. Vidal, “Learning finite-state models for machine translation,” *Machine Learning*, vol. 66, no. 1, pp. 69–91, 2007.

[8] M. Mohri, F. Pereira, and M. Riley, “AT&T FSM LibraryTM and Finite-State Machine Library,” 2003.

[9] A. Pérez, M. I. Torres, and F. Casacuberta, “Speech translation with phrase based stochastic finite-state transducers,” in *Proc. IEEE 32nd International Conference on Acoustics, Speech, and Signal Processing* 2007, vol. IV, pp. 113–116, IEEE.

[10] M. Creutz and K. Lagus, “Inducing the morphological lexicon of a natural language from unannotated text,” in *Proc. International and Interdisciplinary Conference on Aadaptive Knowledge Representation and Reasoning*, 2005.

[11] V. G. Guijarrubia, M. I. Torres, and R. Justo, “Morpheme-based automatic speech recognition of basque,” in *Proc. 4th Iberian Conference on Pattern Recognition and Image Analysis*, 2009, pp. 386–393, Springer-Verlag.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proc. 40th Annual Meeting of the Association Computational Linguistics*, 2002, pp. 311–318.

[13] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in ASR performance evaluation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1, pp. 409–412.

ON MORPH-BASED LVCSR IMPROVEMENTS

Balázs Tarján [1], *Péter Mihajlik* [1,2]

[1] Department of Telecommunication and Media Informatics,
Budapest University of Technology & Economics, Hungary
[2] THINKTech Research Center Nonprofit LLC, Hungary
tarjanb@tmit.bme.hu, mihajlik@tmit.bme.hu

ABSTRACT

Efficient large vocabulary continuous speech recognition of morphologically rich languages is a big challenge due to the rapid vocabulary growth. To improve the results various subword units - called as morphs - are applied as basic language elements. The improvements over the word baseline, however, are changing from negative to error rate halving across languages and tasks. In this paper we make an attempt to explore the source of this variability. Different LVCSR tasks of an agglutinative language are investigated in numerous experiments using full vocabularies. The improvement results are compared to pre-existing other language results, as well. Important correlations are found between the morph-based improvements and between the vocabulary growths and the corpus sizes.

Index Terms — speech recognition, rich morphology, morph, language modeling, LVCSR

1. INTRODUCTION

The most commonly used LVCSR (Large-Vocabulary Continuous Speech Recognition) systems apply words as basic lexical units. Word-based recognition of morphologically rich languages, however, can result in well-known problems [1]: very large vocabularies, high OOV (Out Of Vocabulary) rate, and inaccurate language model parameter estimation due to large number of distinct word forms. These phenomena are handled typically by changing the base units from words to sub-word lexical units called as morphs. In this way vocabulary size can be radically decreased and even OOV words can be recognized. Thus, recognition accuracies can be significantly improved over the word baseline [2-4]. The LVCSR improvement can be outstandingly high in the case of read speech in certain agglutinative languages such as Finnish and Estonian [2]. On the other hand, the reported improvements are much smaller in the case of other agglutinative languages like Turkish, Arabic and Hungarian [4-7]. Besides, improvement for spontaneous speech recognition is very seldom [4] or the results are even worse [8], as compared to the classical word-based approach. Thus, morph-based improvement seems to be not only language but speech genre dependent, as well.

So far, few efforts have been made in order to explain the high variability of improvement due to morph-based speech recognition. A major study compares statistical morphs based LVCSR results across four languages [8]. The difficulty in evaluating these results is that the speech recognition technique was not the same across languages. In [6] the conclusion for Arabic broadcast news recognition is that the improvement of morph-based approach can be eliminated if appropriately large word vocabulary is chosen. [9] also compares morph-based LVCSR to very large vocabulary word-based one but the significant improvements are preserved for Finnish and Estonian. [9] suggests that the relatively worse improvements of others are possibly due to the low order (n<4) of the applied morph n-gram models. All of these work apply empirical cutoffs on the word and morph vocabularies, which make the cross language and across task comparisons difficult. Our previous work was our first attempt to make clear evaluation of morph-based LVCSR across speech genres [10]. It concluded that the vocabulary size at a given training corpus size can be a good indicator for the morph-based improvement. However, some of the Hungarian improvement results were too optimistic due to an unattended cutoff in the word vocabulary and so they became outliers in the across-language comparison.

All in all, in the earlier publications there were always ad-hoc vocabulary cutoffs applied (in the word- or in the morph-based approach or in both cases); therefore the morph and word system comparisons were not entirely fair. In this study, all the results of various LVCSR tasks are measured strictly with full vocabularies. Not only the results of [10] are corrected but important conclusions are sharpened and new ones are found that can be useful for speech recognition of morphologically rich and/or under-resourced languages.

In this paper, the same speech recognition system and the same algorithms are used and optimized separately for three LVCSR tasks: for spontaneous (conversational) speech, for press conference speech, and for classical broadcast news speech. Improvements are measured with well- and less-resourced training text corpora. All experiments are performed in Hungarian – as one of the languages with high morphological complexity – so that cross-lingual effects do not bias the comparison. The conclusions are extended for other languages, as well.

2. TASKS

Since morph-based speech recognition results scatter heavily on a speech genre scale – from read to spontaneous conversational speech – our concept was to measure the improvements due to morph-based speech recognition on a spontaneity scale. Three points on this scale corresponding to three Hungarian language LVCSR tasks are examined.

2.1. Spontaneous speech – MALACH task (SP)

The Hungarian MALACH task was chosen as the spontaneous end of the scale since no other spontaneous Hungarian database was available for us. The MALACH corpus contains interviews with elderly people and is detailed in [4,11]. The recordings are made typically in normal home environment and their content is carefully transcribed. Only transcriptions are used to train the language models, 160K words in sum. The amount of test data is 4 hours, 19K words (matched data set in [11]).

2.2. Mostly planned speech – Press Conference task (PC)

The press conference audiovisual material of the Hungarian government is publicly available. What makes this LVCSR task attractive is that all the transcriptions of press conferences are open for the public for years – altogether 1.2 million words. However, the transcriptions are not always exact, disfluencies and noises are not marked and ungrammatical sentences are corrected. Questions from press people and answers are included in the data, only unintelligible recordings are removed. The amount of test data is 80 minutes, 9.4K words.

2.3. Planned speech – Broadcast News task (BN)

We used publicly available broadcast news audiovisual data of a Hungarian TV channel specialized for news. Unfortunately no transcriptions are available, but a relatively large amount of broadcast news text data is placed on the website of the channel (5.6 million words). The recordings consist of basically clean speech. 1 hour of speech corresponding to 7.7K words is used as test data in the experiments.

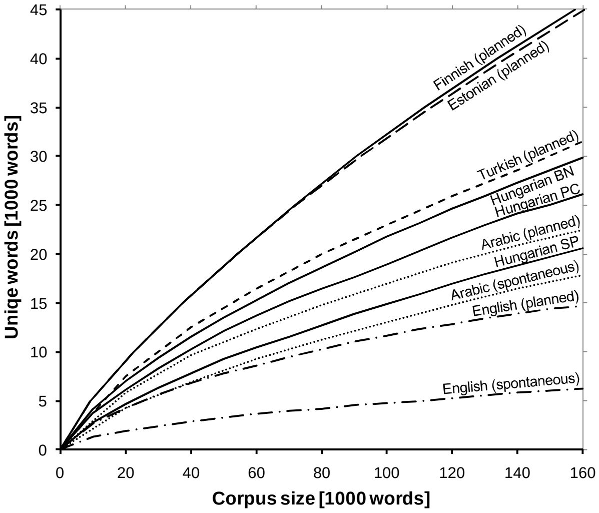


Figure 1: Number of unique words as a function of corpus size. Hungarian curves are calculated on the given databases. Non-Hungarian curves are reproduced from [8] with permission.

The morphological complexities of these tasks can be compared to each other and to other language tasks in Fig. 1.

In the followings word- and morph-based speech recognition approaches and the results of the three LVCSR tasks are presented and analyzed. The aim is to explore the dependencies of the improvement due to morph-based speech recognition.

3. METHODOLOGY

For building morph-based recognizers various vocabulary decomposition algorithms are applied. The differences between morph- and word-based speech recognition results are measured in several experimental setups. In each setup both word- and the morph-based systems are built and optimized on the given, task specific “in domain” training text database as follows.

3.1. Text corpus preparation

Whereas no extraordinary corpus preparation is required for word-based speech recognition, morph-based systems need special treatment of the given training text data. In our approach, first word boundary symbols <w> are placed into the text, after each word, and are considered as separate morphs. (<w> symbols are required for the reconstruction of word boundaries in the decoder output [12]). Then a core word list is collected leaving out all special tokens like acronyms, abbreviations, etc. Morph segmentation is performed on this core word list resulting in a “word-to-morph” dictionary. The corpus for a morph-based speech recognition system is obtained by replacing each word of the corpus by the corresponding morph sequence. The words not presented in the word-to-morph dictionary remain intact in the corpus and treated as simple morph tokens in the subsequent operations. The average numbers of morphs composing a word are given in Fig. 2.

3.2. Speech recognition models

3.2.1. Language model

In each setup, both word- and morph-based n-gram language models are built on the correspondingly processed task specific corpora with *full vocabularies* applying the modified, interpolated Kneser-Ney smoothing technique [13] implemented by the SRILM toolkit [14]. Depending on the task, on the type of morphs and on the training corpus size word and morph vocabulary sizes are in the range of 20k – 285k and 5k – 80k, respectively (see Fig. 2). By default, full 3-gram language models are built for the words and full 4-gram models for the morphs (ignoring 3 and 4 grams found only once). The only exception is at the 5.6M BN corpus, where entropy-based pruning [15] is applied both on the word and morph 3-grams resulting in roughly equal language models in terms of occupied operative memory size.

3.2.2. Pronunciation model

Simple grapheme-to-phoneme rules [16] and exceptions are applied on each lexicon separately in order to obtain word- and morph-to-phoneme mappings. Automatic phonetic transcription of both morphs and words can result in pronunciation errors especially at morph boundaries. However, according to our former experiences this kind of errors occur rarely [4]. Weighted alternative pronunciations are used only for the SP (MALACH) task, though as [4,17] showed, their effect is minimal on the recognition accuracy.

While there is a virtual "os = optional silence" model at the end of each word's pronunciation (with similar aim to the so-called "sp" model [18]), no such model is attached to the pronunciation of morph models. Instead, the <w> symbol itself is mapped to the "os" model.

3.2.3. Context dependency model

As equation (1) shows, triphone context expansion is performed after the integration of higher level knowledge sources, so that context dependency is modeled across word- and morph-boundaries, with respect to inter-word optional silences, as well.

3.2.4. Acoustic models

Speaker independent decision-tree state clustered cross-word triphone models were trained using ML (Maximum Likelihood) estimation [18]. Three state left-to-right HMM's were applied with GMM's (Gaussian Mixture Models) associated to the states. For the SP task, 26 hours of "in domain" training speech was used for training 3000 HMM states with 10 Gaussian per state, based on PLP (Perceptual Linear Prediction) features [17]. For both the PC and BN tasks the acoustic models were trained on the MRBA database [19] augmented with about 10 hours of transcribed PC speech. In that case the number of states was about 2500 and 8 Gaussians were used per state. The feature type was MFCC (Mel-frequency Cepstral Coefficients) with delta and delta-delta, calculated on 8 kHz bandwidth speech and blind channel equalization [20] was also applied.

3.3. Off-line recognition network construction

The WFST (Weighted Finite State Transducer) [21] recognition network is computed on the triphone-level:

$$\mathrm{wred}(\mathrm{fact}(\mathrm{compact}(C \text{ o } S \text{ o } \mathrm{compact}(\mathrm{det}(L \text{ o } G)')))) \qquad (1)$$

where capital letters stands for transducers, others for operators detailed below. First the language model (G) and pronunciation model (L) is composed and determinized. Then some auxiliary symbols are removed and a suboptimal minimization procedure – called as compaction – is applied that does not need the argument to be determinizable. Then each "os" model is replaced to a null-transition and to a normal silence model switched parallel by using a simple (S) transducer. The next step is the triphone context expansion (C), then the WFST network is compacted, factorized and the weights are redistributed resulting in a stochastic transducer suitable for the WFST decoder.

3.4. Evaluation

One-pass decoding was performed by the frame synchronous WFST decoder called as VOXerver – developed in our laboratories. RTF (Real Time Factor) of a morph- and the corresponding word-based system were adjusted to be close to equal using standard pruning techniques. RTF values were about 1 for small and midsized training text corpora for the PC and BN tasks, and about 4 for the largest corpora and for the SP task – measured on the same 3GHz, 1 core CPU.

The SP and the BN test sets contain only the speech of previously unseen speakers. All the PC and BN test speeches arose later in time than the related training text data.

Though in case of morphologically rich languages WER (Word Error Rate) – to some extent – shows a pessimistic picture of the speech recognition performance [8], we used it as the basis of evaluation since it is the most widely accepted and interpretable measure. Under the term of 'improvement' WER reduction is understood.

Signed-rank Wilcoxon tests with a significance level of 0.05 were applied to judge if a morph-based improvement is significant over the corresponding word-baseline.

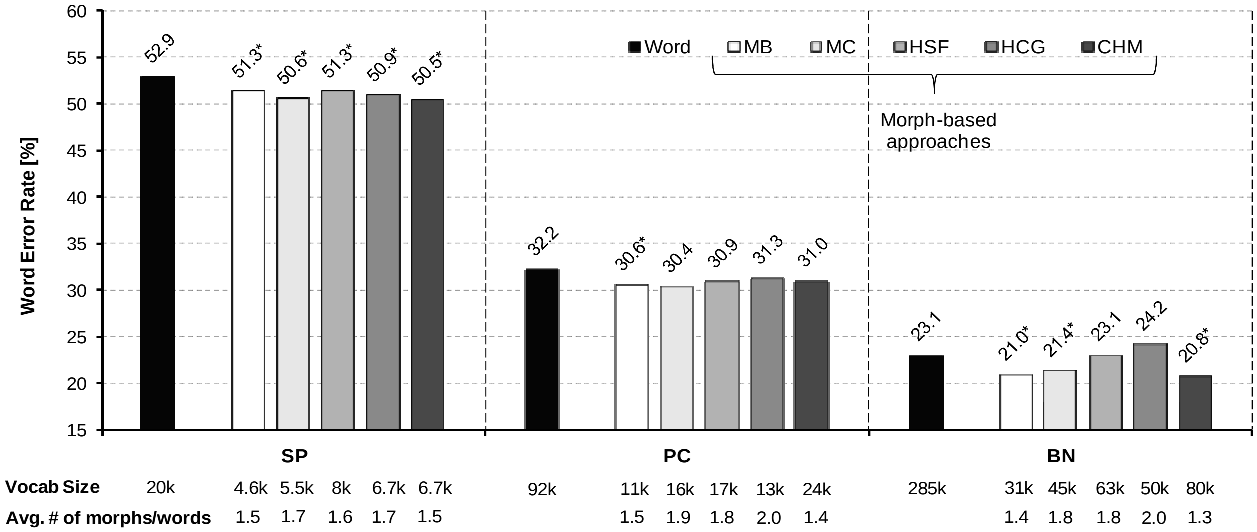


Figure 2: *Full-scale WER results with various morph types*
(* signs indicate significant improvement as compared to word baseline results)

4. RESULTS

First full-scale results are presented on the three Hungarian LVCSR tasks with various morph types. Then, the effect of lower resources – in terms of training text – is investigated. Finally, the influence of acoustic model quality is measured by applying adapted acoustic models.

4.1. Full-scale results of various morph types

The following morph types – in term of the applied vocabulary decomposition algorithm – are evaluated on all speech genres exploiting full training text corpora.

- ***Statistical morphs:*** selected words are segmented to morphs by using the unsupervised MB (Morfessor Baseline, [22]) and MC (Morfessor Categories-MAP, [23]) algorithms.
- ***Grammatical morphs:*** morphs were obtained by applying an affix-stripping method implemented in the Hunmorph system [24,25]. Two methods are used, a grammatically strict HSF (Hunmorph Strict Fallback) and a less strict, more heuristic HCG (Hunmorph Compound Guessing).
- ***Combined morphs:*** CHM (Combined Hunmorph Morfessor) the MB algorithm is used to disambiguate the multiple morph analyses of Hunmorph system [4,11]

More detailed description of these morph types can be found in [4], [11] and in [17].

As the results in Fig. 2 show, there are significant improvements due to the morph-based LVCSR approach in each task, although the improvements are definitely smaller than the Finnish or Estonian ones [8]. In general, the more planned is the speech the higher is the error rate reduction. Nevertheless, the morph modeling technique does matter, especially in the BN task, where grammatical methods fail to outperform the word baseline. In average, the best results are obtained with the CHM method, but only the MB technique achieves consistently significant word error rate reduction. Moreover, the Morfessor Baseline word-to-morph segmentation method provides the smallest vocabulary sizes, therefore only the MB morph modeling technique is investigated further.

Table 1. Speaker independent speech recognition results with various training text corpora sizes

Task	# of training words	# of word forms	OOV rate [%]	Word WER [%]	MB WER [%]
SP	160k	20k	15.6	52.9	51.3
PC	160k	26k	14.1	43.0	38.4
PC	1.2M	92k	6.3	32.2	30.6
BN	160k	30k	16.4	41.8	35.3
BN	1.2M	105k	7.2	26.4	23.5
BN	5.6M	285k	3.6	23.1	21.0

Table 2. Speech recognition results with acoustic model adaptation and with various training text corpora sizes

Task	# of training words	Word WER [%]	MB WER [%]
SP – adapt.	160k	47.6	43.8
PC – adapt.	160k	40.3	36.0
PC – adapt.	1.2M	30.7	29.1
BN – adapt.	160k	39.8	31.5
BN – adapt.	1.2M	24.3	21.4
BN – adapt.	5.6M	20.6	18.9

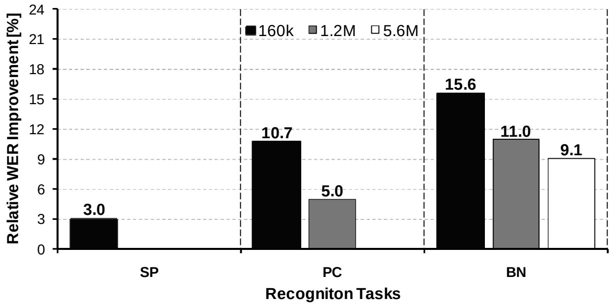


Figure 3: Relative improvement rates of speaker independent WER's measured on different training text corpora sizes

4.2. Effects of down-scaled training text corpora

In order to eliminate the effect of differently sized training texts, we made additional experiments based on equal-size training text corpora. The most recent texts are left in the reduced training databases of the PC and BN tasks.

As Table 1 shows – and as it is expected – the larger is the amount of training text data the lower is the recognition error rate. Looking at the improvement results (Fig. 3), a much less expected phenomenon can be observed. Namely, the reduction of training data dramatically increased the improvement rates. This may mean that morph-based speech recognition can be useful tool if the language resources are strongly limited.

By comparing the improvement results of the three tasks measured with equally sized training text it can be seen that there is no direct correspondence between the OOV rates and the improvements (see Table 1 and Fig. 3). The dependence of morph-based error reduction from the speech genre, however, is even more characteristic.

4.3. Effects of acoustic model adaptation

In this experimental setup we aim at modifying the acoustic model quality beside the language model. The previous experiment is repeated using unsupervised MLLR (Maximum Likelihood Linear Regression) acoustic model adaptation instead of applying speaker independent models. Speaker dependent acoustic model transformations were trained only for the SP task, i.e. only one transformation is used per task both in the PC and BN adaptation setups.

By comparing the results of Table 1 and 2, it can be seen that i) acoustic model adaptation was always effective in the reduction or recognition errors; ii) the relative improvements due to morph-based modeling can be significantly larger with more accurate acoustic models (Fig. 3 and 4).

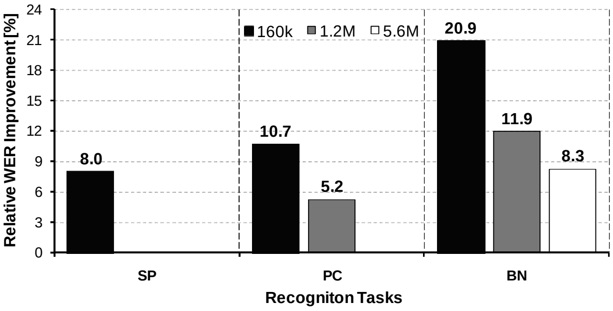


Figure 4: Relative improvement rates of WER's measured with acoustic model adaptation on different training text corpora sizes

5. DISCUSSION

The results suggest that the improvements due to morphs-based modeling correlate greatly with the speech genre. The language dependency assumption is augmented, too: the best improvement for Hungarian is only the half of the best Finnish one [8] achieved with similar techniques.

We suppose that the differences between the examined three speech genres and languages are manifested partially in the different vocabulary growths. Obviously, the number of word forms at a given corpus size is definitely different for the three Hungarian LVCSR tasks as well as for other agglutinative language tasks, see Fig. 1.

In Fig. 5, besides Hungarian, other language relative improvement results from [8] are shown in the function of number of different word forms at 160k words (sub)corpus sizes. All plotted approaches apply the MB algorithm –

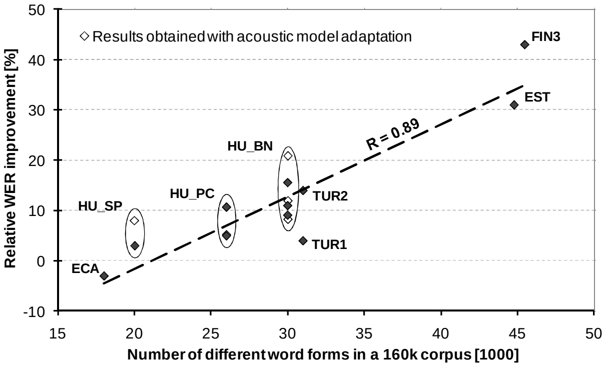


Figure 5: *Illustration of the relation between the relative WER reductions in various language LVCSR setups and the number of unique words at a given amount of training text. (Non-Hungarian results are from [8], ECA stands for spontaneous Arabic, other setups present more or less planned speech.)*

though in different ways – and use context dependent phone models. Although test data is not considered in this comparison we assume that test and training data are matched for good recognition accuracies.

The correlation between the number of different word forms and the relative improvements is 0.89 for the whole set, and 0.93 for the whole set but the results obtained with acoustic model adaptation.

6. CONCLUSIONS

Based on the morph-based improvement results of different Hungarian LVCSR tasks and on their comparison to each other and to other agglutinative language results it is possible to draw several conclusions. First, – independently from language and speech genre – the more rapid is the vocabulary growth of the given task the higher improvement can be expected from the application of morph-based speech recognition approach. Second, for vocabulary decomposition, a well-known and publicly available unsupervised statistical method (MB) seems to be a feasible first choice. Furthermore, the results suggest that neither the OOV rate nor the n-gram orders are crucial factor of the improvement, but acoustic model quality may do matter. Finally, an unforeseen conclusion is that morph-based speech recognition can be more beneficial in the case of less resourced tasks. Or, vice versa, using ample data and gigantic word vocabularies may eliminate the error rate reduction due to vocabulary decomposition. To verify and extend this assumption to other languages further researches and resources are needed even though similar phenomenon is observed in [6] evaluating Arabic broadcast news recognition results.

7. ACKNOWLEDGEMENTS

The work was founded partially by national grants: NKFP-2/034/2004, OM-00102-2007, KMOP-1.1.1-07/1-2008-0034, and by TAMOP-4.2.2/08/1/KMR.

8. REFERENCES

[1] K. Kirchoff and R. Sarikaya, “Processing Morphologically-Rich Languages” Tutorial at *INTERSPEECH 2007*, Antwerp, Belgium

[2] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimaki, J. Pylkkonen, T. Alumae and M. Saraclar, “Unlimited vocabulary speech recognition for agglutinative languages,” in *HLT-NAACL*, New York, USA, June 5-7, 2006.

[3] O.-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” *Speech Communication*, vol. 39, Issue 3-4, pp. 287-300, Feb. 2003.

[4] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh and T. Fegyó, “Improved recognition of spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task,” *IEEE Transactions on Audio, Speech, and Language Processing* – in press

[5] E. Arisoy, D. Can, S. Parlak, H. Sak and M. Saraclar, “Turkish Broadcast News Transcription and Retrieval,” *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):874-883, July 2009.

[6] G. Choueiter, D. Povey, S.F. Chen, and G. Zweig, ”Morpheme-based language modeling for Arabic LVCSR,” in *Proc. ICASSP’06*, Tolouse, France, 2006.

[7] M. Afify, R. Sarikaya, H.-K. J. Kuo, L. Besacier and Y. Gao, “On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition,” in *INTERSPEECH-2006*, paper 1444.

[8] M. Creutz et. al., “Morph-Based Speech Recognition and Modeling Out-of-Vocabulary Words Across Languages,” *ACM Transactions on Speech and Language Processing*, vol. 5, Issue 1, Article no. 3, December 2007.

[9] T. Hirsimäki, J. Pylkkönen and M. Kurimo, „Importance of High-Order N-gram Models in Morph-Based Speech Recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, 17(4): 724-732, May 2009.

[10] P. Mihajlik, B. Tarján, Z. Tüske and T. Fegyó, “Investigation of Morph-based Speech Recognition Improvements across Speech Genres,” in *Proc. Interspeech*, Brighton, United Kingdom, 2009, pp. 2687-2690.

[11] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske and V. Trón, “Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages,” in *TSD 2007*, Pilsen, Czech Republic, September 2007.

[12] T. Hirsimaki and M. Kurimo, “Decoder issues in unlimited Finnish speech recognition” In Proceedings of the Nordic Signal Processing Symposium *NORSIG 2004*, Espoo, Finland, 2004.

[13] S.F. Chen and J.T. Goodman, ”An Empirical Study of Smoothing Techniques for Language Modeling” Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[14] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *Proc. Intl. Conf. on Spoken Language Processing*, pp. 901–904, Denver, 2002.

[15] A. Stolcke, “Entropy-based pruning of backoff language models,” in *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, 270–274

[16] M. Szarvas, T. Fegyó, P. Mihajlik and P. Tatai, “Automatic Recognition of Hungarian: Theory and Practice,” *International Journal of Speech Technology*, 3:277-287, December 2000.

[17] P. Mihajlik, T. Fegyó, Z. Tüske and P. Ircing, “A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – Like Hungarian,” in *INTERSPEECH-2007*, pp. 1497-1500.

[18] S. Young, D. Ollason, V. Valtchev and P. Woodland, *The HTK book.* (for HTK version 3.2.), March 2002.

[19] MRBA – Hungarian Language Speech Database, http://alpha.tmit.bme.hu/speech/hdbMRBA.php

[20] L. Mauuary. “Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition,” in *Proc. of EUSPICO’98*, Vol.1, pp. 359-363, 1998.

[21] M. Mohri, F. Pereira and M. Riley, “Weighted Finite-State Transducers in Speech Recognition,” *Computer Speech and Language*, 16(1):69-88, 2002.

[22] M. Creutz and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.," in *Comp. and Inf. Sci.*, report A81, March 2005.

[23] M. Creutz and K. Lagus, "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text," in *Proc. of AKRR'05, Espoo*, Finland, 15-17 June, 2005.

[24] V. Trón, L. Németh, P. Halácsy, A. Kornai, Gy. Gyepesi and D. Varga, "Hunmorph: open source word analysis," in *Proc. ACL 2005 Software Workshop*, pp. 77-85.

[25] V. Trón, P. Halácsy, P. Rebrus, A. Rung, E. Simon and P. Vajda, "morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis," (in Hungarian) in *MSZNY Conf.*, Szeged, 2005.

POOLING ASR DATA FOR CLOSELY RELATED LANGUAGES

Charl van Heerden, Neil Kleynhans, Etienne Barnard and Marelie Davel

cvheerden@csir.co.za, ntkleynhans@csir.co.za, ebarnard@csir.co.za, mdavel@csir.co.za

ABSTRACT

We describe several experiments that were conducted to assess the viability of data pooling as a means to improve speech-recognition performance for under-resourced languages. Two groups of closely related languages from the Southern Bantu language family were studied, and our tests involved phoneme recognition on telephone speech using standard tied-triphone Hidden Markov Models. Approximately 6 to 11 hours of speech from around 170 speakers was available for training in each language. We find that useful improvements in recognition accuracy can be achieved when pooling data from languages that are highly similar, with two hours of data from a closely related language being approximately equivalent to one hour of data from the target language in the best case. However, the benefit decreases rapidly as languages become slightly more distant, and is also expected to decrease when larger corpora are available. Our results suggest that similarities in triphone frequencies are the most accurate predictor of the performance of language pooling in the conditions studied here.

Index Terms— speech recognition, under-resourced languages, data pooling

1. INTRODUCTION

When developing automatic speech recognition (ASR) systems for under-resourced languages, the amount of training data available is an important limiting factor. Although a wide variety of approaches to this issue have been studied (see Section 2 below), it is safe to say that data scarcity remains the most significant obstacle to the development of high-quality ASR systems.

Many of the existing approaches to dealing with data scarcity utilize similarities between some or all of the phonemes in different languages in order to improve the accuracy of ASR. Typically, it is assumed that sufficient data from a well-resourced language is available, and that data is employed in various configurations to improve the performance of ASR in an under-resourced language. In the current contribution, we investigate a somewhat different approach, namely data sharing among groups of related languages that are all lacking in resources. In particular, we wish to investigate how similar languages need to be for straightforward pooling of ASR training data to be beneficial.

If this approach proves to be useful, it can be used widely, since the vast majority of the actively spoken languages occur in clusters of more or less closely related families. However, there is good reason to suspect that only very closely related languages will benefit from pooling in this way – and even then, only if the amount of training data in the target language is severely limited. The main evidence for these concerns comes from experience with dialects of well-resourced languages: it is well known, for example, that ASR systems trained on American English perform poorly when presented with British English, and that combining training data from these two dialects generally leads to a deterioration in performance.

We therefore experiment with two groups of languages that are strongly related, as discussed in Section 3. In order to assess the effect of data pooling without any confounding influences, we only pool phones with identical IPA symbols in different languages – details are provided in Section 4, which also describes the recognizers employed . In Section 5 we analyse the pooled data according to a number of distance measures and report results for a phoneme-recognition task, showing that the relatedness of languages is indeed crucial to the success of this approach. Our conclusions are summarized in Section 6, which also suggests further work.

2. RELATED WORK

Once ASR systems were being developed for resource-scarce languages, research related to the possibility of supplementing target language data with that of additional languages soon followed. The rationale is simple: since the statistical methods being employed during acoustic modelling require more data than is available for the target language in question, borrow additional matching data from "donor languages" where possible.

In this section we review the main approaches to data combination that have been investigated in the literature. We describe strategies for data combination, different approaches to model mapping, and prior studies dealing specifically with the languages relevant to this paper.

2.1. Data combination strategies

Several different data combination strategies have been investigated, with results strongly influenced by the amount of target language data available, the acoustic diversity of the available databases, as well as the acoustic and phonotactic similarity between the target and donor language(s). Approaches include:

- *Cross-language transfer*: using an existing recognizer without any adaptation to the target language. Predictably, this strategy typically provides poor results, and is only considered if the languages in question are closely related [1], or if no target language data is available. In the latter case, multilingual acoustic models (built from a number of different languages and simultaneously able to recognize any of these) have been shown to yield better results than monolingual donor models [2, 3].

- *Cross-language adaptation*: adapting an existing mono- or multilingual recognizer using limited training data and techniques such as Maximum Likelihood Linear Regression (MLLR) or Maximum A Posteriori (MAP) adaptation [4, 5, 6]. These techniques can produce better results than cross-language transfer, and if target language data is very limited, can also outperform bootstrapping (see below).

- *Data pooling*: combining data from different sources by pooling the data directly. Such multilingual models were first developed in the context of language identification [7] but are also used in speech recognition, especially as initial models from which to adapt or bootstrap [3] or, to a lesser extent, when bilingual speakers are being recognized [8].

- *Bootstrapping*: initially demonstrated in [9], acoustic models are initialized using models from a donor language (or languages) and then rebuilt using target language data only. In [10], bootstrapping from multilingual models was shown to outperform adaptation when both approaches were evaluated using approximately 15 minutes of (Portuguese) target speech. While useful gains were obtained using bootstrapping, accuracy only approached that of a monolingual target language system (developed using 16.5h of target language data) once improved alignments of 90 minutes of target speech were used. (These improved alignments were assumed to be available, but typically are not.)

The methods described above can also be combined. For example, data pooling can be used to create multilingual models as seed models for bootstrapping [10], or a donor language can be adapted to a target language prior to data pooling [11].

Whichever method is used, cross-language data sharing has only been shown to compensate for limited target language data, and improvements soon dwindle as more target language data becomes available.

2.2. Approaches to model mapping

Before applying any of the data combination approaches described above, some mapping is usually required between the acoustic models of the donor languages and those of the target language. These mappings can be based on linguistic knowledge, data analysis or a combination of the two approaches.

Linguistic knowledge is typically encoded in the phoneme inventory of each of the languages, and the phoneme sets mapped directly based on IPA or SAMPA equivalences [8, 3], or other prior phonetic knowledge. Data-driven mappings are based on some distance (or similarity) measure, various of which have been utilized [12, 13, 14]. Good results are obtained using "hierarchical clustering", employing linguistically motivated categories within which data-driven (within-category) clustering is performed [13, 15]. Note that while most of these experiments were applied to context-independent models, similar techniques are applicable to context-dependent models, as well as to sub-phonemic models.

Hierarchical clustering at the sub-phoneme level can be integrated with the standard decision tree building process typically used to cluster and combine context-dependent triphones during Hidden Markov Model-based (HMM-based) model building: data samples are tagged with their source language and this additional information made available during data-driven clustering, resulting in improved results [10].

2.3. Data sharing of Southern Bantu languages

None of the above studies dealt specifically with data from any of the Southern Bantu languages. In the only cross-lingual adaptation study that includes these languages (that we are aware of), monolingual systems in isiXhosa and isiZulu were compared with a multilingual system developed using IPA-based data pooling of the two languages, with language-specific questions added during tying of triphones [16]. The multilingual system outperformed the monolingual systems, but gains were small. (Optimal phoneme accuracies for both approaches ranged between for 60.5% and 61.3%.)

3. CORPUS AND LANGUAGES

Our experiments are based on the Lwazi ASR corpus which was developed as part of a project that aims to demonstrate the use of speech technology in information service delivery in South Africa [17]. The corpus contains data from each of the eleven official languages of South Africa – approximately 200 speakers per language (2,200 speakers in total), contributed read and elicited speech, recorded over a telephone channel. Each speaker produced approximately 30 utterances; 16 of these were randomly selected from a phonetically balanced corpus and the remainder consist of short words and phrases.

For the purposes of the current research, we concentrate on two subsets of this corpus each containing a group of closely related languages. The three Sotho-Tswana languages (Sepedi, Setswana and Sesotho) form the first group, and three of the four Nguni languages (isiZulu, isiXhosa and isiNdebele) the second. (The fourth Nguni language in the Lwazi corpus, Siswati, was not included in the current study for reasons explained below.) These languages all belong to the Southern Bantu family of languages. Although they are used as first language by relatively large populations of speakers (all are considered as first language by several million speakers, with the exception of isiNdebele, which has only 700 000 first-language speakers), very few linguistic resources are available for these languages.

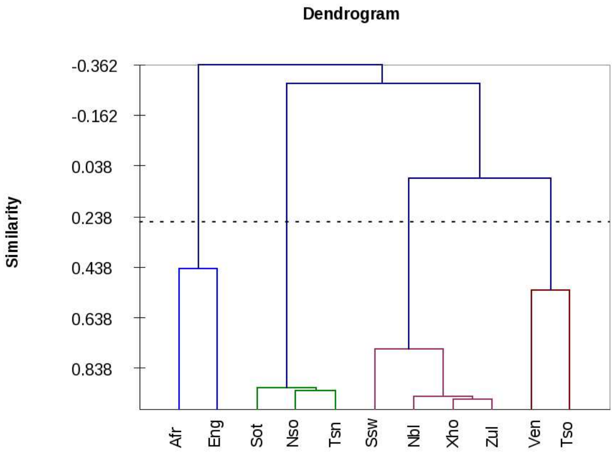


Fig. 1. Dendrogram calculated from confusion matrices for a multi-lingual text-based SVM classifier.

These languages all belong to the Southern Bantu family of languages [18]. We have previously studied their relationships using both orthographic and acoustic measures [19]. A typical dendrogram of the measured distances between the languages is shown in Fig. 1 (based on orthographic or text-based data); it can be seen that the two groups of languages selected here are indeed very closely related by these measures, and are therefore worthy candidates for the type of pooling considered here. Note also that Siswati is not as closely related to the other Nguni languages by these measures – it was therefore not included as a target language in the current study.

Language	# total min	# training min	# testing min
isiNdebele (Nbl)	609	517	92
isiZulu (Zul)	529	447	82
isiXhosa (Xho)	536	454	82
Sepedi (Nso)	548	465	83
Sesotho (Sot)	425	359	66
Setswana (Tsn)	443	379	64
Siswati (Ssw)	663	-	-

Table 1. Size of training and testing sets (in minutes) per language.

4. METHOD

4.1. ASR system overview

The ASR system developed to evaluate the effect of data pooling follows a standard Hidden Markov Model (HMM) design. Acoustic models consist of cross-word tied-state triphones modelled using a 3-state continuous density HMM. Each HMM state distribution is modelled by a 7-mixture multivariate Gaussian with diagonal covariance. The 39-dimensional feature vector consists of 12 static Mel-Frequency Cepstral Coefficients (MFCCs) with the 0'th cepstra, 13 delta and 13 delta-delta coefficients appended [20]. The final preprocessing step applies Cepstral Mean Normalization (CMN) which calculates a per utterance bias and removes it [21]. The different HMM state distributions were estimated by running multiple iterations of the Baum-Welch re-estimation algorithm. Once the triphone acoustic models were trained, a 40-class semi-tied transform [22] was estimated to further improve acoustic model robustness.

Our data pooling experiments are performed using the Lwazi ASR Corpus [17] and the Lwazi pronunciation dictionaries [23], as briefly described in Section 3. Table 1 indicates the amount of speech data in minutes for the different language-specific training and testing sets. Each language testing set was created by choosing 30 speakers at random, which were then excluded from the training data. In each case, we employed both the phonetically balanced sentences and the short phrases in our training and testing data.

4.2. Data combination approach

Our initial step in data pooling is to partition the languages into two groups: The *Nguni* group consists of isiZulu, isiNdebele and isiXhosa, while the *Sotho-Tswana* group includes Sepedi, Sesotho and Setswana. To increase our training data we systematically add speech data from languages in the same group to the target language.

Cross-language mapping is performed at the phoneme level based on the IPA-mapping described in the Lwazi phoneme set version 1.2., a phoneme set that is still undergoing further refinement [23].

Language combinations	# distinct phonemes
Sepedi	43
Sepedi + Setswana	46
Sepedi + Setswana + Sesotho	49
Sepedi + Setswana + Sesotho + isiZulu	65
Sesotho	41
Sesotho + Setswana	42
Sesotho + Setswana + Sepedi	49
Sesotho + Setswana + Sepedi + isiZulu	65
Setswana	34
Setswana + Sesotho	42
Setswana + Sesotho + Sepedi	49
Setswana + Sesotho + Sepedi + isiZulu	65

Table 2. The number of distinct phonemes for each Sotho-Tswana language cluster.

Table 2 shows the increase in the number of distinct phones when languages from the Sotho-Tswana group are added together (and also the count if isiZulu is added to these languages). Similarly, Table 3 shows the increasing count of distinct phones for the Nguni group. Column 1 in Tables 2 and 3, indicate the data pooling combinations which were used in the various ASR experiments.

5. COMBINATION RESULTS

In order to assess the performance of our combined ASR systems, phone recognition on the "base" languages is performed for all combined systems. We also measure several distances in order to quantify how far the languages are apart from one another.

5.1. Inter-phone comparisons

We investigate the "closeness" of languages by measuring several distances: the Bhattacharyya distances between overlapping multivariate normal distributions of monophone and triphone models, the Euclidean distance between overlapping phone durations and the cosine of the angle between phone-count vectors.

5.1.1. Comparison of acoustic similarities

The Bhattacharyya distance for multivariate Gaussian distributions,

$$D_B = \frac{1}{8}\left(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\right)^T \boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\mu_1}-\boldsymbol{\mu_2}\right) + \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma_1\Sigma_2}|}}\right) \quad (1)$$

Language combinations	# distinct phonemes
isiNdebele	48
isiNdebele + isiZulu	54
isiNdebele + isiZulu + isiXhosa	61
isiNdebele + isiZulu + isiXhosa + Siswati	64
isiZulu	47
isiZulu + isiNdebele	54
isiZulu + isiNdebele + isiXhosa	61
isiZulu + isiNdebele + isiXhosa + Siswati	64
isiXhosa	53
isiXhosa + isiZulu	57
isiXhosa + isiZulu + isiNdebele	61
isiXhosa + isiZulu + isiNdebele + Siswati	64

Table 3. The number of distinct phonemes for each Nguni language cluster.

is used to calculate distances between corresponding states of corresponding phones in pairs of languages. In (1), $\boldsymbol{\mu_i}$ denotes the mean vector of a particular multivariate distribution, $\boldsymbol{\Sigma_i}$ the corresponding covariance matrix and

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma_1}+\boldsymbol{\Sigma_2}}{2} \quad (2)$$

The Bhattacharyya distance is calculated for all monophones and triphones shared among languages. In order to obtain a single distance for both monophones and triphones, weighted sums are calculated, with each phone being weighted by the sum of its prior probabilities in the intersection of the languages being compared. The weighted sums, referred to as the acoustic distances, are displayed in tables 4 and 5.

5.1.2. Comparison of (tri)phone frequencies

Another way to assess the closeness of languages is to measure the similarities in the frequencies at which common monophones and triphones occur in those languages. We quantify these similarities in terms of the angle between the vectors containing the frequencies of all monophones / triphones in each of the languages:

$$\cos\left(\angle\left(\mathbf{x},\mathbf{y}\right)\right) = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}|\cdot|\mathbf{y}|}, \quad (3)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors containing the phone / triphone frequencies in two different languages.

The higher this value is, the more overlap exists between both phones and phone counts in these languages. Tables 4 and 5 summarize these measurements for the two language groups studied here.

5.1.3. Comparison of phone durations

A third way to measure how close two languages are, is to consider the similarity between the durations of phones common to both languages. Phone durations are obtained by forced alignment, using tied-state triphone models together with the semi-tied transforms. The mean durations of common phones within each language are then compared, calculated as the sum of the squared differences between these mean durations, normalized by the sum of the mean durations of the phonemes in the pair of languages under consideration. (Normalization is performed to prevent differences between longer phone classes such as vowels to dominate the analysis.) These normalized distances are also presented in tables 4 and 5.

Distance	Nso-Sot	Nso-Tsn	Sot-Tsn
Acoustic distance	1.024	1.157	1.167
Similarity of frequencies	1.162e-02	1.148e-02	1.581e-02
Distance between durations	0.110	0.097	0.090

Table 4. Distance measures between the South African Sotho-Tswana languages, as described in Section 5.1

Distance	Nbl-Xho	Nbl-Zul	Xho-Zul
Acoustic distance	1.316	1.232	1.144
Similarity of frequencies	1.396e-02	1.649e-02	1.301e-02
Distance between durations	0.100	0.097	0.095

Table 5. Distance measures between the Nguni languages employed in this study, as described in Section 5.1

5.2. Recognition results

Figure 2 summarizes the phone-recognition accuracies that were obtained by pooling different sets of data. (In all cases, a flat language model was employed - that is, each phone was allowed to transition to any other phone. As a point of reference, our baseline recognizers were found to have word error rates ranging between 2% and 12% on a ten-word speaker-independent recognition task.) It can be seen that all languages seemed to benefit from the addition of data from closely related languages, except Sepedi. isiZulu in particular performed much better with the addition of isiNdebele and isiXhosa, with an improvement of approximately 2.6% absolute. To assess the magnitude of this improvement, one needs to keep in mind that asymptotic phone-recognition accuracy (with unlimited training data) using only bigram constraints is substantially less than 100%. In earlier work [17] we used parametric fits of accuracy against the amount of training data to estimate asymptotic phone-recognition accuracies for these languages. Based on those calculations, we estimate that the additional accuracy achieved by adding isiNdebele data to the isiZulu training data (our most beneficial pooling) is similar to the benefit that would be achieved by adding approximately another 250 minutes of isiZulu training data. Similarly, the the addition of the Sesotho data to the Setswana recognizer is found to be equivalent to the addition of approximately 110 minutes of Setswana data.

We also see that adding languages from other sub-families (such as isiZulu to the Sotho-Tswana languages) degrades performance significantly, and that the addition of Siswati data to the other Nguni languages is also detrimental in all cases.

Comparing these results with the distance measures shown in tables 4 and 5 suggests that similarity in triphone frequencies is the best predictor of how well data pooling will work. Sepedi, for example, is further away from Sesotho and Setswana than any of the other languages by this measure, and this correlates with the phone recognition results in figure 2, where Sepedi does not add any value to either Sesotho or Setswana. Sesotho and Setswana both improve when adding data from one to the other, as do the Nguni languages, with the angle between the isiZulu and isiNdeblele phone-count vectors being particularly small. The comparison of phone durations is somewhat aligned with the observed recognition accuracies (compare, for example, the relationship between Sepedi and Sesotho), but the measure of acoustic differences that we have employed does not seem to predict the behaviour of data pooling at all. This measure does not correlate with either the assessments of the other two measures (which are fairly comparable in ordering the six languages studied here) or the recognition results observed.

6. CONCLUSION

In this paper we investigated the effectiveness of pooling speech data to improve ASR system performance of resource-scarce languages. We have shown that for both the Nguni and Sotho-Tswana language groups, a non-negligible improvement in ASR correctness can be achieved by combining appropriate speech data sourced from closely related languages. In the best case, approximately 520 minutes of isiNdebele training data is found to improve accuracy to a similar extent as would be expected from approximately 250 minutes of isiZulu data. The next best improvement, to Setswana from approximately 420 minutes of Sesotho data, was seen to be equivalent to approximately 110 additional minutes of Setswana data. These provide rough guidelines for the benefit that can be achieved from pooling speech data from closely related languages families – namely, that two to four hours of cross-language data can give similar benefit to one hour of target-language data.

The factors that influence data combination, as described in Section 2, should of course be kept in mind. It would there-

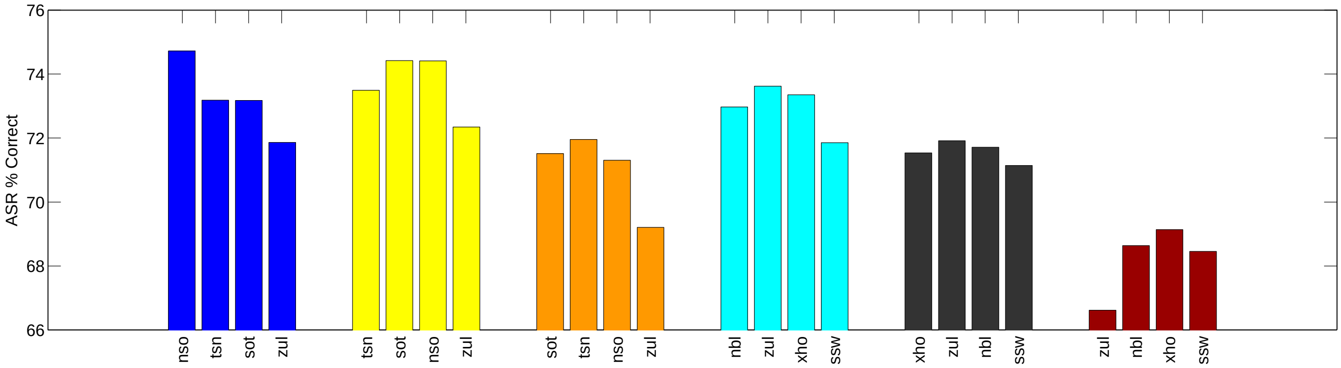


Fig. 2. ASR phone-recognition accuracies for Sepedi, Setswana, Sesotho, isiNdebele, isiXhosa and isiZulu. In each "cluster", the first bar indicates the baseline phone correctness for the particular language being recognized. Each subsequent bar is labelled with the language from which additional training data was added, in addition to all training data used for the previous bar. In this way, the 3^{rd} bar from the 2^{nd} (yellow) cluster, indicates the phone correctness obtained when recognizing Setswana, having used training data from Setswana, Sesotho and Sepedi.

fore be very interesting to repeat the comparisons performed here with different amounts of target and donor data, and also to investigate other language families with greater or lesser language similarities. It will also be interesting to see whether more elaborate data combination strategies can produce larger benefits from the combination of data from closely related languages.

Our results suggest that similarity in the frequencies of the various triphones is the best predictor of data-pooling performance amongst those measures studied here. This suggestion should be evaluated on data from other language families, and it may be fruitful to search for other measures that are even better predictors.

7. REFERENCES

[1] A. Constantinescu and G. Chollet, "On cross-language experiments and data-driven units for ALISP," in *Automatic Speech Recogntion and Understanding (ASRU)*, 1997, pp. 606–613.

[2] U. Bub, J Kohler, and B Imperl, "In-service adaptation of multilingual Hidden-Markov-Models," in *ICASSP*, Munich, Germany, 1997, pp. 1451–1454.

[3] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding*, Landsdowne, VA, 1998, pp. 259–252.

[4] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusumy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *ICASSP*, Adelaide, Australia, 1994, pp. 237–240.

[5] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," Seattle, WA, 1998.

[6] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *ICSLP*, Sydney, Australia, 1998, pp. 1819–1822.

[7] P. Dalsgaard and O. Andersen, "Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organising neural network," in *ICSLP*, Banff, Canada, 1992, pp. 547–550.

[8] U. Ackerman, B. Angelini, F. Brugnara, M. Federico, D. Giuliani, R. Gretter, G. Lazzari, and H. Niemann, "Speedata: Multilingual spoken data-entry," in *ICSLP*, Philadelphia, PA, 1996, pp. 2211–2214.

[9] L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna, "Testing generality in JANUS: A multilingual speech to speech translation system," in *ICASSP*, San Francisco, CA, 1992, vol. 1, pp. 209–212.

[10] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, Aug. 2001.

[11] C. Nieuwoudt and E. C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101–113, September 2002.

[12] O. Anderson, P. Dalsgaard, and W. Barry, “On the use of data-driven clustering techniques for language identification of poly- and mono-phonemes for four european languages,” in *ICASSP*, Adelaide, Australia, 1994, pp. 121–124.

[13] J. Kohler, “Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks,” Leusden, Netherlands, 1999, pp. 79–84.

[14] B. Imperl, Z. Kacic, B. Horvat, and A. Zgank, “Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones,” *Speech Communication*, vol. 39, no. 3-4, pp. 353–366, 2003.

[15] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*, Elsevier, 2006.

[16] T. Niesler, “Language-dependent state clustering for multilingual acoustic modeling,” *Speech Communication*, vol. 49, pp. 453–463, 2007.

[17] E. Barnard, M. Davel, and C. van Heerden, “ASR corpus design for resource-scarce languages,” in *Interspeech*, Brighton, UK, 2009, pp. 2847–2850.

[18] M. Paul Lewis, *Ethnologue: Languages of the World, Sixteenth edition*, SIL International, 2009.

[19] P.N. Zulu, G. Botha, and E. Barnard, “Orthographic measures of language distances between the official south african languages,” *The Literator: Journal of literary criticism comparative linguistics and literary studies*, vol. 29, pp. 185–204, April 2008.

[20] Steve Young, “Large vocabulary continuous speech recognition: a review,” in *of INCIS Project, Schedule 6 in (Small)*, 1996.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2009.

[22] M. J. F. Gales, “Semi-tied covariance matrices for hidden markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, May 1999.

[23] M. Davel and O. Martirosian, “Pronunciation dictionary development in resource-scarce environments,” in *Interspeech*, Brighton, UK, 2009, pp. 2851–2854.

VIETNAMESE MULTIMODAL SOCIAL AFFECTS: HOW PROSODIC ATTITUDES CAN BE RECOGNIZED AND CONFUSED

Dang-Khoa Mac [1,2], *Véronique Aubergé* [2], *Albert Rilliard* [3], *Eric Castelli* [1]

[1] International Research Center MICA, CNRS-UMI 2954, Hanoi, Vietnam
[2] Laboratory of Informatics of Grenoble (LIG), CNRS, France
[3] LIMSI-CNRS, Orsay, France

{dang-khoa.mac, eric.castelli}@mica.edu.vn
veronique.auberge@imag.fr, albert.rilliard@limsi.fr

ABSTRACT

Social affective expression is a main part of face-to-face interaction and it is highly linked to the language through the culture. This paper presents a study on Audio-Visual prosodic attitudes in Vietnamese, an under-resourced tonal language. Based on an audio-visual corpus of 16 attitudes, perception experiments were carried out with Vietnamese and French participants. The result analysis shows the relative contribution of audio, visual, and audio-visual information in attitude perception. It also shows how native and non-native listeners recognize and confuse the attitudes, thus allows us to investigate the cultural specificities and cross-cultural common attitudes in Vietnamese.

***Index Terms*—** Audio-visual corpus, Prosodic social affects, Cross-cultural perception, Vietnamese

1. INTRODUCTION

In speech communication between humans, the expression of mental, intentional, attitudinal, emotional states is a main information channel that is often used by both speaker and listener. Some theoretical models of affect claim that affective expression in speech communication may be controlled at different levels of cognitive processing [1], from the involuntarily controlled expressions of emotion to the intentionally, voluntarily controlled expressions of attitudes. According to [2], attitudinal expressions can be distinguished from emotional expressions by the nature of speaker's control on its expressivity (voluntary vs. involuntary). Some types of expressivity may be expressed as either an attitude or an emotion. For example, "surprise" can be considered an attitude when expressed during a voluntary process; otherwise it can be considered an emotion.

Attitude expression carries the intention and points of view of the speaker (ex: surprise, confirmation, politeness etc.). An utterance without any attitude means that the speaker does not give his opinion in this utterance. Attitudes are constructed for a language and a culture and they need to be learned by children or by second language students. As social affects, attitudinal expressions can vary amongst languages. Some specific attitudes in a language may not be recognized or may be ambiguous in other language. The understanding of this phenomenon may benefit from some cross-cultural studies [3][4].

Vietnamese is a tonal language; therefore the acoustic parameters which are implied in the linguistic and affective functions of prosody play an important role at the phonemic level for lexical access. The Vietnamese tone system has 6 tones: level (1), falling (2), broken (3), curve (4), rising (5) and drop (6). Tone 5b and 6b correspond to tone 5 and 6 on a syllable ended by a stop consonant. A special feature of the Vietnamese tone system is the co-occurrence of glottalization during the production of tone 3 and tone 6. For example, tone 3 is accompanied with harsh voice quality due to a glottal stop (or a rapid series of glottal stops) around the middle of the vowel. Tone 6 has the same kind of harsh voice quality as tone 3; however, it is distinguished by dropping very sharply and it is almost immediately cut off by a strong glottal stop [5]. These phenomena of voice quality cues also happen in the morphology of some attitudes (and emotions) in other languages [3][6].

This paper presents our study of Vietnamese multimodal social affect in a Vietnamese and French cross-cultural context. Because of the contrast of language characteristics (non-tonal vs. tonal language) and the long geographic and cultural distance (West European vs. East-Asian), French was chosen for this cross-cultural study of Vietnamese social affect. This study was done not only in audio modality but also in visual modality in order to investigate the relative contribution of audio, visual, and audio-visual information in the perception of attitudes for both Vietnamese and French participants.

After presenting the Vietnamese corpus construction and the attitude selection, the perception experiment is described. The experimental results are then presented and analyzed. The results show how the native and non-native listeners can recognize and confuse the Vietnamese attitudes. After the discussion, this paper ends with some conclusions and perspectives.

2. PERCEPTION EXPERIMENT

2.1. Attitude selection

Prosodic social affects have been studied in different languages such as English, French, and Japanese [3,4,7]. For these languages, attitudes have been selected thanks to the foreign language didactics' literature. Unfortunately, as Vietnamese is an under-resourced language, there is very little research on Vietnamese expressive speech. We have found only one study [8] dealing with this topic. From this study, we selected 16 attitudes to be examined in Vietnamese speech (Table 1).

Table 1: *16 selected Vietnamese attitudes, with their abbreviations*

Declaration	DEC	Irritation	IRR
Interrogation	INT	Sarcastic irony	SAR
Exclamation of neutral surprise	EXo	Scorn	SCO
Exclamation of positive surprise	EXp	Politeness	POL
Exclamation of negative surprise	EXn	Admiration	ADM
Obviousness	OBV	Infant-directed speech	IDS
Doubt-Incredulity	DOU	Seduction	SED
Authority	AUT	Colloquial	COL

These 16 attitudes were selected in order to investigate their existence and their realization in Vietnamese. The "*exclamation of surprise*" was divided into three sub-types: "*neutral*", "*negative*" and "*positive*" to verify whether or not they can be distinguished in Vietnamese.

2.2. Corpus construction

The corpus was constituted of 125 skeleton sentences without specific affective meaning in order to be produced naturally in all 16 attitudes. To observe the effects of tone and tonal co-articulation on attitudinal expression, the corpus contains 8 sentences of one-syllable length, which correspond to 8 representations of Vietnamese tones, and 72 sentences of two-syllable length, which correspond to all combinations of two tones among the 8 Vietnamese tones. The remainder of the corpus is based on 45 sentences from 3- to 8-syllable length and systematically varied in their syntactic structure: single word, nominal group, verbal group and a simple structure "subject-verb-object".

One male speaker, native of Hanoi (standard pronunciation), was chosen to record the corpus. A training phase was carried out in order to ensure that the speaker expressed each attitude as naturally as possible. The corpus was recorded in a sound-proof room. A high quality microphone (AKG C1000S) was placed approximately 40 cm from the speaker's mouth. The microphone was connected to a computer outside the room through an USB sound device. The speech was recorded at 44.1 kHz, 16bits. During the recording, a digital DV camera (Sony DXC990) recorded the speaker's performance. The video clips were encoded with IndeoVideo codec at 784x576 pixels resolution. Vocal fold's vibrations were also measured using an electroglottograph. To control the speaker performance, a specialist in expressive speech and a native Vietnamese speaker observed the recording process from outside the room, through a video system. They could require the speaker to re-produce a stimulus if they thought that it was not performed satisfactorily. The speaker pronounced all 125 sentences in 16 attitudes. The complete corpus contains 2000 stimuli. It corresponds to more than 90 minutes of audio-visual signal after post-processing.

2.3. Experimental protocol

Three skeleton sentences of one-, two- and five-syllable length were chosen from the corpus for the perception experiment. We note that most of Vietnamese words are mono-syllabic or bi-syllabic [8]. As mentioned above, the Vietnamese tone system has certain characteristics that have been shown to be used in the morphology of some attitudes. Therefore the perception of attitude can be affected by tones. In order to limit the complexity of the test, the influence of tone was not investigated in this experiment (it will be studied in another experiment). The three selected sentences include no tone variation: all syllables are based on tone 1 (the level tone). These sentences were then presented in 16 attitudes and in three modalities (audio-only, visual-only and audio-visual). Thus, there were 3*16*3=144 stimuli in the perception test.

Forty listeners participated in this experiment: 20 Vietnamese (10 males and 10 females with a mean age of 25) who speak the same dialect as the speaker; and 20 French (10 males and 10 females with a mean age of 35) who have no experience on Vietnamese language. Both of these Vietnamese and French participants were separated into two groups. The first group listened to the audio-only stimuli first, then watched the video-only stimuli, and finally watched the audio-video stimuli. The second group started with the video-only stimuli, continued with the audio-only stimuli and ended with the audio-video stimuli. For each listener, the stimuli in each modality were chosen randomly in order to counterbalance a possible effect of stimuli presentation order.

The perception tests were carried out in a quiet room, using a high-quality headset (Sennheiser HD 25-13) at a comfortable hearing level. The testing program interface gave the label and the explanation of the 16 attitudes (in the native language of the listener). No listener expressed any difficulty in understanding the concepts of these 16 attitudes. All subjects listened to (and/or watched) each stimulus only once. After each stimulus, they were asked to indicate the perceived attitude among the 16 presented attitudes.

3. EXPERIMENT RESULT

3.1. Attitude recognition

Figure 1 presents listeners' recognition rates of 16 attitudes in three modalities. Globally, most of attitudes were recognized above chance level, and native listeners have higher recognition scores than foreign ones. Some attitudes were well recognized by both Vietnamese and French listeners, such as DEC, EXp, DOU, AUT, IRR, SCO, SED attitudes. The INT, IDS and COL attitudes were well recognized by Vietnamese listeners but were almost not recognized by the French listeners. In case of ADM attitude, the French listeners' recognition rate is higher than that of Vietnamese listeners.

The modality (Audio only, Visual only and Audio-visual) has a strong effect on attitude perception. As expected, for most attitudes, the average score in audio-visual modality is better than that in audio-only or visual-only modality. For the Vietnamese listeners, the audio information is very important to recognize the DEC, EXo, OBV, AUT and COL attitudes and the visual information play an important role to recognize the EXp, DOU, SCO, POL attitudes. With the French listeners, the audio information is more important to recognize the AUT and IRR attitudes, and the visual information is much more necessary to recognize the DEC, EXp, SCO and ADM attitudes.

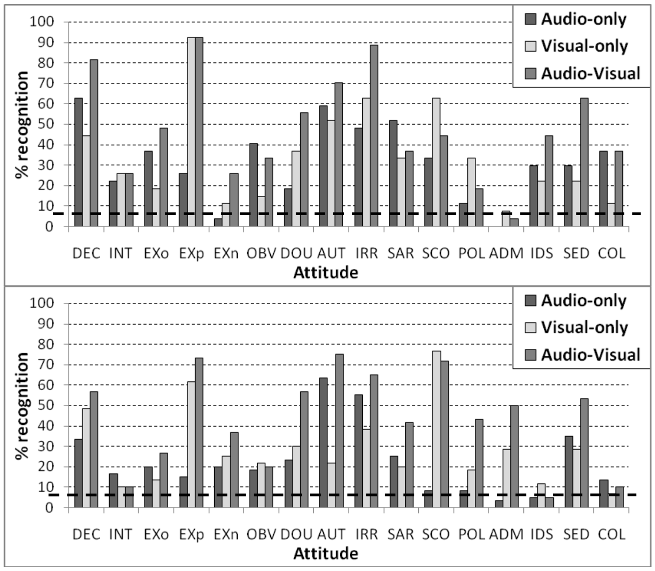


Figure 1: *Recognition rate of each attitude in each modality for Vietnamese listeners (top) and French listeners (bottom). The dash lines: chance level (6.25%)*

3.2. Attitude confusion

From the confusion matrices, confusion graphs were built based on all the confusions higher than twice the chance level (≥ 12.5%). Figure 2 shows the graphic presentation of the confusion among 16 attitudes, in three modalities for Vietnamese and French listeners.

With the audio-only information, the ADM was not recognized by both Vietnamese and French listeners. This attitude was confused with EXo (in case of Vietnamese listeners) and confused with COL and IDS (in case of French listeners). The Vietnamese listeners also did not recognize the EXn and the French listeners did not recognize the IDS. The Vietnamese listeners made a mutual confusion between some pairs/groups of attitudes, such as SAR and SCO; POL and DEC; EXo EXn and DOU. The French listeners have the mutual confusion between AUT and IRR; DOU and EXn; DOU and EXo.

With only the visual information, all attitudes were recognized above the chance level, with both Vietnamese and French listeners. The Vietnamese listeners have the mutual confusion of EXn and DOU; SED and COL; DEC and OBV. The French listeners have the mutual confusion between: EXn and DOU; COL and SED; ADM and EXp. They also strongly confused the SAR with SCO (60%).

As expected, the confusion graph in the case of audio-visual shows less confusion than in case of Audio and Video only. However, several attitudes have the recognition rate below the chance level (ADM for Vietnamese listeners and IDS for French listeners). The Vietnamese listeners confuse between EXn and DOU; SAR and SCO; POL and OBV. The French listeners have also the mutual confusion between SED and COL; EXn and DOU; SAR and SCO.

5. DISCUSSION

According to experimental results, although the mean intensity scores obtained by French listeners are lower than those of Vietnamese, they are fairy coherent with the result of Vietnamese listeners. For both groups of listeners, some attitudes were well recognized: DEC, Exp, DOU, AUT, IRR and SED. It supposes that the concepts and the expressions of these attitudes are similar in the two languages and the two cultures. So they can be seen as cross-cultural social affects (for Vietnamese and French).

Some pairs of attitudes (such as SAR and SCO; EXn and DOU) show a mutual confusion. In the audio channel, this confusion can be explained by the similarity of prosodic characteristic in the expression of these attitudes (the F0 contour, the intensity or the voice quality characteristics). Figure 3 gives an example of the F0 contour of two attitudinal expressions (DEL and POL) with the same sentence. The prosodic forms of these attitudes look nearly similar. Therefore, it is very difficult to distinguish these attitudes with only audio information.

Some attitudes (INT, IDS and COL) are recognized quite well by native listeners, however they are nearly not recognized by non-natives. Perhaps, the prosodic performances for these concepts of Vietnamese are not shared with French and they need to be learned by foreign students.

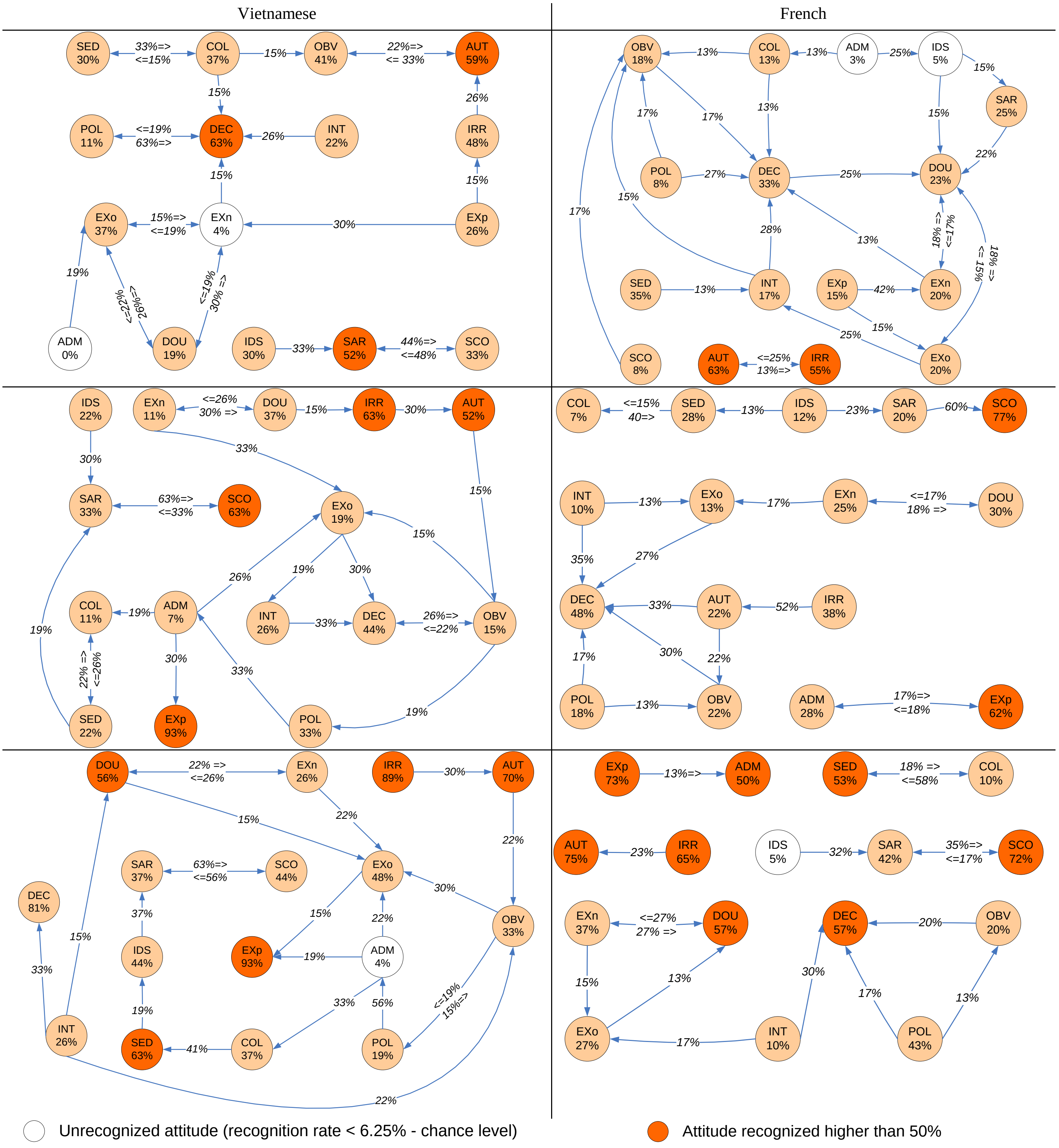


Figure 2: *Confusion graph for the 16 attitudes of Vietnamese and French listeners in Audio only (top), Video only (middle) and Audio visual (bottom)*

Similar conclusions were already discussed for some Japanese attitudes, which are not recognized by French or English [3,7]. An interesting case is the expression of Admiration, which is badly recognized by native listeners but is better recognized by the non-native ones (in visual and audio-visual modalities). Perhaps, for Vietnamese, this attitude cannot occur without lexical coherency [8]. Otherwise, in French, this concept exists and it can be expressed and can be perceived easily by speech prosody or/and gesture of speaker's face.

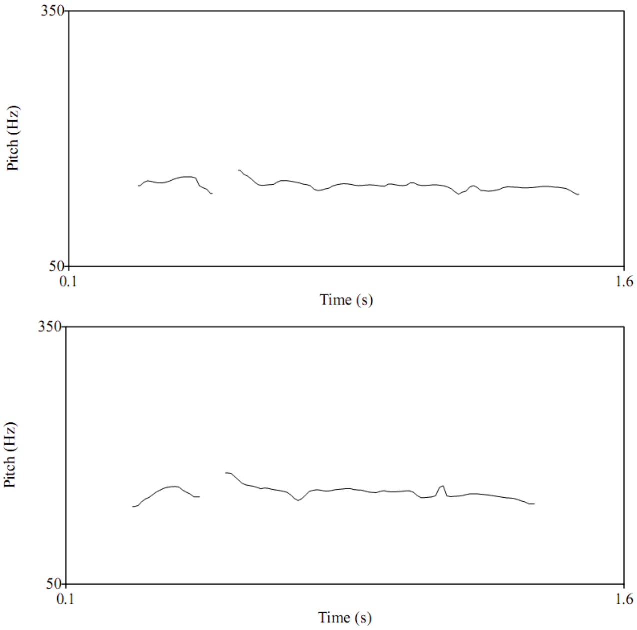


Figure 3: *F0 contour of 5-syllable length sentence in 2 attitudinal expressions: DEC (top) and POL (bottom)*

6. CONCLUSIONS AND PERSPECTIVES

Using the cross-cultural perception of audio and visual social affect in Vietnamese, the speaker's performance for 16 Vietnamese attitudes was quite well evaluated by native and non-native listeners. Experimental results reveal the influential factors on the attitudinal perception: the modality of presentation and the attitudinal expression itself. These results allow us to investigate the cultural specificities and the cross-cultural perception of Vietnamese attitudes, and also raise interesting questions for future researches as well as for educational purposes – mostly in the field of foreign language teaching.

However, the results need to be further validated by a deeper prosodic analysis to find out the acoustical and visual parameters that lead to the perception of these social affects. Other perception experiments including variations of Vietnamese tones are scheduled in order to explore the importance of such a tonal system on the perception of attitudes not only for native, but also for foreign speaker without any linguistic knowledge of a tonal language: will they be able to separate tonal from attitudinal information?

7. ACKNOWLEDGEMENTS

We are deeply grateful to Christophe Savariaux for his efficient technical contribution, as well as to the subjects of our experiments. This study was done in the framework of the KC.03.15/06-10 project.

8. REFERENCES

[1] Scherer, K.R., & Ellgring, H., "Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? ", Emotion, 7(1), 2007, pp. 158-171

[2] Aubergé, V., "A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP", Speech Prosody, 2002.

[3] Shochi, T., Aubergé, V., and Rilliard, A., "How prosodic attitudes can be false friends: Japanese vs. French social affects", Speech Prosody, Dresden, 2006, pp. 692-696.

[4] Shochi, T., Aubergé, V. & Rilliard, A. (2007). Cross-Listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes. Proceedings of ICPhS, Saarbrucken, Germany, 2007, pp. 2097-2100

[5] Do T.D., Tran T.H. & Boulakia G., "Intonation in Vietnamese", in Intonation systems: A survey of 22 languages, D. Hirst and A. Di Cristo, Eds.: Cambridge University Press, 1998, pp. 395-416.

[6] Shochi, T., Erickson, D., Rilliard, A., and Aubergé, V., "Recognition of Japanese attitudes in Audio-Visual speech", in Speech Prosody, Campinas, Bresil, 2008, pp. 689-692.

[7] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D. "Intercultural Perception of English, French and Japanese Social Affective Prosody", in The role of prosody in Affective Speech, ed. S. Hancil, Linguistic Insights 97, Peter Lang AG, Bern, 2009, pp.31-59.

[8] Le T.X., "Etude contrastive de l'intonation expressive en français et en vietnamien", PhD thesis of Linguistic and Phonetic, Université Paris 3, 1989.

FROM TONE TO PITCH IN SEPEDI

Etienne Barnard[1], *Sabine Zerbian*[2]

[1] Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
[2] Department of Linguistics, University of Potsdam, Germany
ebarnard@csir.co.za, szerbian@uni-potsdam.de

ABSTRACT

We investigate the acoustic realization of tone in continuous utterances in Sepedi (a language in the Southern Bantu family). Human labelers marked each of the 271 syllables in a 15-sentence corpus produced by a single speaker as "high" or "low". Automatic pitch extraction was then used to estimate the fundamental frequencies of the voiced segments of each of these syllables. Statistical analysis of the resulting pitch contours confirms that the mean pitch frequencies of the syllabic nuclei serve as the primary indicator of tone, with the relative frequencies of successive syllables being the most relevant measure. Our analysis also suggests that additional factors may play a role in the production and perception of tone.

Index Terms— Tone languages, pitch contours, Sepedi, Southern Bantu

1. INTRODUCTION

Southern Bantu languages are tone languages in which word-level pitch variations generally convey both lexical and grammatical meaning. In contrast to tone languages like Chinese, they are agglutinative languages, i.e. several morphemes are joined together in a word. Although most Southern Bantu languages only have two level tones, namely high tone (H) and low tone (L), modeling of their prosody is complicated by the agglutinative morphology, the significant influence of grammar and the occurrence of tone sandhi within and across words. Given the role of word-level prosody in processes such as semantic interpretation and the production of natural speech, it is important that a detailed and systematic account of the prosody be given. Such an account is complicated by the fact that tonal information is not indicated in the orthography of many Bantu languages (including Sepedi, which is the focus of the current study).

We have recently presented an overview of intonation in the Southern Bantu languages [6], from which we concluded that a detailed understanding of the tone system of these languages is especially important for the creation of natural-sounding text-to-speech (TTS) systems. Such an understanding will require progress in two areas, namely (a) deriving tone assignments from text and (b) understanding the relationship between physical parameters (such as pitch frequency) and the tone levels. It is the second of these tasks that is the focus of the current investigation (initial work on the first task was presented in [8]).

Below, we briefly review a number of pertinent facts on tone in the Sotho-Tswana languages (of which Sepedi is a representative). We also summarize the goals of the current study in more detail, and present the experimental methodology that was followed in pursuit of these goals (Section 3). Our results are contained in Section 4, and Section 5 contains a discussion of our main conclusions and future work that is required to complete the current investigation.

2. TONE IN THE SOTHO-TSWANA LANGUAGES

Most Southern Bantu languages are tone languages whose surface tones can be captured by two level tones, namely high (H) and low (L) [3]. The high tone is the active tone in Sotho-Tswana languages such as Sepedi, as it participates in tone spread and is subject to positional restrictions. As is the case for most Bantu languages, the Sotho-Tswana languages show an asymmetry in the tonal characteristics of its noun and verb system with nouns being more tonal than verbs: whereas nouns can contrast tone on every syllable, verbs only contrast tone on their stem-initial syllable.

By definition, the primary distinctive feature of a level tone is the value of the pitch frequency within the nucleus of a given syllable, with H generally having a higher pitch frequency than L. This general observation was confirmed in our earlier investigations [7], which focused on the temporal alignment of a single high tone within the verbal domain. (As is common practice, we measure the fundamental frequency (F0) as a physical indicator of the pitch frequency.)

However, the more general question of how these pitch values are related to one another in a complete utterance, as well as the details of the temporal trajectories of F0 within and between syllables, have not been investigated systematically. The main aim of the current paper is to present initial findings on how these physical quantities are related to surface tone values in Sepedi. That is, we seek to understand how a speaker of Sepedi chooses to express the difference between H and L tone levels, given the considerable latitude inherent in the specification that H should carry "higher pitch" than L.

3. METHODS AND CORPUS

Our analysis is based on the speech of a 30-year old male speaker, who was selected for the development of a Sepedi corpus for a concatenative text-to-speech system [5]. As part of that development, it was ensured that the speaker employs the standard Sepedi dialect, and he then recorded 299 sentences that give a balanced coverage of the most common diphones in Sepedi. In accordance with the requirements for TTS development with a limited corpus, the speaker was requested to speak naturally, but with relatively flat intonation.

Of these sentences, 15 were selected for analysis (based on factors such as the absence of loan words and proper nouns, and limitations on the mood of the verb to limit the influence of dialectal variations). These sentences were automatically aligned using a Hidden Markov Model recognizer. All syllables were subsequently labeled for tone by three labelers independently of each other. The labelers are sensitive to issues of tone but differ in their background and experience regarding Bantu tone. The individual labels were based on perception of the recorded sentences, acoustic analysis using the *Praat* software package [1] or both. The labeled sentences were compared across all three labelers, which revealed unanimous agreement on the tone labels in 72.3% of the cases (196 out of 271 syllables). A final transcription was generated based on the majority vote, i.e. the tone label selected by at least two labelers. (The flat intonation of the corpus might have been one of the reasons for cases of disagreement between labelers' decisions.)

The autocorrelation-based pitch tracker in *Praat* was employed to estimate the pitch contours (that is, the value of F0 as a function of time) for all utterances. As can be seen in Fig. 1, the computed contours are generally quite smooth (and the F0 values are found to be quite accurate). The exceptions generally occur at the edges of the voiced segments, where the voicing is generally less robust and the F0 estimates less accurate. Because of the smoothness of the pitch contours, we describe the F0 values of each syllable in terms of the smoothed initial and final F0 values of the vowel segment. These are calculated using the boundaries found by automatic alignment, as follows:

- The F0 values corresponding to the initial two pitch periods as well as the final two pitch periods are discarded.
- A least-squares linear fit of F0 as a function of time is computed from the remaining values.
- The initial and final values of F0 are estimated as the value of the linear fit at the respective boundaries of the vowel.

Mean pitch values in each segment, as well as the change in pitch across each segment, are estimated based on these linear parameters. Although this processing does markedly improve the robustness of the estimated values, it does not compensate completely for the pitch tracking errors that occur unavoidably. In addition, the automatic segmentation is not completely accurate, and the ambiguity in tone labeling also introduces some uncertainty. For all these reasons, there will be a fair amount of measurement noise in the results reported below; we return to this matter in the Conclusion.

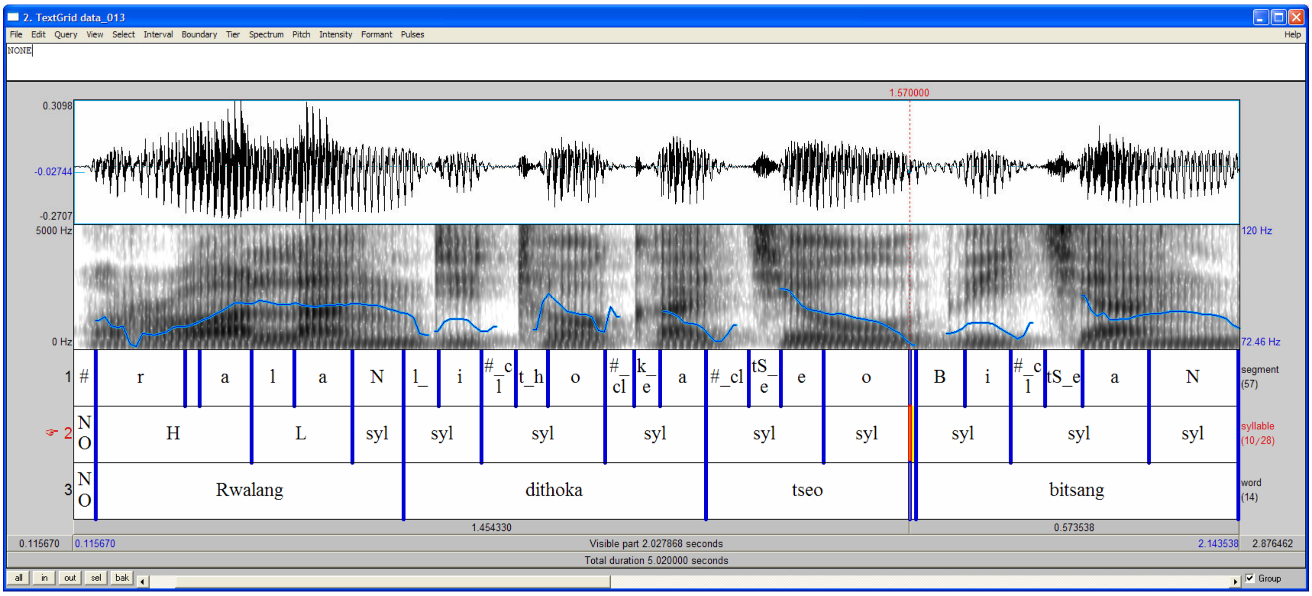

2. TextGrid data_013
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
NONE
1.570000
0.3098
-0.02744
-0.2707
5000 Hz
0 Hz
120 Hz
72.46 Hz
1
# r a l a N l_ i #_cl t_h o #_cl k_e a #_cl tS_e e o B i #_cl tS_e a N
segment (57)
2
NO H L syl syl syl syl syl syl syl syl syl
syllable (10/28)
3
NO Rwalang dithoka tseo bitsang
word (14)
1.454330
0.573538
0.115670
0.115670
Visible part 2.027868 seconds
2.143538
2.876462
Total duration 5.020000 seconds
all in out sel bak
Group

Figure 1: *Spectrogram and segmentation of typical Sepedi utterance used in the current study. The pitch contour is shown in blue, superimposed on the spectrogram.*

4. RESULTS

Figure 2 shows the overall distribution of mean F0 values for all segments in the corpus. As expected, the H segments generally have higher mean F0 values than the L segments, but there is considerable overlap between the two classes. This overlap is a predictable consequence of the fact that pitch generally declines throughout an utterance, so that both H and L pitch values are systematically reduced towards the end of each utterance. (The same tendency was, for example, observed for pitch levels in Mandarin [2].)

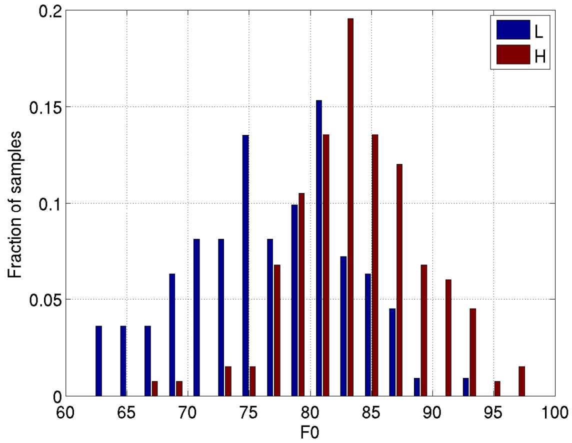


Figure 2: *Distribution of mean F0 values for H and L syllables, respectively*

If the declination in pitch were a complete explanation for the overlap in H and L pitch values, one would expect the relative values between consecutive segments to be a better indicator of the intended tone – such relative F0 values were indeed found to be indicators of F0 perception in Vietnamese [4]. In Figures 3 and 4 we therefore show histograms of the difference between the mean F0 values of successive pairs of vowels, where the first vowel is labeled as H and L, respectively. It can be seen that L-to-H transitions tend to produce an increase or slight decrease in the mean pitch, whereas H-to-L transitions tend to result in a large decrease in mean pitch; L-to-L and H-to-H transitions fall somewhere between these extremes. Statistics confirming these tendencies are presented in Table 1. Note, however, that significant overlaps occur between all four cases, suggesting that the relative mean pitch values do not offer a complete expression of the speaker's intended tone level.

Condition	**Mean change in F0 (Hz)**	**Standard deviation of change in F0**
L-H transitions	4.881	5.650
H-H transitions	-0.183	4.959
L-L transitions	-3.452	5.888
H-L transitions	-6.409	4.576

Table 1: *Statistics of changes in mean F0 values between successive syllables.*

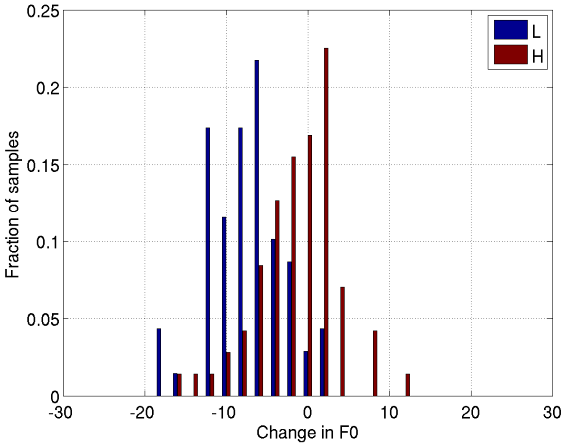


Figure 3: *Distribution of change in mean F0 values* ***between successive*** *syllables, when the first syllable was H.*

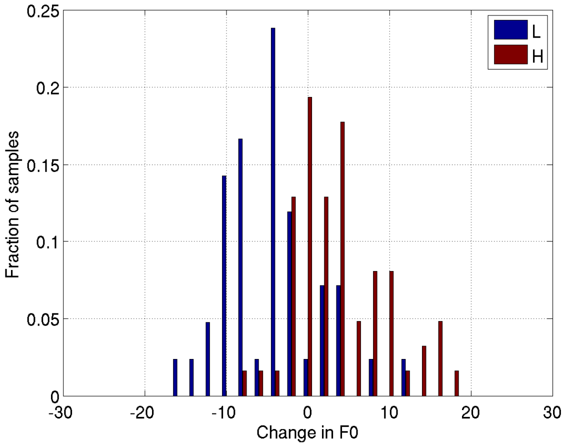


Figure 4: *Distribution of change in mean F0 values* ***between successive syllables****, when the first syllable was L.*

Inspection of pitch tracks such as that shown in Fig. 1 suggests another possible source of distinction between H and L, namely the overall slope of the pitch contour within a syllable (or syllable nucleus). As can be seen in Fig. 5, which represents a histogram of the overall changes in (smoothed) F0 values within each syllable nucleus, this feature does indeed take on somewhat different values for the two tones (though it is not strongly distinctive). The histograms of this feature for the various transitions (Figures 6 and 7) demonstrate that this feature is virtually irrelevant for syllables following an H syllable, but that it is somewhat distinctive for syllables preceded by an L syllable.

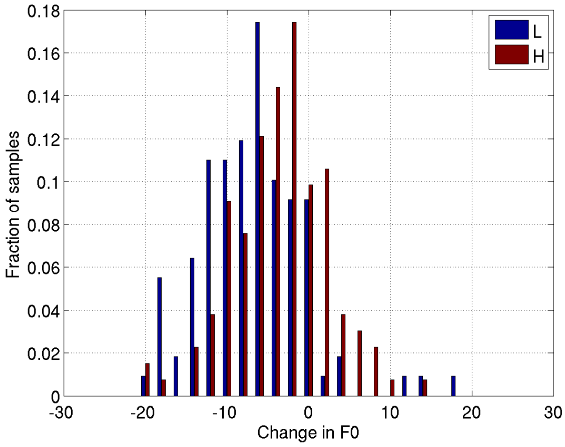


Figure 5: *Distribution of change in F0 within the syllable nucleus for H and L syllables, respectively*

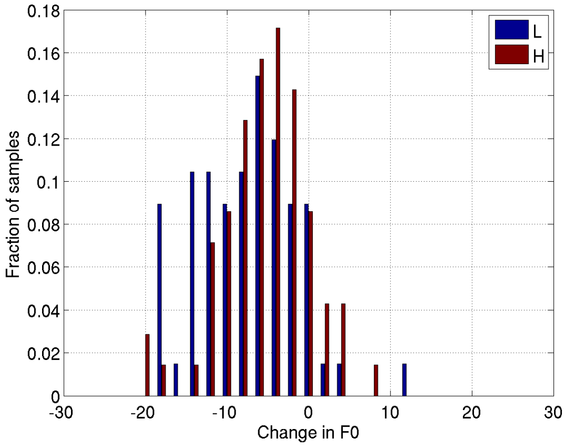


Figure 6: *Distribution of change in mean F0 values* ***within the syllable nucleus****, when the first syllable was H.*

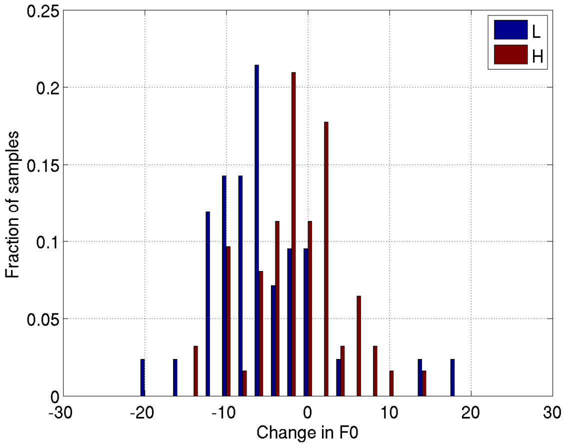


Figure 7: *Distribution of change in mean F0 values* ***within the syllable nucleus****, when the first syllable was L.*

5. CONCLUSION AND OUTLOOK

We have found that the mean pitch within the syllable nucleus is a strong indicator of the tone perceived in Sepedi speech. Not surprisingly, we find that the absolute pitch level is less important than relative pitch, which implies that the *change* in the mean pitch is the strongest indicator of tone amongst the signatures investigated here.

The change in mean pitch is nevertheless not a perfect indicator of tone in our data, as indicated by the overlap of the histograms shown in Figures 3 and 4. It is possible that these overlaps are simply the result of ambiguities in the tone labels and errors in alignment and pitch extraction (as discussed in Section 3). Some of the outliers in our results can certainly be attributed to such factors; however, the large number of syllables with overlapping values for the change in F0 leads us to suspect that other factors may be at stake. Figure 7 suggests that, in some cases, the intra-syllabic trend of F0 may be used to indicate tone, for syllables following an L syllable. Other factors that we have investigated were less promising – for example, consideration of the tone and mean pitch values of surrounding syllables does not produce better separation of the low and high tones. We have seen some evidence that the segmental make-up of a syllable may have an effect on the way that tone is expressed [6], but in the current corpus that influence is not evident.

To resolve these issues, we plan to analyze larger sets of sentences. It will be useful to consider speech from other speakers, to learn whether different speakers employ different strategies to communicate tone. It will also be interesting to perform comparative analyses of other Southern Bantu languages: whereas closely related languages such as Sesotho and Setswana are expected to be quite similar to Sepedi with respect to the phonetics of tone, somewhat more distant languages (e.g. isiXhosa and isiZulu) are likely to display some additional phenomena (e.g. depressor consonants [3]).

The successful application of these insights in speech-technology systems will be strong confirmation of their validity. We are in the process of developing all the components necessary to build a tone-aware TTS system for the Sotho-Tswana languages – the algorithm for tone assignment from text [8] is partially worked out, and the compilation of a sufficiently complete tone-marked pronunciation dictionary remains as the biggest challenge in that regard. The completion of this TTS system will allow us to carry out comprehensive perceptual tests to evaluate our ability to model tonal processes in Sepedi.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance of Ben Khoali, who assisted with the tone marking process and Charl van Heerden, who provided technical assistance. EB was generously supported by a Google Research Award.

7. REFERENCES

[1] Boersma, P., *Praat, a system for doing phonetics by computer.* Glott International, Amsterdam 2001

[2] Chen, S-H and C-C. Kuo, "Perceptual Relevance of Pitch Contours of Mandarin Tones and its Efficacy in Prosody Generation of Speech Synthesis" In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pp. 2669-2672, 2007

[3] Kisseberth CW and D. Odden. "Tone." In Nurse D & Philippson G (eds) *The Bantu languages.* Routledge, London, New York: , pp. 59–70. 2003

[4] Tran D.D, Castelli E., Serignat JF., Le X.H., Trinh V.L., "Influence of F0 on Vietnamese syllable perception", In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2005)*, Lisbon, Portugal, pp. 1697-1700, 2005.

[5] Van Niekerk, D.R. and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages" In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, 2009

[6] Zerbian, S. and E. Barnard. "Phonetics of intonation in South African Bantu languages", *Southern African Linguistics and Applied Language Studies*, 26(2): 235–254, 2008

[7] Zerbian, S. and E. Barnard. "Realizations of a single high tone in Northern Sotho", *Southern African Linguistics and Applied Language Studies* 27(4): 357–379, 2009

[8] Zerbian, S. and E. Barnard, "Word-level prosody in Sotho-Tswana", *submitted for publication*, 2010

DEVELOPMENT OF ANNOTATED BANGLA SPEECH CORPORA

Firoj Alam, S.M. Murtoza Habib, Dil Afroza Sultana, Mumit Khan
Center for Research on Bangla Language Processing, BRAC University
{firojalam, habibmurtoza}@bracu.ac.bd, upama.008@gmail.com, mumit@bracu.ac.bd

ABSTRACT

This paper describes the development procedure of three different Bangla read speech corpora which can be used for phonetic research and developing speech applications. Several criteria were maintained in the corpora development process that includes considering the phonetic and prosodic features during text selection. On the other hand, a specification was maintained in the recording phase as the speaking style is a vital part in speech applications. We also concentrated on proper text normalization, pronunciation, aligning, and labeling. The labeling was done manually – in the present endeavor sentence level labeling (annotation) was completed by maintaining a specification so that it could be expanded in future.

Index Terms **—**Speech corpora, Phonetic research, Speech processing

1. INTRODUCTION

The goal of this paper is to present the development of Bangla annotated read speech corpora which is essential for all kinds of speech processing work starting from acoustic analysis to the development of speech synthesis and speech recognition. These corpora were composed from three different corpora and those were developed for three different purposes.

1. “Corpus for acoustic analysis” was developed for acoustic analysis of Bangla phoneme inventory.
2. “Diphone corpus” was developed for diphone concatenation based speech synthesis.
3. “Continuous speech corpus” was developed for intonation model and unit selection based speech synthesis.

Though, these corpora were developed for different purpose however the use of this resource is innumerable i.e. speech recognition, speaker identification, and spoken information extraction. This resource is also an essential component in linguistic analysis of a language. Compared to other languages, very little work has been done in Bangla. CDAC [1] has developed speech corpora, but there is no published account about the details of their corpora. Since the current Bangla speech synthesis system [2] lacks the naturalness due to the intonation model, hence developing an intonation model from these corpora was one of the primary goals. It is also hoped that, a good unit selection based speech synthesis can be developed from these corpora which may have a naturalistic sound.

A brief literature review is given in section 2, followed by a description of the development in section 3, and section 4 presents the corpus annotation and analysis. Conclusion and future remarks of the study are given in section 5.

2. LITERATURE REVIEW

Several studies show significant improvement on designing, developing and annotating of the corpus. The bases of these are [3][4][5][6][7]. Though depending on the purpose of the corpus, different text pattern have chosen, but the developing process remains the same. Other than the designing and development procedure, significant work [8] [9] has proven the high performance of corpus based synthesizers. This signifies that, developing a phonetically and prosodically rich corpora can lead us to develop better speech applications.

3. DEVELOPMENT PROCEDURE

The development of corpora was done in three steps [3]; such as text selection, speaker selection and voice recording. According to [4] and [5], the following characteristics have to be considered during the development of the corpora:

- Area of speech corpora: Speech synthesis, phonetic research and speech recognition.
- Spoken content: Two approaches considered such as domain and phonological distribution.
- Professional recording studio: This is necessary for a clear acoustic signal from which it is possible to get clear acoustic information.
- Speaking style: Continuous read speech.
- Manual segmentation: Though this leads to significant amount of effort but it also affirm the accuracy of the labeling.
- Recording setup: Supervised onsite recording.

3.1. Text selection

Corpus for acoustic analysis: There are two categories of text in this corpus, one is for vowel and another is for consonants. For consonants phoneme investigation, the list of words selected consists of all possible phonemes with the following two patterns: vCv [iCi] and vCv [aCa], embedded in carrier words to form utterances. To maintain the same context we have embedded the consonants in a carrier utterance. So a total of 35x2 = 70 (35 [12] possible phonemes x 2 patterns) utterances were selected to record the data of the following form.

1. aCa pattern

আমরা কাজ পাই -> ক
amra kaɟ pai -> /k/
1stP.Pl work get.pres
[We get work.]

2. iCi pattern

আমি কিছু পাই -> ক
ami kic^{h}u pai -> /k/
1st.Sg some get.pres
[I get something.]

For vowel phoneme investigation, three different patterns were selected with the nearest number of phoneme segment in each pattern. Each pattern carried two to three syllables. The main intuition of selecting these patterns was duration calculation and formant measurement of vowels. These patterns are:

1. cV.Cv.cvc where V is the target vowel and C is either voiced or voiceless plosive.
2. cV.v.cvc where V is the target vowel
3. cV?V.Cvc where V?V indicates diphthong and C is either voiced or voiceless plosive.

The tricky part was the data collection comprising these patterns. For each pattern we have selected four words from the dictionary [10] to make a balance of the recording data. For the first pattern, the C of second syllable is voiced plosive in two words and voiceless in two words. The reason behind this is that the vowel before voiced is longer than the voiceless plosive [11]. So we will get average duration in both cases. For the second pattern, we were unable to find any word from the dictionary [10]. Then we changed the pattern to v.v.cv.cvc, as the main intention of this pattern was to calculate the duration of the two consecutive vowels appearing in two syllables. However, with this new pattern we found only two words. For the third pattern, four words were selected for each target diphthong. But due to word limitation of this pattern, the first consonant of the second syllable was chosen arbitrarily. In some cases we got only two words then we repeated these two to make it four which was a criterion to make a balance of all phonemes in this analysis. After that, another two carrier words were embedded to form sentences. For example,

এখন গবেষক বলো
ek^{h}on gɔbeʃok bɔlo
Now researcher say.pres
[Now say researcher]

The middle word is our target word. The vowel of the first syllable of the target word is the target phoneme. The list of words selected for this investigation consists of all possible vowel phonemes with the above patterns, embedded in carrier words to form the utterances. A total of 192 utterances were designed for recording with the following form:

1. 14x4 (14 possible phonemes x 4 words) = 56 (pattern cV.Cv.cvc)
2. 1x4 (1 phoneme x 4 words) = 4 (pattern cV.v.cvc)
3. 33x4 (33 possible phonemes x 4 words) = 132 (pattern cV?V.Cvc)

Total = 192 words

The utterances were selected in such a way so that the prosodic variation (such as stress, tone, emphasis and vocal effort) and feature dependent segment duration do not have any effect on the target phoneme. Also, the manner of articulation was considered when these utterances were collected, as the manner of articulation is the usual first basis for segmentation or duration calculation. All listed words were phonetically defined if required, an assertion that was confirmed by linguists. It is proclaim that this corpus has 100% phoneme coverage.

Diphone corpus: According to [12] and [13] Bangla language has 30 consonants and 35 vowels (monophthong, diphthong) phonemes. In general, the number of diphone in a language is the square of the number of phones. Since Bangla language consists of 65 phones, so the number of diphones are (65X65) 4225. In addition, silence to phones are (1X65) 65 and phones to silence are (65X1) 65. So the total number of diphones is 4335. These diphones were embedded with carrier sentences. Though there have been various techniques to embed diphone with carrier sentences, here nonsense words were used to form carrier sentences [14]. It has 100% coverage of phone and diphone

Continuous speech corpus: Language is evolving; everyday new words appear in newspapers, magazines and blogs, which have different spoken variety. So we decided to use the spoken variety of texts. Then, texts were collected from various domains as shown in table 1. The text corpus contains 1,06,860 tokens, ~10K sentences. Some text was encoded in ASCII which was later converted into Unicode using the CRBLP Converter [15]. Then, the spelling and conversion errors were manually corrected. Table 2 shows the token and sentence count of the three corpora.

Category	Tokens	Source
Megazine (weekly)	31296	1
Novel (Beji-Weasel)	30504	2
Legal document (Child)	1909	3
History (Dhaka, Bangladesh, Language movement, 7th March)	10795	4
Blog (interview)	2347	5
Novel (Rupaly Dip)	12160	6
Editorial (prothom-alo)	3963	7
Constitute of Bangladesh	3278	8
News -Prothom alo	10608	9
Total	106860	

Table 1: Different domains of the corpus

Name	Tokens	Token type	Sentences
Corpus for acoustic analysis	602	203	262
Diphone corpus	12,938	2318	4,335
Continuous speech corpus	1,06,860	17,797	10,895

Table 2: Token and sentence count of the three corpora.

3.2. Speaker selection

Corpus for acoustic analysis: Professional and non-professional male and female speakers were selected by considering different ages, heights and the speakers' locality in Bangladesh. Unfortunately, we were unable to include any speaker from the Indian State of West Bengal in this analysis. Four male and four female speakers, with equal numbers of professional vs. non-professional male speakers were selected. The professional speakers' ages ranged from 52 to 54 and non-professional speakers' ages ranged from 25 to 29. Each speaker was given flash cards containing the utterances, and was asked to record each utterance in straight tone/pitch level and without assigning any stress in a word. The education of all speakers is above bachelor degree.

Diphone and Continuous speech corpus: A professional voice talent of a 29 years old male native Bengali speaker was hired for recording.

3.3. Recording

The recording of the utterances was done using the Nundo speech processing software. A professional voice recording studio was chosen to record the utterances. The equipment consisted of an integrated Tascam TM-D4000 Digital-Mixer, a high fidelity noise free Audiotechnica microphone and two high quality multimedia speaker systems. The voice talents were asked to keep a distance of 10-12 inches from the microphone. Optionally a pop filter was used between the speaker and the microphone to reduce the force of air puffs from bilabial plosive and other strongly released stops. The speech data was digitized at a sample rate 44.1 kHz, sample width 24-bit resolution and stored as wave format. After each recording, the moderator checked for any misleading pronunciation during the recording, and if so, the affected utterances were re-recorded.

There were a few challenges in the recording. First, speakers were asked to keep the speaking style consistent. Second, speakers were supervised to keep the same tone in the recording. Since speaking styles varies in different sessions a monitoring were required to maintain the consistency. To keep the consistency of the speaking style, in addition to [3] the following specifications were maintained:

1. Recording were done in the same time slot in every session i.e 9.00 am to 1.00 pm.
2. A 5 minutes break was maintained after each 10 minutes recording.
3. Consistent volume of sound.
4. Normal intonation was maintained without any emotion.
5. Accurate pronunciation.
6. Pre-recorded voice with appropriate speaking style was used as a reference. In each session, speaker was asked to adjust his speaking style according to the reference voice.

4. CORPUS ANNOTATION AND ANALYSIS

4.1. Annotation

There were a few challenges in annotation. The "Corpus for acoustic analysis" and "diphone corpus" was pre-modified. There was no non-standard word (NSW) [16]. That is why no text-normalization was required in those corpora. However, the challenges came in "Continuous speech corpus". A text-normalization tool [17] was required to normalize the text. In case of ambiguous token, the accuracy of the tool is 87% which motivated us to perform a manual check. After the manual checking, phonetic transcription was done using CRBLP pronunciation lexicon [18]. The CRBLP pronunciation lexicon contains all the lexicon entries that are available in a continuous speech corpus. In phonetic form IPA was used. It [18] also proclaims the 100% accuracy of pronunciation form. Later, a script was used to split the corpus into sentences based on punctuation marks such as ?, । and !. Each sentence was assigned a sentence id with orthographic and phonetic form and the same id was used in wave file.

The un-cleaned recorded data was around 24 hours and it has a lot of repetition of the utterances. So in annotation, the recorded wave was cleaned manually using wavlab which tends to reduce the recorded data to 13 hours 32 minutes. Then, it is labeled (annotated) based on id using praat [19]. Praat provides a textgrid file which contains labels (in our case it is wave id) along with start and end time for each label. A separate praat script was written to split the whole wave into individual wave based on id with start and end time. As praat does not support unicode, so id is used instead of text in labeling. Fig 1, 2 and 3 shows the examples of orthographic and phonetic form of the corpora. Fig 4 shows the labeling using praat.

আমি কিছু পাই
ami ki.c^{h}u pai

Figure 1: Orthographic and phonetic of "Corpus for acoustic analysis"

টা ট কা
/ta tɔ ka /

Figure 2: Orthographic and phonetic of "Diphone corpus"

স্কুলটির দিকে দৃষ্টি দিন
/skul.tir d̪ike d̪riʃ.ti d̪in/

Figure 3: Orthographic and phonetic of "Continuous speech corpus"

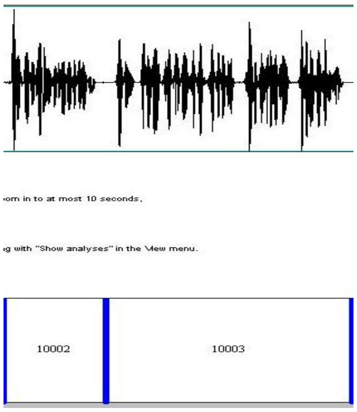
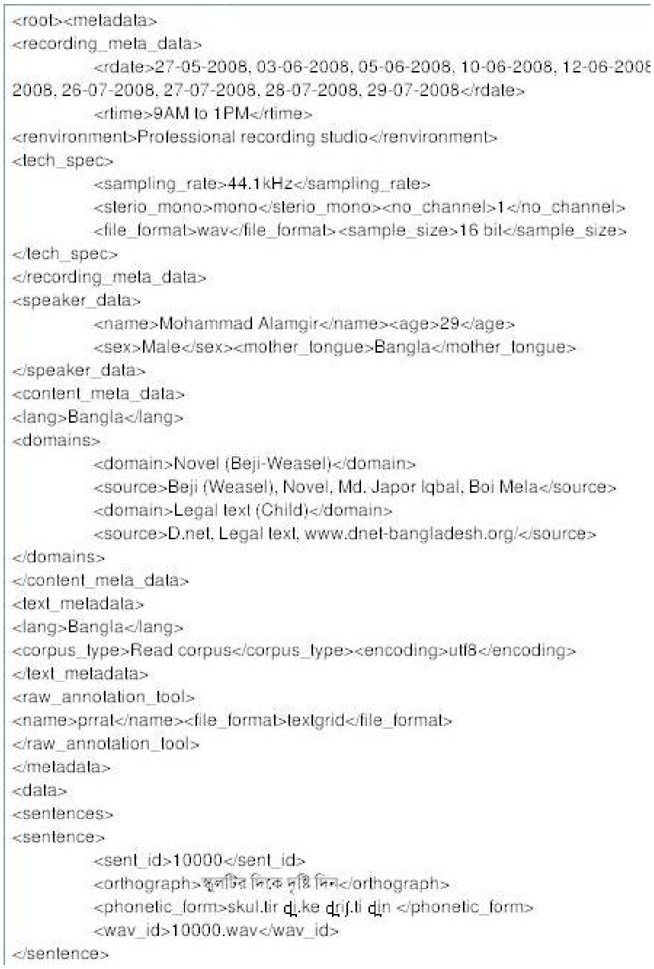

om in to at most 10 seconds.
g with "Show analyses" in the View menu.
10002
10003

Figure 4: Labeling using praat

4.2. Corpus structure

The structure of the corpus was constructed in a hierarchical organization using the XML standard. The file contains meta data followed by data. The metadata contains recording protocol, speaker profile, text, annotation and spoken content. The data contains sentences with id, orthographic form, phonetic form and wave id. The structure is elicited in figure 5.



```
<root><metadata>
<recording_meta_data>
            <rdate>27-05-2008, 03-06-2008, 05-06-2008, 10-06-2008, 12-06-200
2008, 26-07-2008, 27-07-2008, 28-07-2008, 29-07-2008</rdate>
            <rtime>9AM to 1PM</rtime>
<renvironment>Professional recording studio</renvironment>
<tech_spec>
            <sampling_rate>44.1kHz</sampling_rate>
            <sterio_mono>mono</sterio_mono><no_channel>1</no_channel>
            <file_format>wav</file_format><sample_size>16 bit</sample_size>
</tech_spec>
</recording_meta_data>
<speaker_data>
            <name>Mohammad Alamgir</name><age>29</age>
            <sex>Male</sex><mother_tongue>Bangla</mother_tongue>
</speaker_data>
<content_meta_data>
<lang>Bangla</lang>
<domains>
            <domain>Novel (Beji-Weasel)</domain>
            <source>Beji (Weasel), Novel, Md. Japor Iqbal, Boi Mela</source>
            <domain>Legal text (Child)</domain>
            <source>D.net, Legal text, www.dnet-bangladesh.org/</source>
</domains>
</content_meta_data>
<text_metadata>
<lang>Bangla</lang>
<corpus_type>Read corpus</corpus_type><encoding>utf8</encoding>
</text_metadata>
<raw_annotation_tool>
<name>prrat</name><file_format>textgrid</file_format>
</raw_annotation_tool>
</metadata>
<data>
<sentences>
<sentence>
            <sent_id>10000</sent_id>
            <orthograph>স্কুলটির দিকে দৃষ্টি দিন</orthograph>
            <phonetic_form>skul.tir d̪i.ke d̪riʃ.ti d̪in </phonetic_form>
            <wav_id>10000.wav</wav_id>
</sentence>
```

Figure 5: XML Structure of corpus

4.3. Analysis on "Continuous speech corpus"

A small analysis was done on "Continuous speech corpus" to evaluate the corpus. For this reason a statistical analysis has been conducted. Table 3 shows the phoneme, bi-phone and triphone coverage in the corpus. Fig 5 shows the frequency coverage of syllable, phone and biphone[1] in speech corpus. According to the table 3, this corpus lacks 4 phonemes. Besides that it has only 18.11% coverage of diphones and 5.93% coverage of the triphones. The phonemes are four diphthongs (aja, ua, ue, uo). To the best of our knowledge there is no published account about the frequency analysis of Bangla phoneme inventory & phonotactic constrain. This basically limits us to evaluate the phonetic coverage of this corpus. However, we can assert that this corpus has a domain variety. Moreover, this analysis raises a few research questions such as the following:

[1] The word biphone and diphone are used interchangeably.

1. Whether this 18.11% diphone coverage will cover all phonetic space in Bangla or not.
2. Does it make any problem if we omit the diphthongs that are not available in this corpus when designing phonetically balanced corpus?
3. Analysis of biphone set that should not belong to the language e.g. phonotactic constrain.

Pattern type	Possible (unique)	Total in the corpus	Coverage
phone	65	61	93.84%
biphone	4,225	765	18.11%
triphone	2,74,625	16,301	5.93%

Table 3: Phone and biphone coverage in "continuous speech corpus".

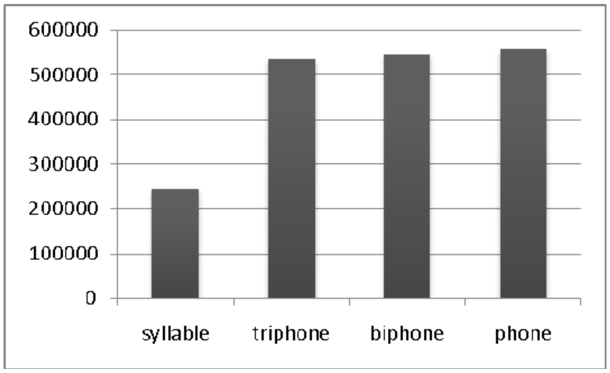


Fig 5: Frequency analysis

S/N	Syllable pattern	Frequency	Percentage
1.	cv	144668	59.0219
2.	cvc	67156	27.39842
3.	v	15460	6.307398
4.	vc	7592	3.097397
5.	ccv	4882	1.991767
6.	cvcc	2395	0.977116
7.	ccvc	2202	0.898376
8.	cvv	344	0.140346
9.	vcc	234	0.095468
10.	ccvcc	62	0.025295
11.	cccv	47	0.019175
12.	cccvc	25	0.0102
13.	cvccc	18	0.007344
14.	vv	18	0.007344
15.	ccvccc	4	0.001632
16.	cvvc	1	0.000408
17.	vccc	1	0.000408

Table 4: Syllable pattern and their frequency in the corpus

Table 4 shows different syllable patterns with their frequency available in the corpus. It is observed that, among these patterns some of the patterns formed from loan words. For example the patterns ccvcc, cvccc , cccvc and ccvccc are appeared in English loan words. Table 5, 6 and 7 shows a fragment of frequency analysis of phone, biphone and triphone.

Phone	Frequency	Percentage	Phone	Frequency	Percentage
a	59907	10.755431	ei	1897	0.340579
e	49200	8.833145	p^h	1671	0.300004
o	48754	8.753072	d	1333	0.239321
r	44180	7.931877	ai	1276	0.229087
i	36337	6.523780	t^h	1255	0.225317
n	32209	5.782658	oi	1094	0.196412
k	26974	4.842790	ã	1035	0.185819
ɔ	25885	4.647276	g^h	888	0.159427
t̪	22855	4.103283	$ɟ^h$	501	0.089947
ʃ	22126	3.972402	ui	484	0.086895
b	20154	3.618358	d^h	371	0.066608
l	16126	2.895189	ou	328	0.058888
m	14514	2.605778	iu	232	0.041652
u	12559	2.254786	eo	223	0.040036
d̪	12490	2.242398	õ	210	0.037702
p	12119	2.175790	ũ	194	0.034830
ɟ	11416	2.049577	ẽ	163	0.029264
j	9860	1.770220	eu	152	0.027289
t	8068	1.448492	ɔo	146	0.026212
h	7985	1.433591	ĩ	86	0.015440
g	6647	1.193372	ɔ̃	77	0.013824
c^h	6466	1.160876	au	55	0.009874
k^h	5467	0.981520	oa	39	0.007002
c	5254	0.943279	æ̃	38	0.006822
æ	5043	0.905397	io	31	0.005566
s	4246	0.762308	ie	11	0.001975
$t̪^h$	4054	0.727837	oe	6	0.001077
b^h	3508	0.629810	ia	2	0.000359
$d̪^h$	3389	0.608446	ea	1	0.000180
ŋ	3315	0.595160	æa	1	0.000180
ɾ	2086	0.374511			

Table 5: Frequency analysis of phoneme in the corpus

Biphone	Frequency	Percentage
a_r	11327	2.07416984
o_n	8205	1.50247758
e_r	7936	1.45321902
o_r	6518	1.19355867
a_n	5979	1.09485843
r_a	5625	1.0300349
r_o	5511	1.00915953
r_e	5264	0.96392955
ʃ_ɔ	5122	0.93792689
t̪_o	5025	0.92016451
n_i	4964	0.90899436
ɔ_r	4805	0.8798787
k_o	4781	0.87548389
n_a	4631	0.84801629
r_i	4591	0.8406916
k_a	4286	0.78484082
t̪_a	4241	0.77660054
a_k	4173	0.76414856
e_n	4170	0.76359921
n_e	3988	0.73027186

Table 6: Frequency analysis of biphone in the corpus

Triphone	Frequency	Percentage
k_o_r	3142	0.58706477
o_r_e	2049	0.382843957
p_r_o	1694	0.316514233
k_a_r	1687	0.315206323
b_o_l	1635	0.30549042
i_j_e	1564	0.292224475
d̪_e_r	1524	0.284750703
k_ɔ_r	1426	0.266439962
a_d̪_e	1386	0.258966191
t̪_a_r	1359	0.253921395
e_c^h_e	1346	0.251492419
o_r_i	1339	0.250184509
n_e_r	1313	0.245326557
a_r_e	1243	0.232247457
o_n_e	1236	0.230939547

Table 7: Frequency analysis of triphone in the corpus

5. CONCLUSION AND FUTURE REMARKS

Here we described the development procedure of Bangla annotated read speech corpora and some statistics of analysis. Corpus building is a continuous process which includes annotation for prosody prediction and annotation in different levels such as word, syllable, biphone and phone level for phonetic research. This is not only required for phonetic research but also in speech applications. Future work includes the following:

5.1. Intonation model and unit selection voice

As mentioned earlier there is an existing Bangla speech synthesis system, which lacks the intonation model. There were two reasons to develop "Continuous speech corpus". One concern was to develop an intonation model. The other reason was to develop a unit selection based speech synthesis system. Though there is no unique approach to design intonation, but the following two approaches consider a set standard to produce intonation for synthetic speech from corpus. One is to produce synthetic speech by concatenating waveform directly which is called unit selection based synthesis. In this process no signal processing is required so it preserves the quality of the original signal. The other approach is to generate intonation from ToBI label which could be generated from the corpus. Our research team is working on both unit selection based synthesis and concentrating on generating ToBI label from corpus.

5.2. Acoustic analysis

An extensive acoustic analysis [12], [13] has done on Bangla phonemes using "corpus for acoustic analysis". However, a significant amount of work need to be done on prosody such as syllable, stress, F0 and accent. So this type of clean and high quality speech corpus will help in acoustic analysis and speech applications.

5.3. Speech technology applications

The multiple uses of these corpora are innumerable. Generating a phonetically balanced corpus is another step which could be done from these corpora. A phonetically balanced corpus is especially important for speech synthesis and speech recognition. Besides these applications, one can use this resource to do research on speaker identification, emotion extraction and spoken information extraction.

6. ACKNOWLEDGMENTS

This work has been supported in part by the PAN Localization Project (www.panl10n.net), grant from the International Development Research Center (IDRC), Ottawa, Canada, administrated through Center for Research in Urdu Language Processing (CRULP), National

University of Computer and Emerging Sciences, Pakistan. We would also like to thank Dr Sarmad Hussain (NUCES), Naira Khan (Dhaka University) and BRAC University students who helped by providing their speech.

7. REFERENCES

[1] CDAC Bangla Speech Corpora, http://www.cdackolkata.in/html/txttospeeh/corpora/corpora_main/MainB.html, last accessed: December 10, 2009.

[2] Bangla Text to Speech Synthesis, http://crblp.bracu.ac.bd/demo/tts, last accessed: December 10, 2009.

[3] W. Zhu, W. Zhang, Q. Shi, F. Chen, "Corpus Building for Data-Driven TTS System," IEEE TTS Workshop, Santa Monica, 2002.

[4] Williams, Briony, "Levels of Annotation for a Welsh Speech Database for Phonetic Research", ISCA SIG SALTMIL, May-1998.

[5] D. Gibbon, R. Moore and R. Winski (eds.), Handbook of Standards and Resources for Spoken Language System. Berlin: Mouton de Gruyter.

[6] A. Nagy, P. Pesti, G. Németh, T. Bőhm: Design Issues of a Corpus-Based Speech Synthesizer, Hungarian Journal on Communications, 2005/6. special issue, pp. 18-24, Budapest, Hungary, 2005

[7] Florian Schiel, Christoph Draxler, Angela Baumann, Tania Ellbogen, Alexander Steffen, The Production of Speech Corpora, Version 2.5 : June 1, 2004

[8] Sinsuke Sakai, Building Probabilistic Corpus-based Speech Synthesis Systems from the Blizzard Challenge 2006 Speech Databases, Academic Center for Computing and Media Studies, Kyoto University

[9] Kishore S. Prahallad, Alan W Black, Rohit Kumar, Rajeev Sangal, Experiments with Unit Selection Speech Databases for Indian Language, Proc. National Seminar on Language Technology Tools: Implementations of Telugu, Hyderabad , India.

[10] Bangla Academy, *Bangla Academy Byabaharik Bangla Abhidhan*, 6th reprint 2005. Bangla Academy, Dhaka, 1992.

[11] Pickett, J.M. *Acoustics of Speech Communication, The: Fundamentals, Speech Perception Theory, and Technology*, Allyn & Bacon, 1998.

[12] Acoustic Analysis of Bangla Consonants, Firoj Alam , S. M. Murtoza Habib and Professor Mumit Khan, Proc. Spoken Language Technologies for Under-resourced language (SLTU'08), Vietnam, May 5-7, 2008, page 108-113.

[13] Firoj Alam, S.M. Murtoza Habib, Mumit Khan, *Research Report on Acoustic Analysis of Bangla Vowel Inventory*, Center for Research on Bangla Language Processing, BRAC University, 2008.

[14] Festvox, http://www.festvox.org/, last accessed: December 10, 2009.

[15] CRBLPConverter, http://crblp.bracu.ac.bd/converter.php, last accessed December 26' 2009.

[16] Sproat R., Black A., Chen S., Kumar S., Ostendorf M., & Richards C, "Normalization of Non-Standard Words: WS'99 Final Report", CLSP Summer Workshop, Johns Hopkins University, 1999, Retrieved (June, 1, 2008). Available: www.clsp.jhu.edu/ws99/projects/normal

[17] Firoj Alam, S. M. Murtoza Habib and Mumit Khan, Text Normalization System for Bangla, Accepted for present on the Conference on Language and Technology 2009 (CLT09), NUCES, Lahore, Pakistan, January 22-24, 2009.

[18] CRBLP pronunciation lexicon, http://crblp.bracu.ac.bd/demo/PL/, last accessed December 26' 2009

[19] Praat. 2007. www.fon.hum.uva.nl/praat/. Version - 4.6.27.

8. APPENDIX

[1] Shaptahik, article, http://www.shaptahik.com/arch_index.php?arc_day_id=2

[2] Beji (Weasel), Novel, Md. Japor Iqbal, Boi Mela

[3] D.net, Legal text, www.dnet-bangladesh.org/

[4] Bangladesh history: http://bn.wikipedia.org/wiki/ঢাকা, http://muktijuddho.wikia.com/wiki/বীর_শ্রেষ্ঠ, http://bn.wikipedia.org/wiki/বাংলাদেশ, http://bn.wikipedia.org/wiki/ভাষা_আন্দোলন, "History of 7 March, 1971", http://susanta.wordpress.com/2008/03/07/আবহাওয়া-নির্মল/, Wikipedia, , Wikipedia, http://bn.wikipedia.org/wiki/

[5] Somewhereinblog, Interview, http://www.somewhereinblog.net/blog/omipialblog/28803654

[6] Rupaly Dip, Humaun Ahmed.

[7] Prothom-alo, Editorial, http://www.prothom-alo.com/archive/headlines_mcat.php?dt=2008-07-24&issue_id=993&mid=Mw==

[8] Govt. of Bangladesh, Constitution of Bangladesh, April, 2008

[9] News, CRBLP-Prothomalo corpus.

MO PIU MINORITY LANGUAGE: DATA BASE, FIRST STEPS AND FIRST EXPERIMENTS

Geneviève Caelen-Haumont[1], Brigitte Cortial[2], Christian Culas[3], Tran Tri Doi[4], Thom Dinh Hong[5], Xuyen Lê Thi[6]
Hung Phan Luong[4,7], Thanh Nguyen Ngoc[5], Emmanuel Pannier[8], Vanessa Roux[2],
Jean-Pierre Salmon[1], Alice Vittrant [2,9], Hoang Thi Vuong[5],Ly A Song[10]

[1]International Research Center MICA, Hanoi University of Technology - CNRS/UMI2954 - Grenoble INP, 1 Dai Co Viet str. Hanoi, Vietnam,
[2]Université de Provence, 29 avenue Robert Schuman 13621 AIX-EN-PROVENCE,
[3] Research Institute on Contemporary Southeast Asia (IRASEC – CNRS – MAEE), Bangkok, Thailand,
[4]University of Social Sciences and Humanities, 336 Nguyen Trai str, Hanoi, Vietnam,
[5]Department of Culture, Sport, and Tourism, Lao Cai Province, Vietnam,
[6]Université Paris 7, Paris, France,
[7]Institute of Linguistics, 9 Kim Ma Thuong, str., Hanoi, Vietnam,
[8] Institute of Sociology, Vietnam Academy of Social Sciences, Hanoi, Vietnam
[9]LACITO, Paris, France,
[10]Nam Thu Thuong (Mo Piu ethnic village)

ABSTRACT

This paper is a first contribution about the Mo Piu language and culture. This ethnic minority is settled in the mountains of the North Vietnam. This culture being not documented at all at the international level, its language is said ‘under-resourced’ in the point of view of the automatic processing.

After a cultural, social and economical presentation of this minority, the paper focusses on the results of the first field ground undertaken in june 2009, and especially on the data basis, and the first experiments on the Mo Piu speech (method and preliminary results). The study in progression is concerning the domain of human recognition of melodic segments in order to try to find out 1° if this language is tonal or not 2° and if so, what are the tonal units.

Index Terms— *Mo Piu, ethnic groups, under-resourced language, endangered language, data basis, prosody, tonal units.*

1. INTRODUCTION

Both in Vietnam and France, this project is based on the collaboration of linguists Vietnamese specialists of ethnic languages, French specialists of speech in the domains of phonetics, phonology, prosody, French and Vietnamese anthropologists, specialists of ethnic groups in Northern Vietnam, Province of Lao Cai, and computer scientists, working all together, which enhances knowledge, skillfulness, and outcomes quality.

Especially in conjunction with Christian Culas, we have laid in 2008 the foundation for a collaboration concerning an endangered ethnic group, the Mo Piu people in Vietnam, living in the mountains in the north, in an area protected from car passing and tourists. This ethnic group is of oral tradition only.

With this strong collaboration of researchers, the willingness of Vietnamese Government to develop the study of minorities languages, the ground of MICA technology in speech and pictures supplying a platform of computing tools and expertise, this project could be the first step to settle in Vietnam and South-Asia, a team (and at the same time a “base camp”) devoted to the study and preservation of endangered languages.

Apart from a humanitarian goal of making all documentation available to the ethnic minority, two other scientific objectives were defined. The first concerns a linguistic and ethno-linguistic study, conducted in conjunction between MICA, linguists specialists of asian languages both at the University of Social Sciences and Humanities, and at Université de Provence, France, and LACITO, the French anthropologists from IRASEC, and the Vietnamese ethnologists from the Department of Culture, Tourism and Sports of Lao Cai. The second objective matches the interests of MICA for the under-resourced languages. These last languages can benefit from the well-resourced languages of the computer resources and developments by the means of transfer and adaptation, particularly in the field of speech recognition and synthesis.

In this paper, we present the Au Co project focussed on the Mo Piu people, the method and the first experiments.

2. ABOUT THE MO PIU PEOPLE

2.1. General features

Though their exonym is *Green Hmong*, their endonym is *Mo Piu* (or *Mo Brieu* depending of graphy). Some of them told us that they can currently speak 7 or 8 languages and concerning the White Hmong they are speaking too, they assert that there is absolutely no familiarity with that one; but because they are probably in instable position about their own identity, we must take this assertion with caution. So some questions have to be settled: for example is "Mo" just a transcription from their language for "Hmong" (which is Vietnamese)? Are they a part of Miao linguistic family – divided in China in Hmong, Hmu, Hmau, Hmu and Qoxiong *Miao* subgroups – and according to Christian Culas hypothesis, a part of one specific group, the *Hmu/Hmou/ or Hmeu* ? [1][2][3][4]. All these questions need to be better investigated by linguistic and ethnographic inquiries.

Only investigations on phonetic, lexical syntactic, cultural and ethno-historical data could identify this ethnic group with more certainty. This population being not documented at all in the international bibliography about Vietnam, we hope nevertheless to find a link with some Chinese ethnic groups. So their origin is a problem due to a lack of knowledge. In fact, this small ethnic group is too small to be listed, and in these conditions either they are listed as Hmong, or they are beyond the census.

From 2003 Provincial Census, Mo Piu or “Green Hmong” in official designation, was 551 people; from June 2009 Census, there are 455 in 2 villages.

By now children speak Mo Piu at 30-50%, and parents 70-80%. Parents are speaking Vietnamese to their children. A primary school works in the village where only the Vietnamese language is spoken and taught. 25% of people have been educated in the village school. No one has graduated college. Moreover nobody can write the Mo Piu language.

The names of the two Mo Piu villages are Nam Tu Thuong and Nam Tu Ha. "Nam" means "river" in Tày language (demographically, the Tày, linguistic family Tay-Kaday, are the most important ethnic group in the large area), "Tu" is the name of the « stream », and "Thuong" means "spring up". So when some families continued their migration lower towards Nam Xe, this new village was called "Ha" meaning "low". In Nam Tu Ha village, several ethnic groups are living together, and the language must certainly bear marks of all these cultures.

As Mo Piu is not documented at all, it is urgent to study several aspects: 1- phonetics and phonology, 2- lexicon, 3- morphosyntax, 4- prosody, tonologic and subjective expression (not only emotional), 5- ethno-history, 6- cultural specificities.

This is needed for compiling Mo Piu text books and dictionaries. This also contributes to popularizing knowledge on Mo Piu people.

2.2. Location

In the Nam Xé commune, 161 families are gathering 890 people, belonging to four ethnic groups Hmông (including White Hmong and Mo Piu or Green Hmong), Dao, Tày and Kinh. Mo Piu are 51% of the population of this commune, but the 2 villages are not in its centre.

The village of Nam Tu Thuong is situated in a sort of circus on the side of a hill, scattered on the left bank of the stream. All the houses are wooden, and without piles as the architectural model of Hmong and Dao. To get from one house to another one, there is no real path, people has to climb boulders left in their natural state, and finds their way sometimes through steep rocks. An organized communication way between the houses is exceptional, wandering most of the time across blocks of stone more or less big, circumventing or avoiding them, climbing up or down the slopes and steep paths. As electricity fails again to the village (but the poles were raised, which presages a next use), the evening traffic from one house to another is even more dangerous for the unfamiliar.

In fact the village of Nam Tu Thuong is divided into several sites: from an older village, Nam Can, located a few miles further up the mountain, went 7 families in 1963, who then have created the present village, now composed of 11 families and 227 persons. The oldest village comprises 12 families. It would also be interesting to investigate also there, if it still remains people.

2.3. Recent history

Formerly, the village of Nam Tu Thuong was that of Red Dao ethnic group (endonym "Ké Mien") who lived there long ago. Then, the Mo Piu arrived in the village of Dao, they used to live and clear the ground. In fact 5 families have left China about 350 years ago, to settle in an area more fertile.

The 5 families have crossed the river Nam Thi “Lang Si” and then the Red River to enter Vietnam at Y Ty (Bat Xat district). Sometimes later, the climate being too hard, people moved to "Mang Pang", now named Khau Bang (Mu Cang Chai District, Yen Bai province), around 50km from the current location. Though we didn’t know exactly why, they could not stay there long, so they left and came to the Nam Xe commune territory (Nam Tu Thuong village, Van Ban District), and there they settled.

2.4. About economy

The area of the village is about 17 hectares, plus 84 ha for agriculture, 13 ha of forest, 126 ha reserved for the annual crop, 8 ha of rice fields from the cooperative, 900 ha for growing in a long-term. The village of Nam Tu Thuong has 27 ha of rice fields farmed privately.

The economy has grown. In the village, each family has a forest where they grow wild cardamom *(Elettaria cardamomum*) under the shelter of a big tree. Some families have up to 0,5 ha. This mountain spice can provide important income for the local farmers; in 2000, dry cardamom was sold in the Sapa wholesale market, 160.000 Vnd/kg, about 10$US/kg. Unquestionably this is the most expensive local product after the end of opium production [5] [4]. Apart from this case, government authorities help and encourage people to grow rice and maize, which thus supply a good productivity. The developing economy leads to increase the quality of life. Though the classification of “poor”, “very poor”, “medium poor” in Vietnam, especially in mountain ethnic area, is still the subject of many debates, according to the Mo Piu authorities, there are only 6 poor families or who do not eat their fill.

The Government invests to support Mo Piu ethnic groups, providing new roofs and water tanks for instance. In the village there are only 3 thatched houses. The machines shelling and husking paddy begin to appear.

3. DATA BASIS

On june 2009, we undertook a first field ground in the village. The team was composed of 2 Vietnamese ethnologists from the Departement of Cuture, Tourism and Sport from Lao Caï, and two scientists from MICA, a linguist and an engineer specialist of audio/video recordings and computer processing. It was a great chance for the team to benefit from such a specialist because linguist and ethnologists could be better involved in their task with the speaker, and thus be better concerned with the scientific aspects and contents of the recordings, while they also benefit from a data basis well structured.

3.1. Method of recording

Before recordings, the linguist with the help of the speaker and of the team, filled up an inquiry form containing the most important information about this speaker.

The audio and video recordings were made around a low table supporting the equipment and the microphones. To the linguist’s left side, stood a first ethnologist speaking English. Both were sharing the same microphone (track 1). Another ethnologist sat near her colleague. To the linguist right side, stood the Mo Piu speaker, then the translator who was translating oral question from the Vietnamese to Mo Piu language (then writing the responses Mo Piu / Vietnamese). The speaker and the translator both spoke in the same microphone (track 2). The engineer watching the monitoring settings was faced with all of us, with all the equipments, computers, camera, sound recorder. He was checking the good position of the microphones, of the camera vis-à-vis the faces, performing the zoom video, capturing recordings on the computer.

Each question (or songs, free speech) corresponds to one file. This facilitates the data distribution and also avoids losing too much time in case of wrong cancellation or system error. Furthermore it gives an instant overview of the richness of the topic.

At a signal from the engineer, in order to filling up the sound file header, the linguist gave the date, the speaker’s name, the theme addressed, the question number (track 1). The anthropologist immediately read the question translated in advance in Vietnamese (same track), which was immediately translated into Mo Piu (2nd track), and the speaker answered (same track). While the person was speaking, the translator wrote in Vietnamese what the person said. All the questions and answers translated into Vietnamese were grouped, then photocopied (later after our leaving, at our arrival in the nearest little town).

This whole methodology has been put in place quickly if not instantly. Once the explanation done on how we wanted to proceed, everyone has fully understood and fully played his/her role.

3.2. Data basis contents

We recorded on the whole 8 hours of films, gathering 1251 photos, 82 video-clips, 7 hours of speech, 1 hour of songs. In detail, the speech corpus length is 423 minutes, the songs or musical pieces, 59 minutes.

Topics	Speakers Number	Recording number	Duration (mn)	TOTAL recordings
1- History	2	5	26	
1-Tales	1	14	79	
1- Folk songs	8	20	59	
1- Music, instruments	2	20	22	
1- Folklore	1	1	15	
1-Past life	2	19	28	79
2- Birth	1	17	16	
2- wedding	2	30	31	
2- Funerals	1	4	15	
2- chamanism	1	28	29	79
3- Agriculture	1	8	7	
3- Animals care	1	2	4	
3- Hunting	1	10	9	
3- Fishing	2	20	17	
3- Tools	1	7	6	
3- Costumes	2	25	31	64
4- Social overview	1	12	9	
4- Village chief tasks	1	11	11	
4- Village rules	1	5	6	
4- Health problems	1	5	5	
4- Agriculture problems	1	8	10	
4- Language survival	1	1	2	42
5- Cooking	1	7	6	
5- Children care	2	34	35	41

Table 1. *Composition of the Data basis.*

The data basis is classified per date / speakers / topics / questions (or songs, tales…), and finally we get 321 sound files and 321 video files. We recorded for speech 4 male and 3 female speakers. The male speakers are between 36 and 66 years, the female ones, between 37 and 72 years. Six are farmers, and among them, one is the village chief, another is working in the commune. The seventh is a chaman.The age range of the 9 singers (7 females, 2 males) spreads from 24 to 70 years. All are farmers.

The Table 1 above presents the topics addressed, the number of speakers per topic, the number of the items recorded (responses, free speech, songs, tales), the total per topic, and the duration in mn of each topic. Grounded on the duration, the most important topics for speech are: tales (79 mn), children care (35 mn), wedding (31 mn), costumes (31 mn), chamanism (29 mn), past life (28 mn), history (26 mn).

4. THE AIMS OF THIS STUDY

As this language is not repertoried at all, this exceptional situation allows to undertake innovative studies both in the domain of speech technologies, and in linguistics.
In this paper, we just want to put forward the linguistic study, its challenge, and the method used.

4.1. The goals of the first study

As we got no cues about the structure of this language, we don't know whether it is tonal or not, and neither if the lexicon is mono- or plurisyllabic. In those conditions, we conduct a double study about tonality. Our goal is twofold:

1- to try to discover if the language is tonal or not,
2- if it is tonal, to try to identify the tonal units, and moreover to try to get rules to segment the language into lexical units.

In fact one more goal can be added to these ones: the need to build up a method and tools to find out the melodic units in the same time where we try to identify these units. No doubt that these goals are true challenges, but eventhough we fail in this tentative, we will get a deeper skill into the relation prosody / lexicon / syllable, that maybe we could not reach if we have been first studying the Mo Piu linguistic system.
In other terms, we are studying this language in the same conditions as computers may work, and thus, if this experience is successful, it could benefit for automatic processing in the domain of human language technologies. One of the applications concern the domain of the under-resourced languages called "PI languages" [6] in a multilingual processing approach aiming at rapidly developing spoken language technologies.

5. FIRST EXPERIMENTS

The acoustic data are analysed under specific speech software such as Praat [7] which is an international standard tool, in order to study formants and time events in the speech signal, MOMEL [8] which following the model of human perception, supplies a continuous intonation line even during the unvoiced speech events such as consonants, and finally an home tool MELISM [9] [10], specialized in the detailed analysis of prosody (F0, tones, duration) at the segmental level (word, syllable, tone, phonetic units), and which offers valuable complements to the previous ones for any kind of languages [11].

5.1. About the method used

As said before, our goal is to try to identify first the shapes of the melodic units (slopes direction, range…).This problem rests on the existence of the repetition of the melodic units invariants. Our experience of prosody and of tonal languages, make us expect the existence at the very least of:

1-kinds of shapes
- plateaux: /P/
- rising slopes: /M/
- falling slopes: /D/

In our perspective, the plateau is considered as such when the rising or falling slope doesn't exceed 25 Hz.
2- kinds of registers. 2 studies have been conducted, one exploring melody with 3 registers:
- high: /h/,
- middle: /m/,
- low: /b/,

and the second one, with 4 levels:
- acute: /a/
- high: /h/
- middle: /m/
- grave: /g/.

In fact, we have to be aware of not confusing the phonologic level where probably only a part of these levels is significative with the phonetic one which describes the structure of the plateau or the slopes on several layers (see below paragraph 5.3.2.).
3- combinations of units (probably 2, please see details paragraphe 4.3.), at least
- plateau + slope
- slope + plateau

4- number of tones (if any): probably less than 10.

If this language is tonal, we could expect to detect some of these units presented above. The regularity of these tonal units, fixed in a few number of shapes, their F0 stability, could prevent us to confuse them with word melody segments which in a no tonal language are far more variable. In fact our task consists in sorting the units in 2 categories: the phonetical and phonologic segments. As this analysis is concerning tones, instead of *phonetics* we could use more appropriately the term of *tonetics,* in opposition with *tonology*.
If the melodic segments are corresponding to the tonologic units, the tonetical ones embrace not only their contextual or speaker variations, but also, if any, the tonetic space between 2 following tones. In fact due to the MOMEL script conception, as explained above, the intonation curve is continuous, even during the unvoiced

segment. In these circumstances, we have also to try to extract the tonological parts from the curve continuum.

5.2. The two experiments

In order to improve objectivity in the detection of the tones (if any), we conduct the experiments in two different ways. The first way is based on the tentative to recognise these melodic segments only by the mean of the ear. This task is undertaken by a linguist musician, and the notation will be made according to the musical mode.

Besides, in order not to influence the final results, we decided not to compare our results before the end of our personnal experiment, except if one of us gives up. As for any challenge, the risk that both of us could dismiss this task was not to exclude. The second method is using ear and vision, receiving a help from specialised tools such as Praat, augmented with the MOMEL and MELISM scripts. Within this method, the intonation curve is segmented and labelled into micro melodic segments units, according to the significative turning points of the curve. Eventhough we cannot access to the word segmentation or meaning, of course we take advantages of all the cues events, such as pauses, breaks, slope ruptures, rhythm variations, occurrence of consonants, creaky voice, melodic symmetry…, but above all, of the repetition of the same phonetic vs. melodic sequences.

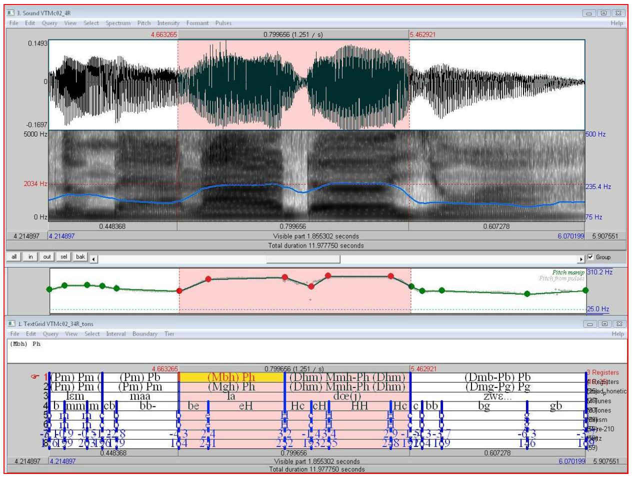


Figure 1. *Illustration of the visua/soundl metho using the MOMEL-MELISM scripts under Praat.*

Another point about the method used concerns the syllable considered as the unit of reference. Since the theory of Frame [12] giving a definition of the syllable at the physical, articulatory and segments organisation, many works at the international level pointed out its importance. Particularly the idea that the more complex consonants clusters occur at the beginning of the syllable instead of the coda, will induce our way of segmenting syllables. On the other hand, this theory is also tuned to the automatic speech recognition method used in MICA which considers the syllable as the unit of modelisation and segmentation. Nevertheless the melodic segmentation will concern the stable part of the vowel, i.e. the vowel nucleus, putting all the other parts of the syllable inside brackets.

In order also to better sustain this experiment, we add phonetic labels in a TextGrid tier. However as this first experiment does not stem on phonetics, we are just using a broad phonetic labellisation aiming simply to be able to compare the syllables with the same phonetic contents.

5.3. Illustration of the visual method (MOMEL-MELISM under Praat)

5.3.1. Symbols and codings

The figure 1 above presents a speech extract from the Mo Piu language (woman VTM, extract VTM02, spontaneous speech), taken after the automatical segmentation and labellisation under Praat / MOMEL / MELISM software[1].

3 windows are open simultaneously: from top to bottom, the sound one with the speech signal and spectrum, then just below, the manipulation one, providing the F0 curve with on one hand the points corresponding to the boundaries of units (candidates for tones and/or words) and sub-units, all manually positionned, and finally the MELISM window with several tiers: from bottom to top, in the frame of the hypothetised syllables,

- the F0 values in Hertz,
- then above their conversion in semi-tones,
- then their alphabetic coding in 9 levels (*g* = *g*rave, level *1* ; *i* = *i*nfra-grave, *2* ; *b* = *b*ottom, *3* ; *c* = *c*entred, *4* ; *m* = *m*iddle, *5* ; *e* = *e*levated, *6* ; *H* = *H*igh, *7* ; *S* = *S*upra-high, *8* ; *A* = *A*cute, *9*),
- then the F0 coding of the space between two boundaries (the so-called melodic syllables, for instance *eH*). In the frame of this present study, we don't use this level of melodic precision (neither the coding in 9 levels nor melodic syllables).
- then the broad phonetic labellisation,
- and finally corresponding to the 2 last tiers, the tentative caracterisation of the supposed tones (everyone put manually) considering for the F0 description 3 or 4 registers (see below 5.3.2.).

In this last tiers, one can read the symbols corresponding to the shapes of F0 segments we precised above (see § 5.1.), such as /P, D, M/ for respectively /plateau, falling slope, and rising slope/, and just afterwards the F0 level indicated in small letters: /b, m, h/ respectively for /low, middle, high/ in the case of 3 registers for the F0 range, and /a, h, m, g/ for /acute, high, middle, grave/ in the case of 4 registers.

In order to simplify the reading and the analysis, all the rising and falling slopes have been labelled with 3 cues:

[1] The overlap of some F0 values, and the absence of some codings in the tonal syllables (for instance H instead of Hm) are caused by the size of the zoom.

after the indication if the slope is rising (M) or falling (D), the second cue corresponds to the melodic level at the beginning of the slope, and then the third cue, to the level at the end of the slope. For example in the figure 1 above /Dhm/ means that the falling slope begins at the high level and reaches the middle one.

In the case of twofold codings, the sequence /Mmh-Ph/ means that the unit begins with a rising slope (M) in the middle register (m) and reaches the high level (h), then continues with a plateau (P) still in the high register (h). Now /Ph-Dhb/ would mean that the sequence starts with a plateau at the high level (Ph), followed by a falling slope beginning in the high level (h), ending in the low one (b).

Last point to clear out: the brackets, for instance /(Dmb)/. This notation corresponds in fact to our method of segmentation based on the syllable and the privilege given to the vowel nucleus. In the brackets, we put what we supposed to correspond to tonetics and not to tonology, such as F0 segments corresponding to consonants, not only the unvoiced ones, but also the voiced ones as they give some F0 modulations, and more generally F0 intervals between 2 melodic units.

The central part of the speech sample above on figure 1 (in blue or grey) is worth noticing, as the structure is symmetric, arousing the question of a right segmentation: for the syllable /dœj/ are they 2 tonologic items (/Mmh-Ph/~/Ph/), or less (if /Ph/ corresponds in fact to the phoneme /j/ and not to the coda of /œ/)?

5.3.2. F0 range: the number of registers in question

In the first phase of this work, we only used 3 registers of the F0 range: /b, m, h/ for respectively /low, middle, high/, wondering whether this partition of the F0 range could be convenient. Examining the data, we saw a clear disproportion of the nomber of plateaux: Pm (72 items), Ph (30), and Pb (17). We thought then that a partition between 4 registers (/g, m, h, a/ for /grave, middle, high and acute/) could be thus more realistic. The new results show that eventhough the distribution is better, the level /m/ is still gathering the most numerous items: Pg (9), Pm (50), Ph (39), Pa (21).

Anyway we have to remind that an adequate number for the tonetic description is not necessary adequate for the tonologic one, as the opposition of melodic height (same form, different F0 levels) could only concern very few tones, if they exist.

These 2 kinds of codings (3 and 4 registers) can be seen as tiers in the figure 1 above, and as columns in the figures 2 and 3 below.

5.3.3. Data bank

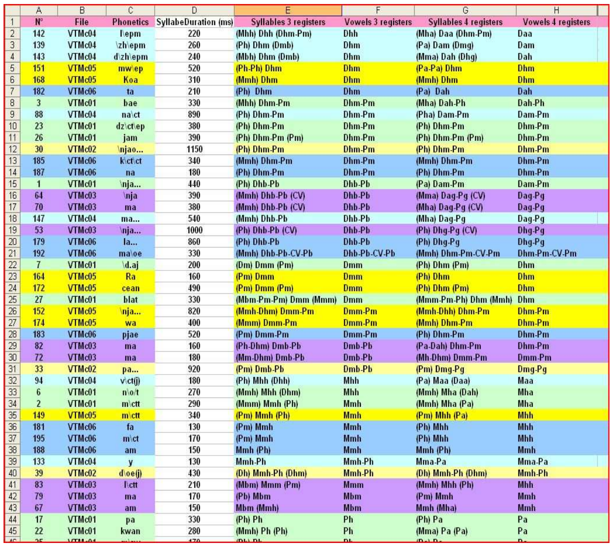
When this phase of segmenting and labelling is over, the MELISM procedure is automatically filling up a data bank containing the list of all the items with their melodic segments. This DB is a very performing tool, easily enabling to compare all the melodic shapes, F0 indices and duration, and then to observe at the best the repetition of the F0 segments and units.

The file automatically filled up by the MELISM software is divided in 3 parts. The first one contains the general F0 information about the sound file (F0 minimum, mean and maximum). The second one gives an overview of the list of the syllables (considered in this hypothesis as melodic units), with the indication of their F0 structure such as plateaux, rising and falling slopes (see above for more details). The third one presents the detail of each melodic segment composing the units. Then from each speaker's file of this kind, a big file is hand made gathering all the data, and enabling to make statistic computings.

Thus the data are automatically sorted in 2 ways. The first way sorts the syllables according to the vowel melodic coding (Figure 2 below). So all the syllables supplying the same melodic content are listed. For improving reading, the same files under excell use the same colour. The number of items supplies indeed relevant information about either the validity of the tone as a melodic unit or as a tonal one: the less numerous items, and the least confidence about the phonologic status.

In the Figure 2, a suite of columns supplies successively from the column 1, the number of the syllable in the list, the name of the file, the phonetic coding (here the symbols are not the IPA ones as these symbols are not known from excell, but the source codings), the syllabe duration in ms, the coding of 1- the whole syllable including the consonant and wovel (3 registers), 2- only the wowel (3 registers), 3- the whole syllable (4 registers), and 4- only the wowel (4 registers).

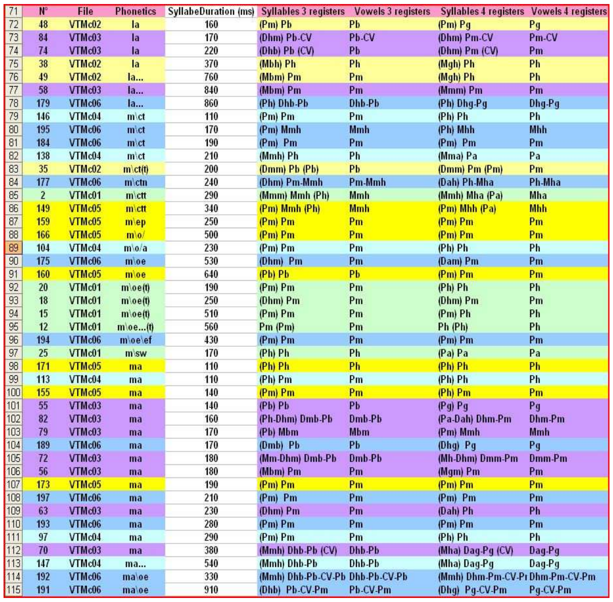
This kind of data asks the question of the identification of the melodic segment.



	A	B	C	D	E	F	G	H
1	N°	File	Phonetics	SyllabeDuration (ms)	Syllables 3 registers	Vowels 3 registers	Syllables 4 registers	Vowels 4 registers
2	142	VTMc04	l\epm	220	(Mhh) Dhh (Dhm-Pm)	Dhh	(Mha) Daa (Dhm-Pm)	Daa
3	139	VTMc04	\zh\epm	260	(Ph) Dhm (Dmb)	Dhm	(Pa) Dam (Dmg)	Dam
4	143	VTMc04	d\zh\epm	240	(Mbh) Dhm (Dmb)	Dhm	(Mma) Dah (Dhg)	Dah
5	151	VTMc05	mw\ep	520	(Ph-Ph) Dhm	Dhm	(Pa-Pa) Dhm	Dhm
6	168	VTMc05	Koa	310	(Mmh) Dhm	Dhm	(Mmh) Dhm	Dhm
7	182	VTMc06	ta	210	(Ph) Dhm	Dhm	(Pa) Dah	Dah
8	3	VTMc01	bae	330	(Mhh) Dhm-Pm	Dhm-Pm	(Mha) Dah-Ph	Dah-Ph
9	88	VTMc04	na\ct	890	(Ph) Dhm-Pm	Dhm-Pm	(Pha) Dam-Pm	Dam-Pm
10	23	VTMc01	dz\ct\ep	380	(Ph) Dhm-Pm	Dhm-Pm	(Ph) Dhm-Pm	Dhm-Pm
11	26	VTMc01	jam	390	(Ph) Dhm-Pm (Pm)	Dhm-Pm	(Ph) Dhm-Pm (Pm)	Dhm-Pm
12	30	VTMc02	\njao...	1150	(Ph) Dhm-Pm	Dhm-Pm	(Ph) Dhm-Pm	Dhm-Pm
13	185	VTMc06	k\ct\ct	340	(Mmh) Dhm-Pm	Dhm-Pm	(Mmh) Dhm-Pm	Dhm-Pm
14	187	VTMc06	na	180	(Ph) Dhm-Pm	Dhm-Pm	(Ph) Dhm-Pm	Dhm-Pm
15	1	VTMc01	\nja...	440	(Ph) Dhb-Pb	Dhb-Pb	(Pa) Dam-Pm	Dam-Pm
16	64	VTMc03	\nja	390	(Mmh) Dhb-Pb (CV)	Dhb-Pb	(Mma) Dag-Pg (CV)	Dag-Pg
17	70	VTMc03	ma	380	(Mmh) Dhb-Pb (CV)	Dhb-Pb	(Mha) Dag-Pg (CV)	Dag-Pg
18	147	VTMc04	ma...	540	(Mmh) Dhb-Pb	Dhb-Pb	(Mha) Dag-Pg	Dag-Pg
19	53	VTMc03	\nja...	1000	(Ph) Dhb-Pb (CV)	Dhb-Pb	(Ph) Dhg-Pg (CV)	Dhg-Pg
20	179	VTMc06	la...	860	(Ph) Dhb-Pb	Dhb-Pb	(Ph) Dhg-Pg	Dhg-Pg
21	192	VTMc06	ma\oe	330	(Mmh) Dhb-Pb-CV-Pb	Dhb-Pb-CV-Pb	(Mmh) Dhm-Pm-CV-Pm	Dhm-Pm-CV-Pm
22	7	VTMc01	\d.aj	200	(Dm) Dmm (Pm)	Dmm	(Ph) Dhm (Pm)	Dhm
23	164	VTMc05	Ra	160	(Pm) Dmm	Dmm	(Ph) Dhm	Dhm
24	172	VTMc05	cean	490	(Pm) Dmm (Pm)	Dmm	(Ph) Dhm (Pm)	Dhm
25	27	VTMc01	blat	330	(Mbm-Pm-Pm) Dmm (Mmm)	Dmm	(Mmm-Pm-Ph) Dhm (Mmh)	Dhm
26	152	VTMc05	\nja...	820	(Mmh-Dhm) Dmm-Pm	Dmm-Pm	(Mmh-Dhh) Dhm-Pm	Dhm-Pm
27	174	VTMc05	wa	400	(Mmm) Dmm-Pm	Dmm-Pm	(Mmh) Dhm-Pm	Dhm-Pm
28	183	VTMc06	pjae	520	(Pm) Dmm-Pm	Dmm-Pm	(Ph) Dhm-Pm	Dhm-Pm
29	82	VTMc03	ma	160	(Ph-Dhm) Dmb-Pb	Dmb-Pb	(Pa-Dah) Dhm-Pm	Dhm-Pm
30	72	VTMc03	ma	180	(Mm-Dhm) Dmb-Pb	Dmb-Pb	(Mh-Dhm) Dmm-Pm	Dmm-Pm
31	33	VTMc02	pa...	920	(Pm) Dmb-Pb	Dmb-Pb	(Pm) Dmg-Pg	Dmg-Pg
32	94	VTMc04	v\ct(j)	180	(Ph) Mhh (Dhh)	Mhh	(Pa) Maa (Daa)	Maa
33	6	VTMc01	n\o/t	270	(Mmh) Mhh (Dhm)	Mhh	(Mmh) Mha (Dah)	Mha
34	2	VTMc01	m\ctt	290	(Mmm) Mmh (Ph)	Mmh	(Mmh) Mha (Pa)	Mha
35	149	VTMc05	m\ctt	340	(Pm) Mmh (Ph)	Mmh	(Pm) Mhh (Pa)	Mhh
36	181	VTMc06	fa	130	(Pm) Mmh	Mmh	(Ph) Mhh	Mhh
37	195	VTMc06	m\ct	170	(Pm) Mmh	Mmh	(Ph) Mhh	Mhh
38	188	VTMc06	am	150	Mmh (Ph)	Mmh	Mmh (Ph)	Mmh
39	133	VTMc04	y	130	Mmh-Ph	Mmh-Ph	Mma-Pa	Mma-Pa
40	39	VTMc02	d\oe(j)	430	(Dh) Mmh-Ph (Dhm)	Mmh-Ph	(Dh) Mmh-Ph (Dhm)	Mmh-Ph
41	83	VTMc03	l\ctt	210	(Mbm) Mmm (Pm)	Mmm	(Mmh) Mhh (Ph)	Mhh
42	79	VTMc03	ma	170	(Pb) Mbm	Mbm	(Pm) Mmh	Mmh
43	67	VTMc03	am	150	Mbm (Mmh)	Mbm	Mmh (Mha)	Mmh
44	17	VTMc01	pa	330	(Ph) Ph	Ph	(Ph) Pa	Pa
45	22	VTMc01	kwan	280	(Mmh) Ph (Ph)	Ph	(Mma) Pa (Pa)	Pa

Figure 2. *Extract of the data bank sorting the items according to their identical melodic labellisation by the column F (3 registers), and then by the column H (4 registers).*

For instance Figure 2 above, the twofold segment /Dhb-Pb/ (line 20, item n° 179) and /Dhm-Pm/ (line 8, item n° 3) are they the same one, one of them being a phonetic variation of the other one? If not, 3 registers zones are they enough? This problem can be easily solved by considering the 4 registers labelling. This one actually makes a clear opposition of the melodic levels /Dhg-Pg/ and /Dah-Ph/, which plaides definitively for 2 distincts melodic forms.



71	N°	File	Phonetics	SyllabeDuration (ms)	Syllables 3 registers	Vowels 3 registers	Syllables 4 registers	Vowels 4 registers
72	48	VTMc02	la	160	(Pm) Pb	Pb	(Pm) Pg	Pg
73	84	VTMc03	la	170	(Dhm) Pb-CV	Pb-CV	(Dhm) Pm-CV	Pm-CV
74	74	VTMc03	la	220	(Dhb) Pb (CV)	Pb	(Dhm) Pm (CV)	Pm
75	38	VTMc02	la	370	(Mbh) Ph	Ph	(Mgh) Ph	Ph
76	49	VTMc02	la...	760	(Mbm) Pm	Pm	(Mgh) Ph	Ph
77	58	VTMc03	la...	840	(Mbm) Pm	Pm	(Mmm) Pm	Pm
78	179	VTMc06	la...	860	(Ph) Dhb-Pb	Dhb-Pb	(Ph) Dhg-Pg	Dhg-Pg
79	146	VTMc04	m\ct	110	(Pm) Pm	Pm	(Ph) Ph	Ph
80	195	VTMc06	m\ct	170	(Pm) Mmh	Mmh	(Ph) Mhh	Mhh
81	184	VTMc06	m\ct	190	(Pm) Pm	Pm	(Pm) Pm	Pm
82	138	VTMc04	m\ct	210	(Mmh) Ph	Ph	(Mma) Pa	Pa
83	35	VTMc02	m\ct(t)	200	(Dmm) Pb (Pb)	Pb	(Dmm) Pm (Pm)	Pm
84	177	VTMc06	m\ctn	240	(Dhm) Pm-Mmh	Pm-Mmh	(Dah) Ph-Mha	Ph-Mha
85	2	VTMc01	m\ctt	290	(Mmm) Mmh (Ph)	Mmh	(Mmh) Mha (Pa)	Mha
86	149	VTMc05	m\ctt	340	(Pm) Mmh (Ph)	Mmh	(Pm) Mhh (Pa)	Mhh
87	159	VTMc05	m\ep	250	(Pm) Pm	Pm	(Pm) Pm	Pm
88	166	VTMc05	m\o/	500	(Pm) Pm	Pm	(Pm) Pm	Pm
89	104	VTMc04	m\o/a	230	(Pm) Pm	Pm	(Ph) Ph	Ph
90	175	VTMc06	m\oe	530	(Dhm) Pm	Pm	(Dam) Pm	Pm
91	160	VTMc05	m\oe	640	(Pb) Pb	Pb	(Pm) Pm	Pm
92	20	VTMc01	m\oe(t)	190	(Pm) Pm	Pm	(Ph) Ph	Ph
93	18	VTMc01	m\oe(t)	250	(Dhm) Pm	Pm	(Dhm) Pm	Pm
94	15	VTMc01	m\oe(t)	510	(Pm) Pm	Pm	(Pm) Ph	Ph
95	12	VTMc01	m\oe...(t)	560	Pm (Pm)	Pm	Ph (Ph)	Ph
96	194	VTMc06	m\oe\ef	430	(Pm) Pm	Pm	(Pm) Pm	Pm
97	25	VTMc01	m\sw	170	(Ph) Ph	Ph	(Pa) Pa	Pa
98	171	VTMc05	ma	110	(Ph) Ph	Ph	(Ph) Ph	Ph
99	113	VTMc04	ma	110	(Ph) Pm	Pm	(Ph) Ph	Ph
100	155	VTMc05	ma	140	(Pm) Pm	Pm	(Ph) Pm	Pm
101	55	VTMc03	ma	140	(Pb) Pb	Pb	(Pg) Pg	Pg
102	82	VTMc03	ma	160	(Ph-Dhm) Dmb-Pb	Dmb-Pb	(Pa-Dah) Dhm-Pm	Dhm-Pm
103	79	VTMc03	ma	170	(Pb) Mbm	Mbm	(Pm) Mmh	Mmh
104	189	VTMc06	ma	170	(Dmb) Pb	Pb	(Dhg) Pg	Pg
105	72	VTMc03	ma	180	(Mm-Dhm) Dmb-Pb	Dmb-Pb	(Mh-Dhm) Dmm-Pm	Dmm-Pm
106	56	VTMc03	ma	180	(Mbm) Pm	Pm	(Mgm) Pm	Pm
107	173	VTMc05	ma	190	(Pm) Pm	Pm	(Pm) Pm	Pm
108	197	VTMc06	ma	210	(Pm) Pm	Pm	(Pm) Pm	Pm
109	63	VTMc03	ma	230	(Dhm) Pm	Pm	(Dah) Ph	Ph
110	193	VTMc06	ma	280	(Pm) Pm	Pm	(Pm) Pm	Pm
111	97	VTMc04	ma	290	(Pm) Pm	Pm	(Ph) Ph	Ph
112	70	VTMc03	ma	380	(Mmh) Dhb-Pb (CV)	Dhb-Pb	(Mha) Dag-Pg (CV)	Dag-Pg
113	147	VTMc04	ma...	540	(Mmh) Dhb-Pb	Dhb-Pb	(Mha) Dag-Pg	Dag-Pg
114	192	VTMc06	ma\oe	330	(Mmh) Dhb-Pb-CV-Pb	Dhb-Pb-CV-Pb	(Mmh) Dhm-Pm-CV-P	Dhm-Pm-CV-Pm
115	191	VTMc06	ma\oe	910	(Dhb) Pb-CV-Pm	Pb-CV-Pm	(Dhg) Pg-CV-Pm	Pg-CV-Pm

Figure 3. *Extract of the data bank sorting the items according to their identical broad phonetic labellisation.*

The second way is sorting the syllables according to their phonetic contents (Figure 3 below). This Table presents an extract of the data bank where the data have been sorted by the phonetic contents of the syllables (column 3 "Phonetics"), and their corresponding melodic labels (3 and 4 registers). This extract was choosen because of the several items with the same phonetic content, for instance /la/, /ma/…

If two or several syllables sharing the same phonetic contents, share also the same melodic coding, thus the codings have some chance to be validated. If not, it could mean either that perhaps the difference is due to the speaker variation (tonetics), or to a drawback in the number of the melodic ranges, or to an approximation of the melodic coding, or finally that there may exist 2 different syllables / words with 2 melodic segments units.

Of course each coding allows to cross the data sortings for a better understanding. For all the questions arising, the ressource is of course to increase the number of data, and also to consult the corresponding F0 values in question both under Praat-MELISM and the excell files where all the F0 values are supplied.

5.4. First results

In the overall, the Mo Piu intonation is a very melodic one, and often we wonders whether the person is speaking or singing. In our experience, it is the first language even heard of that kind. This impression comes from the presence of some special indices such as the use of big and rapid contrats both at the pitch level and at the duration one, even if they don't occur necessary at the sime time, and moreover the existence of sorts of motives or ritornellos based on symmetric notes.

In other respects, the samples analysed show the regular use of a lengthening before a pause, which may be a very long one. As the limits of the syllable are right now under discussion, in order to supply quantification in ms, we only chose in each case the duration of the lenghtened vowel before a pause. For the 24 samples observed, this vowel lenghtening spreads from 110 ms to 1000 ms, mean, 591 ms which is very long. Depending on a few examples, this result however is just of course a preliminary one.

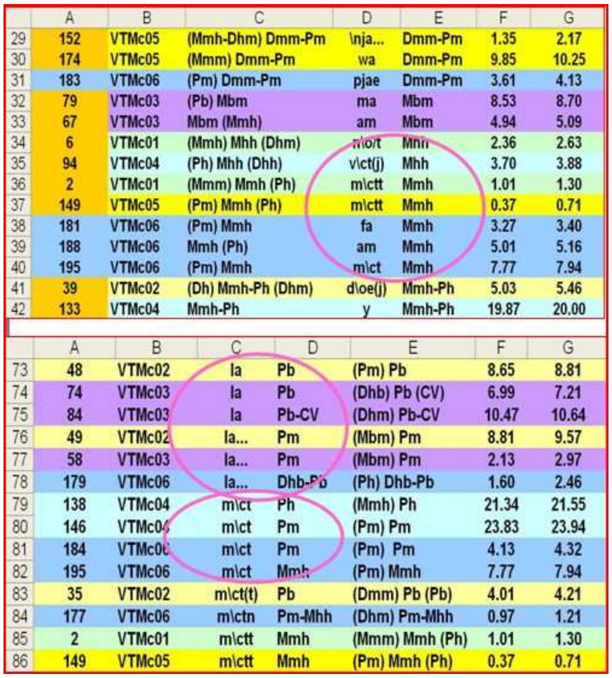
The second point concerns the question whether the Mo Piu language is mono- or plurisyllabic. An argument in favour of this thesis stems on the length of the mean duration of the syllables: 378 ms. This duration is a long one, corresponding in fact to the length of plurisyllables in other languages. Moreover this mean duration encompasses great variations (from 40 ms to 1380 ms) in the same way as the mean duration of the wovels lengthening before pauses. Another argument is also the abundance of plateaux, as it is difficult to imagine a lot of plurisyllabic words made of successive and long plateaux. Grounded on these findings, the hypothesis of a monosyllabic language, seems the most reliable.This issue joins the following point.

The third point is now concerning the melodic segments and their components. As said before, the first evidence is the great use of plateaux. Over the 173 syllables, 68% are corresponding to plateaux, and among them, 42% of the 173 syllables are middle plateaux /Pm/. Based on this regularity, and the great amount of plateaux at different levels, we finally incline to thinking that the Mo Piu language is *tonal*.

In fact, we find (see Figure 4 below) either different syllables corresponding to the same tone (Figure 4, see circles Table above) or different tones for the same syllable (Figure 4, see circles Table below), which greatly confirms that Mo Piu is a tonal language. Moreover the great variation in duration suggests that this parameter plays a significative role at the lexical meaning.

Besides 80% of the labelled vowels have a simple form. The fact that the complex forms are rare (20%) argues also for the role of duration (short ~ long) as a relevant parameter for meaning.

We are now considering the tonal content of the melodic forms (Table 2 below). In the Table 2 below, the tonal codings concerning the 173 vowels (syllable nucleus) have been reported, with put side by side, in the left columns, the codings corresponding to the 3 registers, and in the right columns, the 4 registers ones with the new distribution of the same data. In a grey shade, the items with very few occurrences being in fact the tonetic variations of the other ones, can be easily discarded. They are many.



	A	B	C	D	E	F	G
29	152	VTMc05	(Mmh-Dhm) Dmm-Pm	\nja...	Dmm-Pm	1.35	2.17
30	174	VTMc05	(Mmm) Dmm-Pm	wa	Dmm-Pm	9.85	10.25
31	183	VTMc06	(Pm) Dmm-Pm	pjae	Dmm-Pm	3.61	4.13
32	79	VTMc03	(Pb) Mbm	ma	Mbm	8.53	8.70
33	67	VTMc03	Mbm (Mmh)	am	Mbm	4.94	5.09
34	6	VTMc01	(Mmh) Mhh (Dhm)	m\o/t	Mhh	2.36	2.63
35	94	VTMc04	(Ph) Mhh (Dhh)	v\ct(j)	Mhh	3.70	3.88
36	2	VTMc01	(Mmm) Mmh (Ph)	m\ctt	Mmh	1.01	1.30
37	149	VTMc05	(Pm) Mmh (Ph)	m\ctt	Mmh	0.37	0.71
38	181	VTMc06	(Pm) Mmh	fa	Mmh	3.27	3.40
39	188	VTMc06	Mmh (Ph)	am	Mmh	5.01	5.16
40	195	VTMc06	(Pm) Mmh	m\ct	Mmh	7.77	7.94
41	39	VTMc02	(Dh) Mmh-Ph (Dhm)	d\oe(j)	Mmh-Ph	5.03	5.46
42	133	VTMc04	Mmh-Ph	y	Mmh-Ph	19.87	20.00

	A	B	C	D	E	F	G
73	48	VTMc02	la	Pb	(Pm) Pb	8.65	8.81
74	74	VTMc03	la	Pb	(Dhb) Pb (CV)	6.99	7.21
75	84	VTMc03	la	Pb-CV	(Dhm) Pb-CV	10.47	10.64
76	49	VTMc02	la...	Pm	(Mbm) Pm	8.81	9.57
77	58	VTMc03	la...	Pm	(Mbm) Pm	2.13	2.97
78	179	VTMc06	la...	Dhb-Pb	(Ph) Dhb-Pb	1.60	2.46
79	138	VTMc04	m\ct	Ph	(Mmh) Ph	21.34	21.55
80	146	VTMc04	m\ct	Pm	(Pm) Pm	23.83	23.94
81	184	VTMc06	m\ct	Pm	(Pm) Pm	4.13	4.32
82	195	VTMc06	m\ct	Mmh	(Pm) Mmh	7.77	7.94
83	35	VTMc02	m\ct(t)	Pb	(Dmm) Pb (Pb)	4.01	4.21
84	177	VTMc06	m\ctn	Pm-Mhh	(Dhm) Pm-Mhh	0.97	1.21
85	2	VTMc01	m\ctt	Mmh	(Mmm) Mmh (Ph)	1.01	1.30
86	149	VTMc05	m\ctt	Mmh	(Pm) Mmh (Ph)	0.37	0.71

Figure 4. *Extract of the data bank showing that a same melodic segment may correspond to different syllables, and conversely.*

The words being monosyllabic, we can now support the theory that the syllable segmentation corresponds in fact to the word segmentation (eventhough the precise limits are still under discussion). In this case, some word tones seem to be good candidates:

- simple forms
    - for the plateaux (P)
        - 3 registers: /Ph/, /Pm/, /Pb/,
        - or in the 4 registers version: /Pa/, /Ph/, /Pm/, /Pg/,
    - for the falling tones (M)
        - /Dhm/ both for 3 and 4 registers. A question remaining to explore is for instance for the 3 registers, whether /Dhm/ could be merged with /Dmm/.
    - for the rising tones
        - /Mmh/ for 3 registers,
- complex forms
    - /Dhm-Pm/ both for 3 and 4 registers, which lead to the conclusion that /Dhb-Pb/ could be a tonetic variation for /Dhm-Pm/ (3 registers) as /Dhm-Pm/ (4 registers) suggests it.

For the complex forms, we have also to explore whether this opposition of forms /Dhm-Pm/ ~ /Dhm/could not be resolved as well by a simple opposition in duration (for instance /Dhm-Pm/ resulting in fact to /Dhm/ long tone as opposed to /Dhm/ short one).

3 melodic registers		4 melodic registers	
Melodic segments	Population	Melodic segments	Population
Ph	30	Pa	21
Pm	72	Ph	39
Pb	17	Pm	50
		Pg	9
Dhh	1	Daa	1
Dhm	5	Dah	2
Dmm	4	Dam	1
		Dhm	6
Mhh	2	Maa	1
Mmh	5	Mha	2
Mmm	1	Mhh	4
Mbm	2	Mmh	3
Pb-CV-Pm	1	Pa-Dag	1
Pb-Mbm	1	Ph-Maa	1
Ph-Dhb	1	Pm-Mmm	1
Pm-Mhh	1	Pg-CV-Pm	1
Ph-Dhm-Pm	2	Pa-Dam-Pm	1
Pm-Dmb-Pb	2	Ph-Dhm-Pm	1
Pb-Mbm-Pm	1	Ph-Dmg-Pg	1
		Ph-Dhg-Pg	1
		Pg-Mgh-Ph	1
Pb-Mbm-Dmb-	1	Pm-Mbh-Dhm-	1
Pb-Mbh-Dhb-	1	Pm-Mma-Dam-	1
Pm-Mmh-Dhm-	1	Pg-Mgh-Dhg-	1
Dhb-Pb	7	Dah-Ph	1
Dhm-Pm	7	Dag-Pg	3
Dmm-Pm	3	Dam-Pm	2
Dmb-Pb	3	Dhm-Pm	10
		Dhg-Pg	2
		Dmm-Pm	1
		Dmg-Pg	1
Mmh-Ph	2	Mma-Pa	1
		Mmh-Ph	1
Total	173		173

Table 2. *Population of the tonal segments according to the 3 (left columns) and 4 registers codings (right columns).*

Grounded on the argument of the tonetic variations due to the melodic context or the speaker, a deeper analysis of the other melodic items with a smaller population has to be undertaken, in order to minimise the number of the different forms.

6. CONCLUSION

This paper focussed on a first presentation of a language and a culture which have never been studied before. This study is very attractive and tempting because we work under exceptional circonstances: we attempt to discover the intonational / tonal system of a language without having no prior knowledge of it. Our experience of the prosody and melody of various languages is our unic tool and safeguard.

We undertook this study carefully, step after step. For more security we based our investigation on the syllable considered as a stable reference unit. If the Mo Piu language is plurisyllabic, the remaining task consists in joining the syllables as simple bricks of the word unit. But we don't rely on this hypothesis because it is impossible that any language could supply successive words with several syllables, each of them being most of the time long duration plateaux. If conversely, this language is monosyllabic, as our arguments seem to prove it, the word segmentation is near to be effective (it just remains to clear out the right limits of the syllable).

Based on this first experiment, the analysis of the melodic segments inclines us to establish that the *Mo Piu language is tonal and monosyllabic.*Moreover the first results show also clearly the repetition of a few tonal candidates, corresponding to

- several levels of plateaux (low, middle and high, i.e. 3 registers: **/Ph/, /Pm/, /Pb/**, or in the 4 registers version, acute, high, middle and grave: **/Pa/, /Ph/, /Pm/, /Pg/**),
- a F0 falling slope simple (**/Dhm/**) or more complex in two parts (**/Dh-Pb/**) if this opposition doesn't finally stem on a simple opposition of duration,
- and for the F0 rising slopes, to the **/Mmh/** one.

Concerning the other candidates, less numerous, we have to study either they are simply tonetic variations to the other ones, due to the phonetic and melodic context, or due to the speaker, or true tones but naturally with less occurrences in this language.

At this state of the study, it still remains to find out how many tonal oppositions are relevant for the Mo Piu tonologic system, and what are their definitive shapes.

For the moment, it is important to pay attention to all the cues which can supply right information about human recognition, as it could provide interesting paths and unexpected milestones for the needs of the automatic processings.

7. REFERENCES

[1] C. Culas, "The ethnonyms of the Hmong in Vietnam: Short history (1856-1924) and practical epistemology", in C. Culas and F. Robinne (eds.), *Interethnic Dynamics in Asia. Ethnic Relationships through Ethnonyms,Territories and Rituals*. Routledge London, pp 13- 42, 2009.

[2] C. Culas, *Le Messianisme hmong XIXe et XXe siècle. La dynamique religieuse comme instrument politique*. Paris, Éditions du CNRS, 380 p., 2005.

[3] J. Lemoine, "Les ethnies Non Han de la Chine. Les Miao-Yao", in Poirier J. (ed.), *Encyclopédie de la Pléiade.* Ethnologie Régionale II. Paris, Gallimard, pp. 731-922, 1978.

[4] Ph. Klein, *La Saga des Miao/Hmong*, Tome 1, à paraître.

[5] C. Culas, "Study of discourses on local knowledge and practices on environment management in Vietnam mountains: An anthropological perspective", *Modernity and Dynamics of Tradition in Vietnam: Anthropological Approaches*, Social Sciences Edition, 21 p. [in Vietnamese and English], forthcoming.

[6] V. Berment, "Méthodes pour informatiser des langues et des groupes de langues peu dotées", *Thèse de doctorat*, Université Joseph Fourier, Grenoble 1, 2004.

[7] P. Boersma, D. Weenink, 2008 Praat: doing phonetics by computer, <http://www.praat.org/>.

[8] D. J. Hirst, R. Espesser, "Automatic labelling of fundamental frequency using a quadratic spline function", *Travaux de l'Institut de Phonétique d'Aix, 15,* 71-85, 1993.

[9] G. Caelen-Haumont, C. Auran, "The Phonology of Melodic Prominence: the Structure of Melisms", *Proceedings of Speech Prosody 2004*, Nara, Japan, 143-146, 2004.

[10] G. Caelen-Haumont, Manuel d'utilisation de la procédure MOMEL-MELISM sous Praat, vol 1 (p. 1-28) et 2 (p. 29-57), version 2 <http://www.lpl.aix.fr /~lpldev/MELISM/>, 2008.

[11] G. Caelen-Haumont, "Emotion, emotions and prosodic structure : an analysis of the melisms patterns and statistical results in the spontaneous discourse of 4 female speakers from four generations", book chapter, in Sylvie Hancil éd., *The Role of Prosody in Affective Speech* , Peter Lang, li97, 95-138.

[12] P. F. MacNeilage "The Frame/Content theory of evolution of speech production", *Behavioral and brain sciences,* Vol. 21, pp. 499-546, 1998.

ADAPTATION TECHNIQUES FOR SPEECH SYNTHESIS IN UNDER-RESOURCED LANGUAGES

Gopala Krishna Anumanchipalli, Alan W Black

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213
{gopalakr,awb}@cs.cmu.edu

ABSTRACT

This paper presents techniques for building speech synthesizers targeted at limited data scenarios - limited data from a target speaker; limited or no data in a target language. A resource sharing strategy within speakers and languages is presented giving promising directions for under-resourced languages. Our results show the importance of the amount of training data, the selection of languages and the mappings across languages in a multilingual setting. The objective evaluations conclusively prove that the presented adaptation techniques are well suited for building voices in resource-scarce conditions.

Index Terms: Speech Synthesis, Adaptation, Voice conversion, under-resourced languages.

1. INTRODUCTION

In today's digital age, there is an increasing use and acceptance of text-to-speech(TTS) technologies in the internet, mobile phones and dialogue systems. Besides, the use of speech as an output modality also enables information access for low-literate and visually impaired users. There is a compelling case for the development of speech synthesis technology in possibly all languages of the world. However, most languages have little or no resources required for building synthesis systems. Even for languages rich in speech and language resources, there is a need for efficient strategies for user-customization. Eliciting limited data ($<$ 2 mins) from the subject should sufficiently allow adaptation of an existing synthesizer to his voice. In this paper, we address both these situations as resource-scarce scenarios for bilding acceptable quality speech synthesizers.

While there is no definite notion of the minimum amount of resources required for training, availability of at least one hour of clean speech recordings is the norm for building high-quality functional speech synthesizers. This is in addition to phonetic and linguistic knowledge that requires annotated text resources in the language. This can be expensive and non-trivial for most languages. Many languages still have limited or no resources required to build text-to-speech systems. This makes building synthesis systems challenging using existing techniques. While, unit selection [10] continues to be the underlying technique in most commercial systems, its requirement of a large amount of well recorded and labeled speech data to ensure optimal unit coverage makes it prohibitive for under-resource situations. Statistical parametric synthesis [16], on the other hand is more liberal in its requirements, produces a more flexible voice comparable in quality to unit selection synthesis. Hence it is ideal for building voices in resource-scarce conditions.

Section 2 briefly describes our statistical parametric speech synthesis framework. A description of the resources required for building parametric voices follows in Section 3 including strategies for building voices under the resource-scarce conditions. Experiments and results are presented in Section 5.

2. STATISTICAL PARAMETRIC SPEECH SYNTHESIS

We use Clustergen [6], a statistical parametric framework within the Festvox [13] voice building suite. Fig. 1 shows a schematic representation of the training and testing phases in Clustergen. In the training phase, source and excitation parameters of the speech are extracted. Text-normalization and letter-to-sound(LTS) rules are applied on the transcription. The speech and phonetic transcriptions are automatically segmented using Hidden Markov Model (HMM) labeling. The speech features are then clustered using available phonetic and linguistic knowledge at a phoneme state level. Trees for duration, spectral (e.g. MFCC) and source (e.g. F0) features are built during the training phase. During testing (i.e. Text-to-Speech) input text is processed to form phonetic strings. These strings, along with the trained models are used to generate the feature parameters which are vocoded into a speech waveform by a synthesis filter (e.g. MLSA for MFCCs).

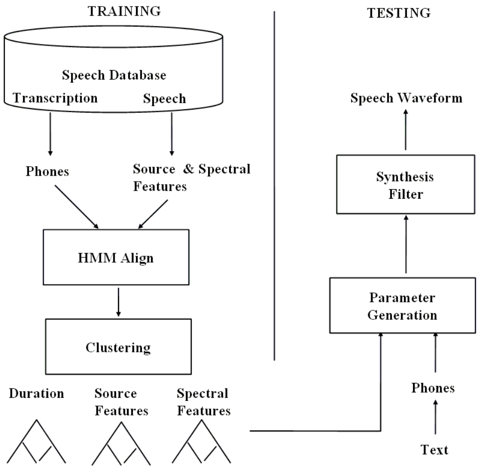


Fig. 1. Schematic diagram of the Clustergen framework

In this framework, models are stored as Classification And Regression Trees (CART) of the phone state. Each phone is realized as a left-to-right Markov chain of three states (roughly corresponding to the initial, middle and final states of a phone). The intermediate nodes of the tree are questions about phonetic and other high levels of contextual information (e.g., parts of speech). At the leaf nodes of the tree are the Gaussian codebooks corresponding to the feature instances falling in that path of the tree. The parametric representation (multi-dimensional Gaussians, in this case) makes transformations feasible via simple matrix algebraic operations. This flexibility of parametric models makes them well suited for adaptations required in under-resource conditions. Although this framework is similar to HTS [1], Clustergen generates the utterance frame by frame, rather than by state, allowing more detailed modeling.

3. BOOSTING RESOURCES FOR VOICE BUILDING

In this section, the resources required for building a voice are described, The specific alternatives for dealing with each kind of resource scarcity—that of limited target speaker data and target language data are presented in subsections 3.1 and 3.2 respectively. According to [11], the issues that need to be addressed while building a voice for a new language are 1) Definition of a phoneme set, 2) Creation of a lexicon and/or Letter-to-Sound(LTS) rules, 3) Text analysis, 4) Building prosodic models and 5) Building a waveform synthesizer.

For languages that have an IPA, SAMPA or a phoneset defined on another standard, they may be adequate to produce synthesizers of acceptable quality. However, for languages that have no established phonesets, it takes a few expert hours to design one based on the acoustic phonetic information of the phonemes in the language. For languages that are fairly phonetic (high grapheme-to-phoneme correspondence), grapheme-based phonesets have been shown to be adequate[17]. It should be noted that there is a certain arbitrariness in the allophonic variations within a language or even among speakers and there is no one best phoneset, optimal for all voices. Similarly, construction of a lexicon and LTS rules is non-trivial and the effort varies across languages, but a rule-based or a data-driven, statistical model for LTS has become commonplace for synthesizers in most languages [18]. In the following sections, the issues with limited amount of speech data are presented.

3.1. Limited data from a target speaker

As mentioned earlier, building a voice for a speaker requires a good amount clean recorded speech. It is thus desirable to have techniques that can work with just a few minutes of speech and produce good quality output. Recalling from Section 2, building a voice implies constructing decision trees for duration, source and spectral features. When the data is limited, phone coverage and contextual converge are both insufficient. This hurts any automatic technique to label the data. Even the estimated parameters (Gaussian means and variances) tend to be unreliable.

To compensate for this, data from one or more speakers may be used to build the 'source model' upon which the adaptation technique can impose the target speaker's identity.

This problem is studied extensively as 'model adaptation' proposed for speech recognition, starting with the work of [19], later also successfully applied for speech synthesis [20]. The selection of the source speakers on which to adapt may also be improved. Techniques involving speaker clustering and cohort selection have previously shown significant gains. There is also related work in voice transformation and feature space transforms [4] that deal with limited target speaker data.

3.2. Limited data in a target language

Lack of sufficient speech data for building speech systems is a common problem for most minority languages of the world. The GlobalPhone [8] project addresses this problem for speech recognition by exploiting existing resources in several languages to create a new language synthesizer. Similar attempts in speech synthesis [2] [14] also succeeded in creating a new language synthesizer sharing resources from several languages. This process is briefly described in the next section.

3.2.1. Multilingual Speech Synthesis

The 'source' voice in case of a target language adaptation is a multilingual voice. The training data for such a voice is speech included from several languages and the processed transcriptions in the respective languages. Since the phonetic properties (and labels) of the languages could be different, a global phoneset is created for the multilingual voice which assigns the same phonetic category to phonemes of different languages with the same acoustic phonetic properties. This strategy optimally shares the speech data across languages wherever appropriate. This also helps 'boost' the phonetic coverage of each language. However, this process requires carefully developed phone mappings between languages. The voice is built in a similar way as a monolingual voice after the mapping.

For the target language, the phoneset is mapped to that of the global set of the multilingual voice. The adaptation follows the same strategy as in a monolingual case transforming only the phonemes appropriate to the data presented for the target language. As shown in our results, the choice of the languages included in the training, and the amount of data in each language also affects the quality of the voice in a target language.

4. EVALUATION OF VOICES

We use Mel-Cepstral Distortion (MCD), a spectral distance measure proposed for evaluating voice conversion performance. It is given by the equation

$$MCD = 10/ln10\sqrt{2\sum_{d=1}^{24}(mc_d^{(t)} - mc_d^{(e)})^2} \quad (1)$$

where $mc_d^{(t)}$ and $mc_d^{(e)}$ are the target and the estimated spectral vectors respectively. MCD is known to correlate well with the quality of a voice [12]. The significance of MCD is quantitatively shown as a function of the training data size. A reduction of 0.12 MCD is shown as being equivalent to doubling the amount of training data used for building the voice. This is shown to be consistent across speakers and languages. The MCD measure is hence relevant both in the limited target speaker and limited new language data in this work.

5. EXPERIMENTS AND RESULTS

In this section, we report our observations of the adaptation techniques in each limited data situation. In all experiments, 50 dimensional Mel-Frequency Cepstral Coefficients (static + delta features) are used as the spectral representation. The features are clustered using phonetic and contextual questions. For growing the CART trees thresholded with a stop value of 50 instances at the leaf node. All adaptations are done only on the spectral features. A simple z-score mapping is done for the fundamental frequency to adjust to the dynamic range of the target speaker.

5.1. Limited target speaker data

To evaluate the limited target speaker data scenario, we use varying amounts of adaptation data of an American male speaker taken from the arctic database [7]. As the source model, we use 41 American English speakers of the Wall Street Journal speech corpus [15]. An ‘average’ voice is built from 3 hours of speech data sampled evenly across 41 speakers. It is shown that such an average voice is closer to an arbitrary new speaker since it has the average characteristics of all training speakers, and tends to be speaker independent.

We report two experiments of voice adaptation, one model based, MLLR adaptation [19] and the other feature based using Joint density GMM-based estimation (GMM-JDE) [3]. Since the target data is limited, adaptation is done only on the Gaussian means and the original variances are retained.

Figure 2 shows the MCD of the estimated speech with respect to the reference data as a function of the amount of data used for adaptation. It can be seen that even with 20 utterances there is a significant improvement in the voice and it is closer to the target speaker. The two techniques begin almost giving same improvements, and begin to converge with increasing adaptation data. The GMM-JDE technique converges more quickly. MLLR outperforms the GMM-JDE technique when more adaptation data is presented. This shows that of the two techniques, MLLR exploits data more effectively for this task.

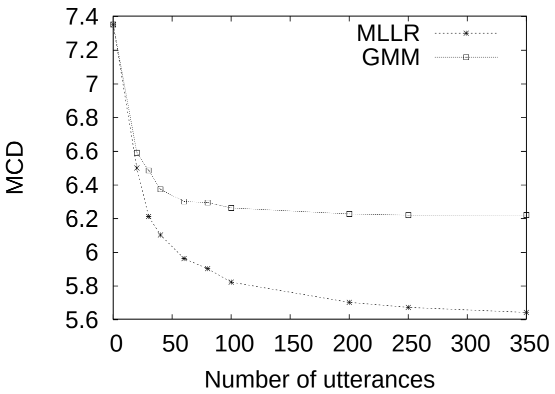


Fig. 2. Performance with increasing size of adaptation data from target speaker

5.2. Limited new language data

For simulating a limited new language data condition, a subset of the Globalphone database is selected. This subset consisted of 10 female speakers, one from each of Chinese (ZH), English (EN), German (DE), Japanese (JP), Croatian (KR), Portuguese (PT), Russian (RU), Spanish (ES), Swedish (SW) and Turkish (TR). Of these, German is set aside as a test target language. The remaining 9 languages are included in different amounts to also study the effect of data size in a multilingual setting. 10% of the sentences are set aside as testing data for each language.

Figure 3 presents the MCDs of the individual languages using the same multilingual voice. The x-axis is the amount of training data contributed by each language. The near-linear pattern of (es, pt, sw ru, en and zh) suggests that MCD (hence, voice quality) is proportional to the training data size, and this holds even in the multilingual setting. The good performance of Turkish and Japanese irrespective of the amount of training data may be explained by their simple phonetic structures.

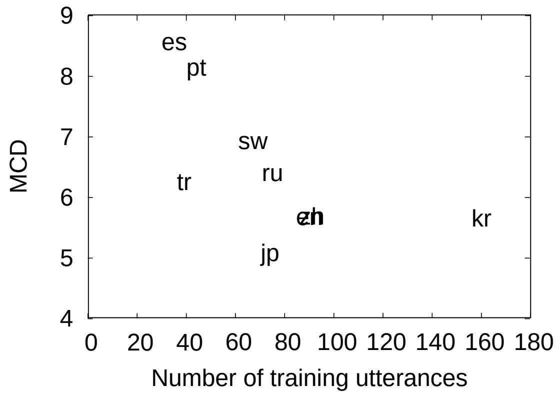


Fig. 3. MCDs of individual languages using a multilingual voice. Note: en/zh have the same amount of training data and the same MCD score.

For testing the new languages, we choose German (DE) and Telugu (TE) languages. The phonemes of these languages are mapped to their closest alternative from any of the nine different languages included as training. The overlap in the acoustic phonetic feature values of these phonemes are used to determine the closeness between phonemes (currently no weight is given to different acoustic phonetic features). The multilingual voice is incrementally adapted with data from the target language. Figure 4 shows the performance of the adaptation as MCD gains as a function of increasing amount of adaptation data. It can be seen that the German voice has a relatively lower MCD than the Telugu voice even without any adaptation. This may be explained by the fact that Telugu belongs to the Dravidian language family which is not represented in the training languages, while European languages are well represented. Informal listening tests also show that while the voices are understandable, they have new accents caused by the training languages.

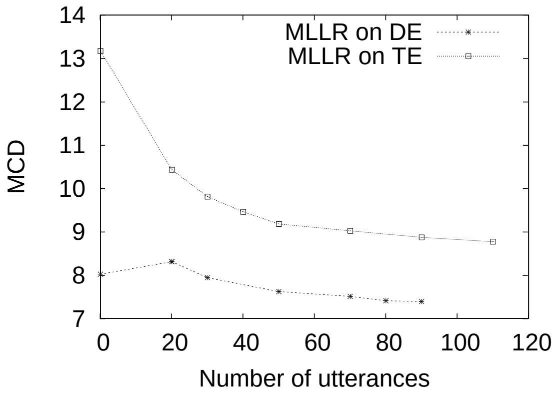


Fig. 4. German (DE) and Telugu (TE) language MCDs with increasing adaptation data

6. LANGUAGE SELECTION EXPERIMENTS

In this section, we report our experiments with changing the subset of languages included in training the multilingual voice. From, the initial subset of 9 languages chosen for training in the previous section, two subsets are created one including all but English and the other consisting of all but Chinese language. The choice of these languages is for two reasons: 1) They are phonetically quite distinct and 2) They contribute the same number of training sentences (as can be seen in the overlayed en/zh tags in the Fig. 3)

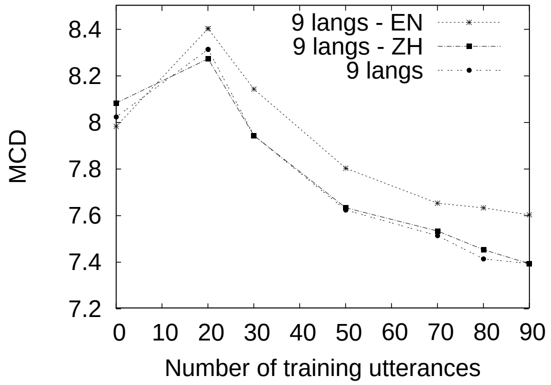


Fig. 5. German adaptation with different Training languages

From 6, as we expected the removal of English, a language phonetically similar to German, gives worse results, while the removal of Chinese, does not make much difference to the quality of German voice.

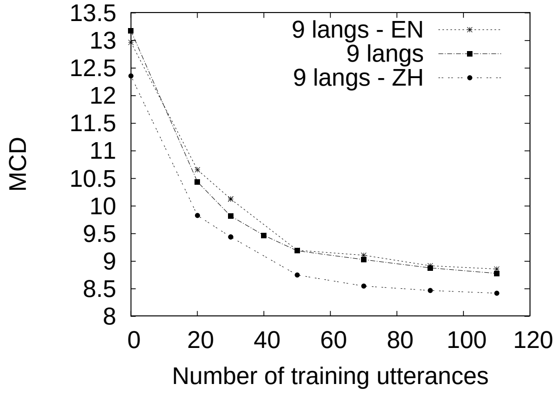


Fig. 6. Telugu adaptation with different Training languages

For the same experiment for creation of the Telugu voice, we find the removal of English does not make much difference, which we believe is due to the fact that Telugu and English are phonetically not very close. The unexpected result though is that the removal of Chinese improves the results. This shows that language selection is clearly important. One hypothesis for this result is the fact that Telugu has a larger number of stop distinctions than English (e.g. aspirated and unaspirated) such allophones do appear in English but are not phonetic. The initial models have these distinctions conflated, but become distinct with more adaptation data. However in Chinese, aspirated and unaspirated allophones do not occur within stops, hence the training data actually biases the initial phone models more and requires more training data to contract.

7. CONCLUSIONS

This work proposes adaptation techniques for under-resourced languages that clearly give promising results. The selection of initial models, although can be done by simple acoustic phonetic feature matching, our results show that more subtle selection of initial phonetic models and the languages that contribute to them may give even better results. We have yet to discover an efficient automatic method to improve these existing techniques.

The second important result is that the resulting synthesis quality seems to be linearly related to amount of training data, even across several languages.

8. ACKNOWLEDGMENT

This work was supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

9. REFERENCES

[1] K. Tokuda, T. Masuko, T. Kobayashi and S. Imai, “Speech Parameter generation from HMM using Dynamic Features”, *ICASSP ’95*, Detroit, USA, 1995.

[2] Javier Latorre, Koji Iwano and Sadaoki Furui, “New approach to the polyglot speech generation by means of an HMM-based speaker adaptable Synthesizer”, *Speech Communication*, 48:1227–1242, 2006.

[3] Tomoki Toda, Alan W Black and Keiichi Tokuda, “Voice Conversion based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, in *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222-2236, 2007.

[4] Yannis Stylianou, O. Cappé and E. Moulines, “Statistical Methods for Voice Quality Transformations”, in *Eurospeech*, Madrid, Spain, 1995.

[5] Viet-Bac Le, Laurent Besacier, and Tanja Schultz, “Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability” in *ICASSP*, 2006, Toulouse, France.

[6] Alan W Black, “CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling”, *Interspeech* 2006, Pittsburgh, PA.

[7] John Kominek and Alan Black “CMU ARCTIC databases for speech synthesis”, Tech Report CMU-LTI-03-177, Carnegie Mellon Unversity `http://festvox.org/cmu_arctic`, 2003.

[8] Tanja Schultz “GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University”, *ICSLP* 2002, Denver, USA.

[9] Yanagisawa K. and Huckvale Mark, “A Phonetic Assessment of Cross-Langage Voice Conversion.”, *Interspeech* 2008, Brisbane, Australia.

[10] A. J. Hunt and Alan W Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, *ICASSP* 1996, Atlanta, USA.

[11] Alan W Black and Kevin Lenzo, “Multilingual Text-to-Speech Synthesis”, *ICASSP* 2004, Montreal, Canada.

[12] John Kominek, Tanja Schultz and Alan W Black, “Synthesizer voice quality on new languages calibrated with mel-cepstral distorion”, *SLTU* 2009, Hanoi, Vietnam.

[13] Alan W Black and Kevin A. Lenzo, “Building Synthetic voices”, `http://festvox.org/`.

[14] Alan W Black and Tanja Schultz, “Speaker Clustering for Multilingual Synthesis”, *ITRW Multilingual Speech and Language Processing*, 2006, South Africa.

[15] D. Paul and J. Baker, “The design for the wall street journal based CSR corpus” *DARPA Speech and Natural Language Workshop*, 1992.

[16] Zen, H,. Black, Tokuda, K., and Black, A., “Statistical Parametric Speech Synthesis” *Speech Communication*, 51(11), pp 1039-1064, November 2009.

[17] Font Llitjos, A and Alan W Black “Unit Selection without a phoneme set” *IEEE TTS Workshop*, Santa Monica, USA, 2002.

[18] Maskey, S. , Tomokiyo, L. and Black, A. “Bootstrapping Phonetic Lexicons for New Languages” *ICSLP*, Jeju Island, Korea, 2004.

[19] Leggetter, C. J. and Woodland, P.C.. “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models” *Computer, Speech & Language*, pp. 171-185, 1995.

[20] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, “Speaker adaptation for HMM-based speech synthesis system using MLLR,” *ESCA/COCOSDA Workshop on Speech Synthesis*, Nov, 1998.

MALAY LANGUAGE MODELING IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION WITH LINGUISTIC INFORMATION

Hong Kai Sze[1,2]*, Tan Tien Ping*[2]*, Tang Enya Kong*[3]*, Cheah Yu-N*[2]

[1] Faculty of Engineering & Science, Universiti Tunku Abdul Rahman, Kuala Lumpur, Malaysia
[2] School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
[3] Universiti Multimedia, Cyberjaya, Malaysia
HongKS@utar.edu.my, TienPing@cs.usm.my, EnyaKong@mmu.edu.my, YNCheah@cs.usm.my

ABSTRACT

In this paper, our recent progress in developing and evaluating Malay Large Vocabulary Continuous Speech Recognizer (LVCSR) with considerations of linguistic information is discussed. The best baseline system has a WER of 15.8%. In order to propose methods to improve the accuracies further, additional experiments have been performed using linguistic information such as part-of-speech and stem. We have also tested our system by creating a language model using a small amount of texts and suggested that linguistic knowledge can be used to improve the accuracy of Malay automatic speech recognition system.

Index Terms— Speech Recognition, Agglutinative Language, Language Modeling, Part-Of-Speech, Stem

1. INTRODUCTION

Malay is spoken by more than 200 million people in Southeast Asia. It is an agglutinative language which allows base words to be combined with affixes to form new words [1]. This may change the meaning and part-of-speech of the base word. For example, “guna” (verb) can be combined with “peng” (prefix) to become “pengguna” (noun), meaning user. There are many studies on how affixation can be taken advantage of in an agglutinative language; one of it is class-based language model [2]. By using class-based language model, each part-of-speech (POS) is modeled as a distinct class and words are linked to their respectively POS to avoid data sparseness problem.

Some studies have suggested using morpheme-based units for modeling agglutinative language to avoid the need to have large dictionaries [3]. In these systems, affixes and based words are treated as morphemes to be re-combined later, like hidden-event language models. Other possibility is to model agglutinative language using factored language model with language features such as part-of-speech and stem information [4]. Using part-of-speech information in factored language model has been found to reduce word error rates (WER) of the agglutinative LVCSR system. There were also other findings that linguistic information does not necessarily improve the system. For instance, the stem information has been found to increase the WER in Estonian LVCSR [3]. In most cases, agglutinative characteristics can be useful in language modeling as it can provide an additional feature that can be used to classify different lexicon entries. For example, by knowing that “pengguna” consists of base word “guna” and prefix “peng”, its maximum likelihood probability can be interpolated to those probabilities of these sub-words [5].

This paper discusses our recent work in Malay LVCSR, particularly on Malay language modeling using agglutinative information. Section 2 discusses about Malay phonemes and agglutinative information. Section 3 describes about factored language models using agglutinative information. Section 4 and Section 5 elaborates on our testing results. Section 6, 7 and 8 includes the discussion, conclusion and future work, respectively.

2. MALAY PHONEMES AND AGGLUTINATIVE INFORMATION

In Malay language, a word can be formed by attaching base words with prefix, suffix, infix, circumfix, proclitic, enclitic and particle [5]. A prefix is located in front of base words, while a suffix is appended at the end of the base words. A compulsory prefix and suffix forms the circumfix. On the other hand, an infix is located within the base word. Enclitics functions as objective pronounce while proclitic functions as subjective pronounce. Particles are used to express various emotions. Only one clitic and one particle are allowed to attach to the base word.

There are 34 phonemes in Malay: 24 original phonemes and 10 borrowed phonemes [6]. From these 34 Malay phonemes, there are 25 consonants, 6 vowels and 3 diphthongs. We are able to use the n-gram language model because Malay sentence structures are fixed, which means that the language users may not switch the positions of words like what is allowed in some other languages like Arab, which requires extra independent word location modeling. [7]

3. FACTORED LANGUAGE MODEL USING MALAY LINGUISTIC APPROACH

Factored Language Model (FLM) [4] is a flexible framework which allows various types of language information to be incorporated. In FLM, the system utilizes different back-off paths to calculate the back-off coefficients. To specify these back-off paths, FLM use a Factored file. A Factored file (a control file for factored language model) is used to describe the specifications of language model that consists of various types of linguistic information.

In our Malay LVCSR system, we use word, stem and POS information. Word information will be removed from the back-off chains first before POS and stem information. Since the probability estimation for having a word in a sentence is more affected by the word's nearest history word, thus the near history word should be retained in the back-off chained. Cutoff point is set to 1 by default. To represent this information in training transcription, the following format is used:

W-pengguna:S-peng:P-noun

Where W denotes word, S denotes stem and P denotes part-of-speech. All words in the sentences were converted to the above format before they are trained using Factored Language Modeling.

4. EXPERIMENTS

In our LVCSR system, we have 34 different phone-like units and 9 non-speech units (including SIL and various other noises, like CLICK and etc). We used Sphinx ASR from CMU [8] to carry out the test. For training the acoustic model, 67 hours of Malay continuous read speech was used [10]. SRILM toolkit [9] and CMU-CAM Language Modeling toolkit were used for language modeling. We evaluated Malay language modeling using different amount of text corpus. The text corpus was collected from various newspapers over several years [10]. The moderate size text consists of 20.5 thousand sentences while the large training text corpus consists of more than 3.5 million sentences with about 71 million words, while testing text corpus contains more than 8 thousand sentences. The training text corpora do not contain any entries from testing text corpus. The vocabulary size is about 32 thousand words.

In Section 4.1, we look at the effect of different n-gram order to the language perplexity. Next, different smoothing methods will be used to evaluate language model perplexity in Section 4.2. In Section 4.3, the language model with lower perplexity will be selected for testing using Malay automatic speech recognition system in different configurations.

4.1. Language Modeling with Different N-Gram Order

In this section, we set up the experiments to train the language models using the moderate corpus described above to avoid computational power and memory constraints for higher order of n-grams. The language model was trained using SRILM toolkit with Good-Turing smoothing method. The orders of n-grams ranged from 1-9 were tested. *Table 1* shows the perplexity and entropy of the language models.

Table 1: Perplexity and Entropy vs. N

N-Gram Order	Perplexity	Entropy
1	1723.6	10.75
2	498.21	8.96
3	417.09	8.7
4	426.41	8.74
5	437.72	8.77
6	443.03	8.79
7	445.47	8.8
8	446.5	8.8
9	447.3	8.81

The results shows that both perplexity and entropy are near to optimal when n=3.

4.2. Language Modeling with Common Smoothing Methods

In this experiment, we utilized CMU-CAM toolkit to train and evaluate Malay language models (trigram) using four common types of smoothing methods. These methods are common in the literature, like Absolute, Good-Turing, Linear and Witten-Bell methods [11]. We tested the approaches using the language models (trigram) created from moderate size text corpus and large size text corpus. The purpose of these experiments is to find out the best language model to be used for our Malay LVCSR system. Different smoothing methods can work well in different configurations.

Table 2: Perplexities vs. Smoothing Methods (Moderate Size Corpus)

Smoothing Method	Perplexity
Absolute	457.388
Good-Turing	448.609
Linear	478.024
Witten-Bell	451.077

Table 2 shows that Good-Turing outperforms other smoothing methods slightly in term of perplexity with moderate amount of texts. We further evaluated the smoothing algorithms using large amount of texts and the results are shown in *Table 3* and *Table 4*. Language models with different n-gram orders from 1-gram to 4-gram were trained using Good-Turing, Witten-Bell and Linear smoothing algorithms. Table 3 shows that tri-gram with Witten-Bell smoothing method has the lowest perplexity compared to other number of grams and smoothing methods.

Table 3: Perplexities of Large Language Models

LM Order	Good-Turing	Witten-Bell	Linear
1	2549.02	2549.02	2549.02
2	385.48	378.44	374.38
3	93.56	73.42	74.36
4	39.09	19.33	22.97

Table 4: Entropies of Large Language Models

LM Order	Good-Turing	Witten-Bell	Linear
1	11.32	11.32	11.32
2	8.59	8.56	8.55
3	6.55	6.20	6.22
4	5.29	4.27	4.52

We are aware of some newest smoothing methods like Kneser-Ney and Modified Kneser-Ney methods, which are reported to perform better.

4.3 Incorporating Statistical LM into Malay LVCSR Baseline System

We discussed about training language models in Section 4.1 and Section 4.2. The purpose of these experiments is to find out the smoothing method that can best model Malay. In this section, we will incorporate these language models into our Malay LVCSR. We trained our acoustic models using two third of available 67 hours of speech sound corpus while the rest were used for testing purpose [10]. Different set of speakers were used for training and testing transcriptions.

The first experiment used was carried out using language model created with the moderate size language model, while the second experiment was using language model created with large text corpus.

Table 5: Malay LVCSR WER using various Language Models

Type of Language Model	Word Error Rate
Moderate Size Training Corpus	20.3
Large Size Training Corpus	15.8

Witten-Bell smoothing method was used for comparison purpose. We observed that large training corpus outperformed moderate training corpus by reducing LVCSR WER from 20.3% to 15.8% (4.5% reduction).

5. LINGUISTIC FEATURES EVALUATION

In this section, we include part-of speech and stem information to evaluate language models built using Malay speech corpus. We used an internal POS tagger and an internal Malay Stem Extractor to parse and tag all Malay training and testing transcriptions. The POS tagger is part of S-SSTC project [12].

The POS tagger is a stochastic-based tagger based on Q-tag while the stem extractor is a JAVA program written with 2554 base-words and a rich set of linguistic segmentation rules.

For preliminary test, we have extracted the linguistic information from about 10106 training sentences and 1072 testing sentences. For each word in a sentence, the POS and stem information is attached.

5.1 Factored Language Models (FLM)

In our Malay language modeling, we utilized word, stem and POS information. The testing text corpus contains 8245 sentences with 2773040 words. The text corpora are tagged with stem extractor and POS tagger.

The experiments are designed in two-folds. First we plan to find out language modeling top limit in text prediction. Secondly, we want to know whether we could reduce the amount of text corpus used for training without sacrificing the WER of ASR. This is because we do not necessary have a large text corpus all the time, especially for under resourced languages. The results for bigram using FLM are shown in *Table 6*. In *Table 6*, W(-1) denotes the immediate previous word, S(-1) denotes the stem of the immediate previous word while P(-1) denotes the part-of-speech of the immediate previous word. For notation purpose, we used W to denote word, S to denote stem and P to denote part of speech. The number in the bracket denotes *n*-th previous word. W(-1) indicates the word before, and W(-2) denotes the word before the previous word.

Table 6: Perplexities with Linguistic Information

Configuration in FLM	**Perplexity**
W(-1) S(-1) P(-1)	1031.7
W(-1) S(-1)	1301.12
W(-1) P(-1)	447.582

The back-off will be performed from the rightmost information to the leftmost information. E.g. for "W(-1),S(-1)" type of FLM, S(-1) will be eliminated first. We used Good-Turing smoothing algorithm with cut-off point of 1 and interpolation at all back-off levels.

With these configurations (*Table 6*), we can see that third configuration using word and POS information in factored language modeling outperformed first and second configurations. The second configuration, on the other hand, also verified that paper [3]'s observation about the incapability of stem information to reduce WER. Besides evaluating the effects of linguistic information on perplexities, we also evaluated different FLM configurations using stem information.

Table 7: Perplexities with Different FLM Configurations

Factored Language File	**Perplexity**	**Perplexity excluding end-of-sentence tokens**
bigram.flm (Use W_{-1} only)	705.466	1139.49
bigram1.flm (Use W_{-1} and S_{-1}, then backoff to W_{-1})	731.614	1186.67
bigram2.flm (Use W_{-1} and S_{-1}, then backoff to S_{-1})	762.414	1240.38
bigram3.flm (Use W_{-1},S_{-1},S_0, then backoff to W_{-1}, S_{-1} finally backoff to S_{-1})	524.083	829.517
bigram4.flm (Use W_{-1}, S_{-1}, S_0, then backoff to W_{-1}, S_{-1}, finally backoff to W_{-1})	499.333	787.544
bigram5.flm (Use S_{-1}, S_0, then backoff to S_{-1})	**275.737**	**416.359**
bigram6.flm (Use W_{-1}, S_{-1}, then backoff to S_{-1})	1231.09	2074.51
trigram.flm (Use W_{-1}, W_{-2}, then backoff to W_{-1})	685.124	1104.27
trigram1.flm (Use W_{-1}, W_{-2}, S_{-1}, then backoff to W_{-1}, W_{-2}, finally backoff to W_{-1})	722.47	1170.76
trigram2.flm (Use W_{-1}, W_{-2}, S_{-1}, then backoff to W_{-1}, S_{-1}, finally backoff to W_{-1})	875.898	1439.58
trigram3.flm (Use W_{-1},W_{-2}, S_{-1},S_{-2}, then backoff to W_{-1}, S_{-1}, S_{-2}, next backoff to S_{-1}, S_{-2}, finally backoff to S_{-1})	1236.85	2084.92
trigram4.flm (Use W_{-1},W_{-2}, S_{-1}, S_{-2}, then backoff to W_{-1}, W_{-2}, S_{-1}, next backoff to W_{-1}, W_{-2}, finally backoff to W_{-1})	836.143	1369.57
trigram5.flm (Use W_{-1},W_{-2}, S_{-1}, S_0, then backoff to W_{-1}, W_{-2}, S_{-1}, next backoff to W_{-1}, W_{-2}, finally backoff to W_{-1})	726.307	1177.44
trigram6.flm (Use W_{-1},W_{-2}, S_0, then backoff to W_{-1}, W_{-2}, finally backoff to W_{-1})	625.477	1002.93
trigram7.flm (Use S_0, S_{-1}, S_{-2}, then backoff to S_{-1}, S_{-2}, finally backoff to S_{-2})	488.727	769.603
trigram8.flm (Use S_0, S_{-1}, S_{-2}, S_{-3}, then backoff to S_{-1}, S_{-2}, S_{-3}, next backoff to S_{-2}, S_{-3}, finally backoff to S_{-3})	885.772	1457.01

For explanation on how to understand the factored file, the reader may refer to [4]. The best FLM specifications are in the first column of *Table 7*. In this experiment, we used 80k sentences of training transcriptions and 8k of sentences of testing transcriptions.

The results are affected by data sparseness problem. We can observe that "S0,S1,W1" can perform better than "S0,S1,S2,W1,W2". The reason for above observation is that FLM increases data sparseness problem in language modeling as there are more varieties of identical word forms. On the other hand, large text corpus may not be feasible to some under-resourced languages. Moreover, processing time and availability of language expert are also the concerns. This has increased the requirement for robust parameter estimation.

The factored language models were evaluated based on above text corpus and their perplexities are lowest when only stems were used. It is due to data sparsity problem occurred for large lexicon. The lowest perplexity that has been obtained is 275.737 using stem solely and bi-gram. We observed that using stem information alone with small text corpus could have the lowest perplexities. From the test, the language model that gives the best result use the following equation:

$$P(w_t | s_t, s_{t-1})$$

It does not necessary conclude that POS works better in huge corpus, but in a large text corpus, most of the linguistic information (like POS and stem) has already been embedded in the word sequence. The stem information will add on to the data sparsity problem as more unique lexical items are available. The reason is for the same size of corpus of Malay language, the identical number of POS is less than stems.

5.2 Different Corpus Sizes Using Linguistic Information for Malay

In this section, we want to examine the effect of corpus size using stem information. We designed the experiments so that the number of sentences was varied from 1000 to 900000. In fact, our text corpus consists of about 3.5 million sentences, but when using more sentences, more times are needed. This is especially true for mobile devices which have limited processing power during decoding stage. The result is shown in *Figure 1*.

First configuration is a trigram using word and POS information only while second configuration is a trigram using word, POS and stem information. We can see that with a limited text corpus, stem information always give better perplexities, which is in contrast with those reported for Estonian language. [3] For all corpus size, perplexities with stem information outperformed perplexities without stem information.

6. DISCUSSIONS

In these experiments, language model built from large corpus using Witten-Bell smoothing algorithm has been found to have the lowest perplexity in Malay ASR. On the other hand, although we found that word and POS information only is adequate enough in *Table 6*, stem information has shown that it can lower the perplexity in *Table 7*.

Error analysis was carried out using SCLITE. Most errors are caused by similar pronunciations between words and word phases. The observations implied that language modeling can play a very important role whenever there are acoustic ambiguities. This has encouraged our researchers to move towards this direction.

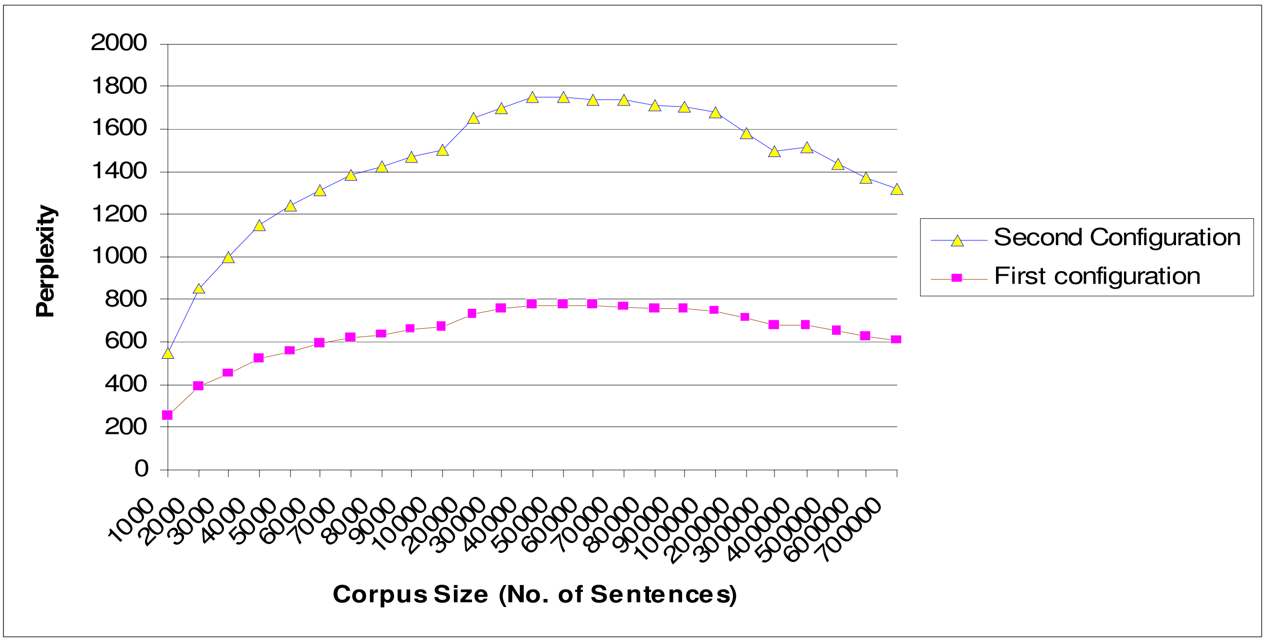


Figure 1: Perplexities vs Corpus Sizes using Linguistic Information

7. CONCLUSION

In conclusion, the development of Malay Large Vocabulary Speech Recognition System has been presented. Various experiments on Malay Language Modeling have been performed and 3-gram with Witten-Bell smoothing is found to be the best configuration. We also concluded that linguistic information like POS and stem can aid in reducing WER/perplexity in resource limited environments.

8. FUTURE WORK

Our proposal is to develop the 2-pass decoder which utilizing the benefits of morphology information. Some literature has shown that agglutinative language tends to create compound words that can be rectified by hidden event language model. Moreover, building n-gram for multi-words and morpheme-like units can further reduce WER for Malay LVCSR system.

In Malay language, we can extract some information like part-of-speech, stem, plural or singular, present-past-future tenses. [1] The combination of the information will be used to improve our speech recognition system. For derived word detection, we can improve the correction reliability by using statistical language modeling techniques, like bigram.

9. ACKNOWLEDGEMENTS

This research work is sponsored by Malaysia MOSTI Science Fund (No: 01-02-11-SF0070) "Malay Phoneme Recognition for Continue Utterance" granted to Universiti Tunku Abdul Rahman. The computing equipments are provided by Universiti Sains Malaysia. Besides, we would like to thank Dr. Tang EK and Mr Chua CC for MMU for sharing their Malay POS tagger.

10. REFERENCES

[1] Nik Safiah Karim, Farid M. Onn, Hashim Hj. Musa, Abdul Hamid Mahmood, "*Tatabahasa Dewan*", 11th Ed., 2006, Dewan Bahasa dan Pustaka, Kuala Lumpur.

[2] Imed Zitouni, "*Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition*", Computer Speech and Language 21 (2007) 88-104.

[3] Tanel Alumae, "*Sentence-Adapted Factored Language Model for Transcribing Estonian Speech*", ICASSP 2006, pp. I429 – I432.

[4] Katrin Kirchhoff, Jeff Bilmes, Kevin Duh, "*Factored Language Models Tutorial*", UWEE Technical Report Number UWEETR-2008-0004, February 2008.

[5] Bali Ranaivo-Malancon, "*Computational Analysis of Affixed Words in Malay Language*", UTMK internal publication, Universiti Sains Malaysia, 2004.

[6] Indirawati Zahid and Mardian Shah Omar, "*Fonetik dan Fonologi*", PTS Professional, 2006.

[7] Ruhi Sarikaya, Mohamed Afify, Yonggang Deng, Hakan Erdogan and Yuqing Gao, "*Joint Morphological-Lexical Language Modeling for Processing Morphologically Rich Languages With Application to Dialectal Arabic*", IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 7, September, pp. 1330 – 1339.

[8] Web Page:
http://cmusphinx.sourceforge.net/html/cmusphinx.php

[9] Web Page:
http://www-speech.sri.com/projects/srilm/

[10] Tien-Ping Tan, Xiong Xiao, Tang EK, Eng Siong Chng, Haizhou Li, "*MASS: A Malay language LVCSR corpus resource*", 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, 10-12 August 2009, pp. 25-30.

[11] Stanley F. Chen and Joshua Goodman, "*An Empirical Study of Smoothing Techniques for Language Modeling*", TR-10-98, August 1998, Havard University.

[12] Al-adhaileh, Mosleh H., Tang Eya Kong and Zaharin Yusoff, "*A Synchoronization Structure of SSTC and its Applications in Machine Translation*", The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan, September, 2002.

Development of a Speech-to-Text Transcription System for Finnish*

Lori Lamel[1] *and Bianca Vieru*[2]

[1]Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
lamel@limsi.fr

[2]Vecsys Research
3, rue Jean Rostand
91400 Orsay, France
vieru@vecsysresearch.com

Abstract

This paper describes the development of a speech-to-text transcription system for the Finnish language. Finnish is a Finno-Ugric language spoken by about 6 million of people living in Finland, but also by some minorities in Sweden, Norway, Russia and Estonia. System development was carried out without any detailed manual transcriptions, relying instead on several sources of audio and textual data were found on the web. Some of the audio sources were associated with approximate (and usually partial) texts, which were used to provide estimates of system performance.

1 Introduction

Traditionally speech-to-text transcription (STT) systems are trained on large amounts of carefully transcribed speech data and huge quantities of written texts. However obtaining the needed transcribed audio data remains quite costly and requires substantial supervision. Several research directions have addressed reducing these costs [6] and much of the recent audio training data, as in the DARPA Gale program are associated with quick transcriptions (QTR) [7]. For certain audio sources, it is possible to find associated texts, ranging from quite accurate, but usually incomplete, transcriptions, to closed captions, summaries or other less closely related texts. A variety of approaches have been investigated most relying on supervision from a language model. The approaches differ in their details: use or not of confidence factors [8] or [9], [10], doubling vs iterative training [11] and the amount of data used.

In this study, system development is also lightly supervised, in that no detailed annotations are available for the development and test data. Initially approximate reference transcriptions were used to assess both acoustic and language models during system development. Only afterward were the transcripts manually corrected in order to have a better estimate of the true performance.

The next section gives an overview of the characteristics of the Finnish language, followed by a description of the approach and corpus used in this study. This is followed by a description of the language models, phone set and acoustic models, after which experimental results are provided.

2 Finnish Language

Part of Uralic languages, Finnish is a Finno-Ugric language spoken by about 6 million of people living in Finland, but also by some minorities in Sweden, Norway, Russia and Estonia.

Finnish shares a basic vocabulary with the other Uralic languages and has various derivational suffixes. It has regular letter-to-sound correspondences, which simplifies the problem of pronunciation modeling. While Finnish has a smaller core vocabulary than English, it allows creation of new words by extensive use of agglutination, resulting in a very large lexical variety.

Most of the reported speech-to-text transcription results for the Finnish language are substantially worse than results reported for more resourced languages such as English or French. A first explanation could be that the extensive use of agglutination in Finnish which has impact on the language modeling difficulties. In [1] it is highlighted that using a 20K word vocabulary in English gives a lower OOV rate than a 500000-word vocabulary in Finnish. For example 40-million-word English corpus contains about 190000 distinct words, while the corresponding Finnish corpus contains about 1.9 million unique words. A proposed solution to this problem is the decomposition of words into morphs as shows in [2, 3].

But another explanation of this poor results is the lack of suitable speech and text training data resources. If in 2002, about 72% of the websites were in English although that

*This work was supported in part by OSEO under the Quaero program and by the European Commission under the VITAL project.

was the language of only a third of the Web users, in 2003 Finnish is found to be used in 1% of a random selection of web pages [4].

The results are less dramatic when one looks at the languages used by people to communicate with each other via the Web: they then prefer to use their mother tongue. Moreover, the recent rise of the blog, reflecting a desire to reach a smaller audience closer to the writer, could allow an increase in the range of languages used on the Internet, in particular French [5].

3 Approach and Corpus

The general approach taken in the work is similar to that of [11, 12, 13] in that a speech recognizer is used to provide "approximate" transcripts for acoustic model training. The audio data is transcribed in batches, and in successive iterations the models are trained on more data. In [14] an analysis of training behavior is compared for supervised and unsupervised approaches.

In contrast to previous studies where audio and text data were available for model training, the first challenge in this study was locating audio and text data in Finnish. Three types of audio data were found. The first data are from a website which we refer to as BN Learning website, diffusing news audio data with close transcriptions targeting an audience of non-native speakers of Finnish. The data on this site use a simplified language so as to be accessible to foreigners. A total of 31 hours of audio with corresponding approximate transcriptions (102k words) have been downloaded since November 2007. A second data set containing 19 hours of audio with approximate transcriptions was downloaded from the Finnish News Agency. These audio correspond to short newswires diffused hourly for native Finnish speakers. The transcripts cover only part of the audio and are not aligned.

Since the initial word error rate estimates were quite low on these data compared to previously published results for Finnish, it was decided to extend the range of data sources and types (general news, special reports, interactive shows).[1] A total of 190 hours of varied broadcast data were collected from a variety of Finnish sources. The audio data used in this study are summarized in Table 3. In addition to the audio data and (when available) associated transcripts, 30M words from text materials were collected.

Initial acoustic and language models were built using just the BN learning corpus. Then the FNA data were added, and finally some of the more general data. In order to provide supervision in acoustic model training, the language models used in the early decoding stages of the audio data were heavily biased, being trained on texts from the same epoch of the BNL transcripts. The language models used for test purposes were initially also only trained on the BNL data, but quickly additional texts were included. It should be noted that both AM and LM development were ongoing, as the text normalization was progressively improved.

Texts	Transcription		Newspapers
	BNL	FNA	
Train	78K	193K	30M
Dev	24K	48K	

Table 1: Text corpora used for language modeling.

4 Language models

Texts from over 20 different sources, mainly newspapers, formed the language model training corpus. As can be seen in Table 1 approximate transcriptions of audio data represent less than 1% of the text corpus. Concerning the transcriptions, it should be noted that the BN learning texts uses a substantially simplified language compared to standard Finnish broadcast news.

In Finland, newspaper articles are written in Finnish or in Swedish, both being official languages. Also sometimes small citations or entire articles can be found in English, Russian, Estonian or other languages in texts downloaded from Internet. Since the language is not always clearly indicated in the texts, text based language identification using the program TextCat [15] was run on each processed paragraph and only Finnish paragraphs were retained.

As is standard practice, the texts were split into sentences and the main punctuation was removed. During normalization all words were converted to lowercase, and words with a dash or a colon were separated, keeping the dash and colon as words. Numbers were transformed to a full, spoken form. This is quite complicated for the Finnish language which has 15 declensions cases and all parts of numbers should be declined. Some cases are constructed by adding suffixes, such as 's' in ordinals, after each component number. Given the complexity of expanding numbers into words for different cases, and our lack of knowledge about the Finnish language, it was decided to first only use the nominative case. After processing, the texts contained a total number of about 30M words, with a vocabulary size of about 1.4M words.

Finnish is an agglutinative language, using suffixes to ex-

[1] These sources were found by a native Finnish speaker who also, after we had developed a system with lightly supervised references, corrected the transcripts of the BN Learning and FNA news data.

press grammatical relations and also to derive new words, so the vocabulary expands rapidly. There is no grammatical gender for nouns however all 15 cases are even even for proper names. This makes the loan words difficult to treat since each word could appear in each of these 15 cases. For example, the following forms were found in the texts:

> ***Bush*** *Bushia Bushien Bushiin Bushilla Bushille Bushilta Bushin Bushissa Bushista Bushit Bushkin*
>
> ***Obama*** *Obamaa Obamaan Obamalla Obamalle Obamalta Obaman Obamassa Obamasta Obamat*

A list of words was selected by interpolation of unigram models trained the normalized texts from different newspapers and the approximate transcripts associated with the audio data. The Morfessor [16] decompounding algorithm is applied to this list to determine possible word decompositions. For example, for the word *elinaikakerroin* (survival factor) Morfessor proposes *elin* + *aika* + *kerroin*, which is mapped to *elin_ _aika_ _kerroin* in order to keep track that the lexical entries result from a decomposition. In order to avoid creating too many small, easily confusable lexical entries, a minimum of 3 characters per unit was imposed. All of the texts are decomposed using the selected decompositions proposed by Morfessor. Since the resulting lexical entries differentiate words from the decomposed forms, the language models decide the appropriate form and the forms in the hypotheses can simply be glued back together. The total number of tokens in the text corpus is increased as a result of word decomposition, but the number of distinct word forms is divided by two.

As mentioned in Section 3 biased n-gram language models were constructed to decode the audio by training on only the associated approximate transcriptions collected from the same period (usually 1 month) in order to provide strong, but flexible supervision. These initial LMs were based on full word lexical entries (no decomposition) and were used only for the first acoustic models.

For the second iteration, language models trained on all the transcripts from the same year and type as the audio data were constructed in order to have a more general LM. The LMs were interpolated with a general language model trained on the entire text corpus, with each component LM having an equal mixture weight.

Different language models were used in speech recognition experiments. For most of the experiments, the language models use a 300k word list optimized on the BNL+FNA dev data. The n-gram language models were obtained by interpolation of backoff n-gram language models trained on separate subsets of the available language model training texts using the modified Kneser-Ney smoothing. The characteristics of the 300k 4-gram language models are summarized in Table 2. The mixture weights were automatically chosen using the EM algorithm to minimize the perplexity of the development data. It can be seen that the perplexity and OOV rates of the BNL data and the FNA dev are much lower than the test data.

Type	BNL_dev	FNA_dev	FNA_test	BN_test
OOV	0.67	1.81	4.01	3.85
ppx	193	386	2418	2668

Table 2: Perplexity (PPX) and Out Of Vocabulary (OOV) rates for the different sets of dev and test data using a 300k LM. The LM mixture weights were tuned on the dev data.

5 Phone Set & Acoustic models

Words of foreign origin excluded, Finnish is written with 8 letters for vowels and 13 for consonants. All the vowels and almost all the consonants can be either short or long sounds. The phone set used in this work is composed of 42 phones: 16 vowels, 27 consonants and three units for silence, breath and filler. The long and short phones are represented with separate symbols and have separate acoustic models. Standard Finnish is basically a phonetic language where each letter corresponds to one and the same phoneme, and each phoneme corresponds to one and the same letter [17]. So, with very few exceptions, the lexicon observes a strict correspondence between letters and phonemes, with a low number of variants (avg 1.1 pronunciations/word).

A multi-language, cross-language bootstrapping [18] was used to initialize the acoustic models. Phones from English, French, German, Italian and Arabic were mapped to Finnish phones, and models extracted from corresponding acoustic model sets served as initial seed models. The first month of BN learning data was decoded using these models and a language model built only on transcriptions of that month (with a 22k word LM). The acoustic models were trained in a lightly supervised manner [13], one month at a time until the full 14 hours of speech from the BN Learning (BNL) corpus was used. For the first stages a 22k LM was used to decode the audio data. Data from the standard BN (FNA) were then progressively added with larger models trained after each step.

The standard cepstral features (perceptual linear prediction - PLP) were used. The PLP feature vector has 39 cepstral parameters: 12 cepstrum coefficients and the log energy, along with the first and second derivatives. The acoustic models are tied-state, left-to-right context-dependent, HMMs with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent, but

Audio	**Learning**	**FNA**	**BN**
Train	19h	11h	170
Dev	7h	4h	-
Test	-	4h	16h

Table 3: Audio corpus (in hours) used for training, dev and test for the Finnish STT system.

Model			Audio corpus	
set	ctx	#Gaussians	Hours	Sources
BN0	8345	190k	26	BNL+FNA
BN1	9713	239k	35	+BN 9 hrs
BN2	10568	272k	42	+BN 16 hrs
BN3	12493	355k	63	+BN 37 hrs
BN4	18268	369k	195	+BN 169 hrs

Table 4: Characteristics of different acoustic model sets.

word position-dependent. The tied states are obtained by means of a decision tree. The acoustic models are gender-independent and speaker-adaptive trained (SAT). Silence is modeled by a single state with 1024 Gaussians. The best model trained on only the BN Learning corpus cover about 5.6k phone contexts, with 3.7k tied states and 32 Gaussians per state. With the additional 11 hours of FNA data, the acoustic models cover 8k contexts and 6k tied states. These models, trained on the pooled data were also then MAP [19] adapted to each audio corpus. As more of the varied BN data was progressively added, larger models were built, with the largest covering about 18k contexts as shown in Table 4.

6 Experimental results

This section reports a series of experiments assessing recognition performance as a function of the available acoustic and language model training data. The system is based on the LIMSI broadcast news transcription [20] was used. It has two main components, the audio partitioner and the word recognizer. During development of the Finnish STT system, all evaluation was done using selected portions of the web transcriptions as references (based on string alignments). These may be inexact and often contain either fewer or more words than in exact transcriptions. After system development, a native Finnish speaker corrected these transcriptions and a real scoring was realized.

Figure 1 shows the recognition results using web transcripts and the corrections made by a native Finnish on the BN learning corpus. In these experiments, acoustic models were built using only the BNL data with a vocabulary of 700k decomposed words and language models built on a 8M text corpus available at the time. As can be seen in the figure, the two error curves closely follow each other, with slightly optimistic results with the approximate transcripts.

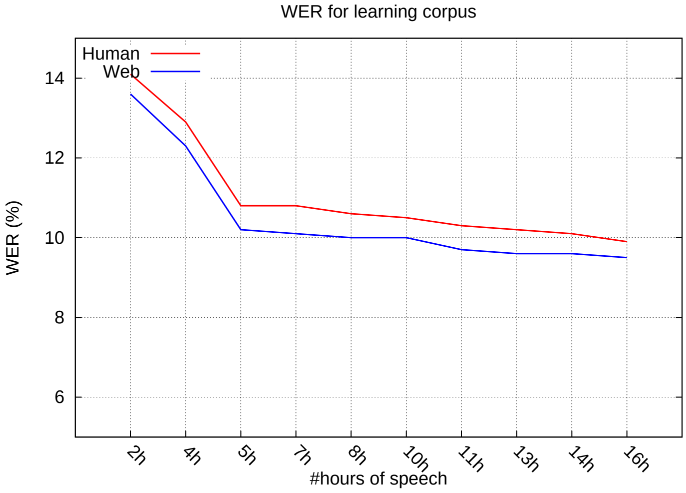


Figure 1: Performance on BN Learning development data using a 700k LM estimated on a 10M word text corpus.

As in [12] a speech recognizer was used to automatically transcribe unannotated data and generating ”approximately” labeled training data. As the amount of training data increases iteratively, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. The data were added progressively, choosing the data with good likelihood scores first [8, 9]. The characteristics of the acoustic models are given in Table 4.

Figure 2 shows that using web references for scoring can give an idea of system performance of different acoustic sets. For each set of curves, the solid line corresponds to scoring with approximate web transcripts and the dotted lines scoring with manually corrected references when available. It can be seen that although the absolute levels are different, the behavior of the curves are quite similar. There is a particularly a big difference for the FNA_test, which is due to the fact that available web transcriptions do not cover all of the audio data, so the insertion rate is very high. In contrast, the curves are very close for the BNL dev data for which close approximate transcriptions are available. It can also be seen that as progressively more varied BN data are included in the training, the BNL and FNA results slowly degrade. The first set of models (BN0) are trained on only FNA and BNL data, so these are closer to the dev and test data. These experiments all used the same 300k word list

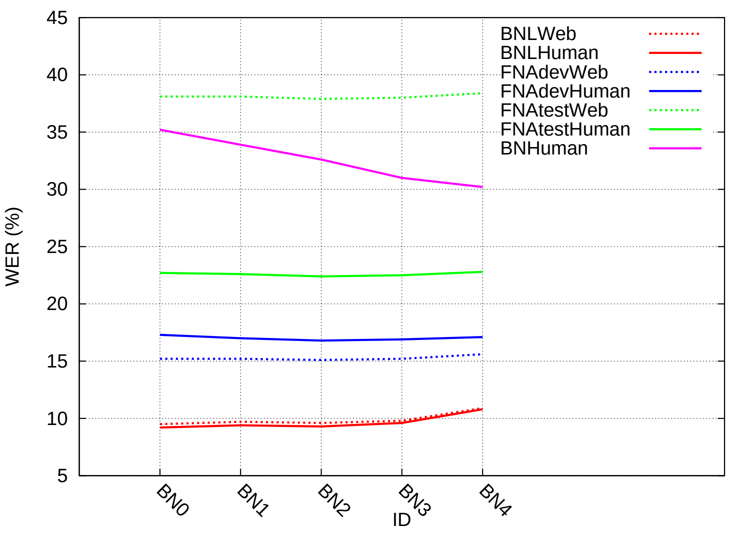


Figure 2: System performance with manual or web references (when available). Acoustic models are built on BNL and FNA, and progressively more BN data. A 300k vocabulary obtained by interpolation of 1-grams on BNL+FNA dev was used.

	BNL	FNA_dev	FNA_test	BN
Web	8.8	14.4	21.9	-
Human	9.1	16.3	37.3	29.4

Table 5: WER with manual references of best system for each type of data with a two pass decoding and unsupervised acoustic model adaptation.

selected by interpolating 1-grams so as to optimize the coverage of the BNL and FNA dev data, and language models trained on the 34M word (decompound words) corpus.

Table 5 gives the best results obtained on different data types. These results are obtained using a 2 pass system, with unsupervised acoustic model adaptation between decoding passes [20]. The acoustic models are also specific to each data type, being MAP [19] with the available audio training data from each audio corpus (using the automatic transcripts).

7 Conclusions

This paper has described the development of a speech-to-text transcription system for the Finnish language. The first task was locating appropriate resources for acoustic and language model training, and system assessment. In doing so the methodology used in lightly supervised or unsupervised acoustic model training has been extended to system development since no carefully transcribed development data was available for model optimization. Transcription word error rates were reported with approximate web transcripts that were used during system development and with manual transcripts that were later created, and although the approximate transcripts give an optimistic estimate of the true word error rates they were found to be useful for system optimization.

References

[1] Teemu Hirsimki, *Advances in unlimited-vocabulary speech recognition for morphologically rich languages*, Phd dissertation, Helsinki University of Technology, Department of Information and Computer Science, 2009.

[2] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *in Proc. Eurospeech*, 2003, vol. 20, pp. 2293–2296.

[3] T. Hirsimaki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkonen, "Unlimited vocabulary speech recognition with morph language models applied to finnish,"*Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, October 2006.

[4] Edward T. O'Neill, Brian F. Lavoie, and Rick Bennett, "Trends in the evolution of the public web, 1998 - 2002," *D-Lib Magazine*, vol. 9, 2003.

[5] Anne de Beer and Grard Blanc, "La diversit des langues sur internet," *Futuribles*, vol. 329, pp. 29–36, 2007.

[6] O. Kimball, C.L. Kao, R. Iyer, T. Arvizo, and J. Makhoul, "Using quick transcriptions to improve conversational speech models," *INTERSPEECH*, pp. 2265–2268, 2004.

[7] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," *LREC*, pp. 69–71, 2004.

[8] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross domain automatic transcription on the TC-STAR EPPS corpus," *ICASSP*, vol. 1, pp. 825–828, 2005.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[10] F. Wessel and H. Ney, “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[11] J. Ma and R. Schwartz, “Unsupervised versus supervised training of acoustic models,” *INTERSPEECH*, pp. 2374–2377, 2008.

[12] L. Lamel, J.L. Gauvain, and G. Adda, “Lightly supervised acoustic model training,” *ITRW ASR*, vol. 1, pp. 150–154, 2000.

[13] L. Lamel, J.L. Gauvain, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[14] S. Novotney and R. Schwartz, “Analysis of low-resource acoustic model self-training,” *INTERSPEECH*, pp. 244–247, 2009.

[15] G. van Noord, “http://www.let.rug.nl/vannoord /textcat/,” .

[16] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0,” Tech. Rep., Helsinki University of Technology, 2005.

[17] F. Karlsson, *Finnish: An Essential Grammar (Routledge Grammars)*, Routledge, 1st edition, 1999.

[18] J. Loof, C. Gollan, and H. Ney, “Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system,” *INTERSPEECH*, pp. 88–91, 2009.

[19] J.L. Gauvain and C.H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[20] J.L. Gauvain, L. Lamel, and G. Adda, “The limsi broadcast news transcription system,” *Speech Communication*, vol. 37, pp. 89–108, 2002.

MORPHEME-BASED AUTOMATIC SPEECH RECOGNITION FOR A MORPHOLOGICALLY RICH LANGUAGE – AMHARIC

Martha Yifiru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel

Department of Informatics, University of Hamburg
Vogt-Kölln Str. 30, D-22527 Hamburg, Germany
abate,tachbeli,menzel@informatik.uni-hamburg.de

ABSTRACT

Out-of-vocabulary (OOV) words are a major source of error in a speech recognition system and various methods have been proposed to increase the performance of the systems by properly dealing with them. This paper presents an automatic speech recognition experiment conducted to see the effect of OOV words on the performance speech recognition system for Amharic (a morphologically rich language). We tried to solve the OOV problem by using morphemes as dictionary and language model units. It has been found that for a small vocabulary (5k) system morphemes are better lexical and language modeling units than words. An absolute improvement (in word recognition accuracy) of 11.57% has been obtained as a result of using a morph-based vocabulary. However, for large vocabularies morpheme-based systems did not bring much performance improvement as they suffer from acoustic confusability and limited language model scope while word-based recognizers benefit much from OOV rate reduction.

Index Terms— Out-of-Vocabulary problem, Morpheme-based speech recognition, Amharic

1. INTRODUCTION

Most large vocabulary speech recognition systems operate with a finite vocabulary. All the words which are not in the system's vocabulary are considered out-of-vocabulary words. These words are one of the major sources of error in an automatic speech recognition system. When a speech recognition system is confronted with a word which is not in its vocabulary, it may recognize it as a phonetically similar in-vocabulary unit/item. That means the OOV word is mis-recognized. This in turn might cause its neighboring words also to be mis-recognized. [1] indicated the fact that each OOV word in the test data contribute to 1.6 errors on the average. Therefore, different approaches have been investigated to cope with the OOV problem and consequently to reduce the error rate of automatic speech recognition systems. One of these approaches is vocabulary optimization [2], where the vocabulary is selected in a way that it reduces the OOV rate. This involves either increasing the vocabulary size or including frequent words in a vocabulary. This approach may work for morphologically simple languages like English where a 20k vocabulary has 2% OOV rate and a 65k one has only 0.6% [3].

However, for morphologically rich languages, for which OOV is a severe problem, a much larger vocabulary is required to reach the 1% OOV rate. [3] indicated the fact that for Russian and Arabic 800k and 400k vocabularies are required, respectively for a 1% OOV rate. Increasing the vocabulary to alleviate the OOV problem is not the best solution especially for morphologically rich languages as the system complexity increases with the size of the vocabulary. Therefore, modeling sub-word units, particularly morphs, has been used for morphologically rich languages. Many researchers [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] did morpheme-based or sub-word based speech recognition experiments.

In this paper, we show the effect of OOV rate on the performance of an Amharic speech recognition system. We investigate options to reduce the OOV problem using morphemes as a lexical and language modeling unit and study its effect on the performance of the system. Section 2 gives a brief description of the Amharic word morphology. After reviewing previous works on morpheme-based speech recognition for Amharic in Section 3, we present the results of our experiments in Sections 4, 5 and 6. Finally, conclusions are drawn and recommendations for future works are derived in Section 7.

2. AMHARIC MORPHOLOGY

Amharic is a member of the Ethio-Semitic languages, which belong to the Semitic branch of the Afro-Asiatic super family [15]. It is related to Hebrew, Arabic, and Syrian. Amharic is a major language spoken mainly in Ethiopia. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as second language throughout different regions of Ethiopia [16].

Like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic 'lexical' meaning. A pattern consists of a set of vowels which are inserted (intercalated) among the consonants of the root to form a stem. The pattern can be combined with a particular prefix or suffix to make a single grammatical form [17] or another stem [18]. For example, the Amharic root *sbr*[1] means 'break', when we intercalate the pattern ä-ä and attach the suffix -ä we get *säbbärä* 'he broke' which is the first form of a verb (3rd person masculine singular in past tense as in other semitic languages) [17]. In addition to this non-concatenative morphological feature, Amharic uses different affixes to form inflectional and derivational word forms.

Some adverbs can be derived from adjectives but, adverbs are not inflected. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. For example, from the noun *lǧ* 'child' another noun *lǧInät* 'childhood'; from the adjective *däg* 'generous' the noun *dägnät* 'generosity'; from the stem *snIf*, the noun *snIfna* 'laziness'; from root *qld*, the noun *qäld* 'joke'; from infinitive verb *mäsbär* 'to break' the noun *mäsbäriya* 'an instrument used for breaking' can be derived.

Case, number, definiteness, and gender marking affixes inflect nouns. Table 1 presents, as an example, the genitive case markers for nouns.

Person	singular		plural
	Vowel ending	Consonant ending	
1^{st}	-ye	-e	-aččn
2^{nd} masculine	-h	-Ih	-aččhu
2^{nd} feminine	-š	-Iš	
2^{nd} polite	-wo	-wo	
3^{rd} masculine	-w	-u	-aččäw
3^{rd} feminine	-wa	-wa	
3^{rd} polite	-aččäw	-aččäw	

Table 1. Genetive Case Markers (Adapted from Titov (1976))

Adjectives are derived from nouns, stems or verbal roots by adding a prefix or a suffix. For example, it is possible to derive *dnIgayama* 'rocky' from the noun *dnIgay* 'rock, stone'; *znIgu* 'forgetful' from the stem *znIg*; *sänäf* 'lazy' from the root *snf* by suffixation and intercalation. Adjectives can also be formed through compounding. For instance, *hodäsäfi* 'tolerant, patient', is derived by compounding the noun *hod* 'stomach' and the adjective *säfi* 'wide'. Like nouns, adjectives are inflected for gender, number, and case [18].

Unlike the other word categories such as noun and adjectives, the derivation of verbs from other parts of speech is not common. The conversion of a root to a basic verb stem requires both intercalation and affixation. For instance, from the root *gdl* 'kill' we obtain the perfective verb stem *gäddäl-* by intercalating pattern ä-ä. From this perfective stem, it is possible to derive the passive stem *tägäddäl-* and the causative stem *asgäddäl-* using prefixes tä- and as-, respectively. Other verb forms are also derived from roots in a similar fashion.

Verbs are inflected for person, gender, number, aspect, tense and mood [18]. Table 2 shows how a perfective Amharic verb inflects for person, gender and number. Other elements like negative markers also inflect verbs in Amharic.

Person	Singular	Plural
1^{st}	säbbärku/hu	säbbärn
2^{nd} masculine	säbbärh/k	säbbäraččhu
2^{nd} feminine	säbbärš	
2^{nd} polite	säbbäru	
3^{rd} masculine	säbbärä	säbbäru
3^{rd} feminine	säbbäräčč	
3^{rd} polite	säbbäru	

Table 2. Inflection of a Perfective Verb

From the above brief description of Amharic word morphology it can be seen that Amharic is a morphologically rich language. It is this feature that makes the OOV problem more serious in Automatic speech recognition system.

3. PREVIOUS WORK

The application of automatic word decomposition (using Harris algorithm) for automatic speech recognition of less-represented languages, specifically Amharic, has been investigated by [12]. In their study, the units obtained through decomposition have been used in both lexical and language models. They reported recognition results for four different configurations: full word and three decomposed forms (detaching both prefix and suffix, prefix only and suffix only). A word error rate (WER) reduction over the base line word-based system has been reported using 2 hours of training data in speech recognition in all decomposed forms although the level of improvement varies. The highest improvement (5.2% absolute WER reduction) has been obtained with the system in which only the prefixes have been detached. When both the prefixes and suffixes have been considered, the improvement in performance is small, namely 2.2%. This might be, as the authors indicate, due to the limited span of the n-gram language models.

Decomposing lexical units with the same algorithm led to worse performance when more training data (35 hours) was used [13]. This can be explained by a higher acoustic confusability. [13] tried to solve this problem by using other modified decomposition algorithms. Their starting algorithm is Morfessor [19] which has been modified by adding different information. They were able to achieve a word error rate reduction only when a phonetic confusion constraint was used

[1] For transcription purposes, IPA representation is used with some modifications.

to block the decomposition of words which would result in acoustically confusable units.

In contrast to [12] and [13], [14] used morphemes only for the language modeling component. They applied a lattice rescoring framework to avoid the influence of acoustic confusability on the performance of the speech recognizer. Lattices have been generated in a single pass recognition using a bigram word-based language model and rescored using sub-word language models. Improvement in the performance of the speech recognition has been obtained. However, this method does not solve the out-of-vocabulary problem since a word-based pronunciation dictionary has been used.

4. WORD-BASED SPEECH RECOGNITION

4.1. The Speech Corpus

The speech corpus used to develop the speech recognition system is an Amharic read speech corpus [20]. It contains 20 hours of training speech collected from 100 speakers who read a total of 10850 sentences (28666 tokens). Compared to other speech corpora that contain hundreds of hours of speech data for training, our models obviously suffer from a lack of training data.

Although the corpus includes four different test sets (5k and 20k both for development and evaluation), for the purpose of the current investigation we have used the 5k development test set, which includes 360 sentences (4106 tokens or 2836 distinct words) read by 20 speakers.

4.2. Acoustic, Lexical and Language Models

The acoustic model consists of 6610 cross-word triphone HMMs each with 3 emitting states. The states of these models and all the cross-word triphone models that are potentially needed for recognition are tied using decision-tree based state-clustering that reduced the number of triphone models from 77658 logical models to 10215 physical ones. Their mixture is added incrementally and 12 Gaussian mixtures have been found to be the optimal.

Vocabulary of the three full-word form pronunciation dictionaries (5k, 20k and 65k) have been prepared by taking the most frequent words from a text corpus consisting of 120262 sentences (2348150 tokens or 211120 types). Table 3 shows the out-of-vocabulary rates of the 5k development test set against these vocabularies. Although we tried to optimize the vocabularies by taking the most frequent words, the OOV rate is still high.

Vocabulary	Token OOV (%)	Type OOV (%)
5k	36.43	51.55
20k	20.41	29.23
65k	9.33	13.36

Table 3. OOV rate of the 5k development test set

In order to minimize the development effort, the pronunciation dictionaries have been encoded by means of a simple procedure that takes advantage of the orthographic representation (a consonant vowel syllable) which is fairly close to the pronunciation in many cases. There are, however, notable differences especially in the area of gemination and insertion of the epenthetic vowel.

The text corpus from which the vocabularies have been selected has also been used to train language models. As we have three dictionaries (5k, 20k and 65k), we have developed three trigram language models one for each vocabulary using the SRILM toolkit [21]. The language models are made open by including a special unknown word token. The modified Kneser-Ney smoothing method has been used to smooth all the language models.

4.3. Performance of Word-based Speech Recognizers

Speech recognition experiment has been performed using the 5k, 20k and the 65k vocabularies. In each case the systems have been evaluated with the 5k development test set. Figure 1 presents the word recognition accuracy for each vocabulary. As it can be seen from the figure, the OOV rate decreases when the vocabulary size increases. As the OOV rate decreases the performance of the speech recognition system increases. The best performance (78.3%) has been obtained for the 65k which has OOV rate of 9.33%. The results show that the OOV rate highly affects the performance of speech recognition systems. To deal with this problem, morphemes instead of words have been considered as dictionary entries and units in language models.

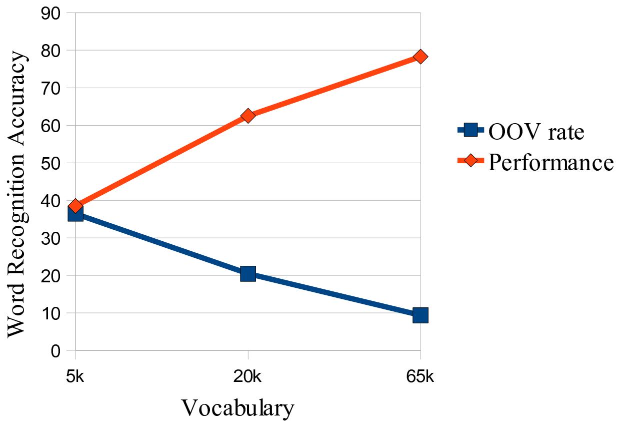


Fig. 1. Word Recognition Accuracy of three Word-based Recognizers.

5. MORPHEME-BASED SPEECH RECOGNITION

5.1. Morphological Segmentation

To use morphemes in speech recognition system a word parser, which splits word forms into their constituents, is needed . Different attempts [22, 23, 24] have been made to develop a morphological analyzer for Amharic using different methods. However, none of the systems can be directly used for our purpose. The systems developed by [22] and [23] suffer from lack of data. The morphological analyzer developed by [24] seems to suffer from a too small lexicon. It has been tested on 207 words and analyzed less than 50% (75 words) of them. Moreover, the output of the system is not directly useful for our study which needs the morphemes themselves instead of their morphological features. Since the source code of the analyzer has not yet been made available, it is not possible to customize it.

An alternative approach is offered by unsupervised corpus-based methods that do not need annotated data. These methods are particularly interesting for resource scarce languages like Amharic. Thus, Morfessor [19] which is a freely available, language independent unsupervised morphology learning tool that tries to identify all the morphemes found in a given word has been used to morphologically segment our text corpus. The morphologically segmented text consists of 15,925 distinct morphs. Figure 2 shows the morph length (in terms of number of characters) distribution of the corpus. As can be observed from the figure, the length of most of the morphs is between four and six characters. In order to facilitate the conversion of morpheme sequences to words, a special word boundary marker has been attached to word boundary morphemes which made the morphemes context-sensitive and consequently increased the number of distinct morphemes to 28,492.

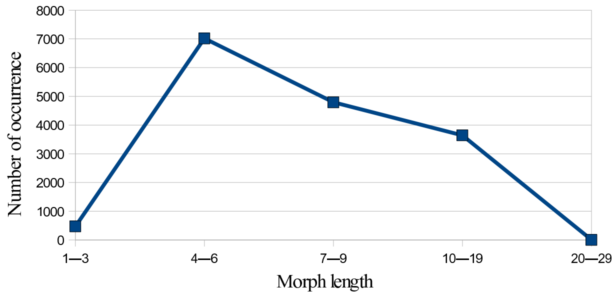


Fig. 2. Morph Length Distribution.

5.2. Acoustic, Lexical and Language Models

The acoustic model has been developed in a similar fashion as for the word-based recognizers. The training data has a set of 6459 cross-morph triphone HMMs each with 3 emitting states. The states of these models and all the possible cross-morph triphone models are tied and, therefore, the number of triphone models is reduced from 57799 logical to 7685 physical models. Similar to the word-based models, 12 Gaussian mixtures have been found to be the optimal.

The entries in the pronunciation dictionary are morphemes. From the morphologically segmented corpus, three dictionaries have been prepared: 5k and 20k by taking the most frequent morphs and 28.4k by considering all the morphemes. The morph-based OOV rates of these vocabularies on the 5k development test set are presented in Table 4 which shows that the OOV rate is highly reduced as a result of using morphs. The token OOV rate for the 5k morph vocabulary, for instance, is only a little bit higher than the token OOV rate of the 65k full-word vocabulary.

Vocabulary	Token OOV (%)	Type OOV (%)
5k	10.75	28.43
20k	0.67	1.83
All (28.4k)	0.03	0.08

Table 4. Morph OOV rate of the 5k development test set

As we have three dictionaries (5k, 20k and 28.4k), we have developed three open vocabulary morph-based trigram language models, one for each vocabulary. Similar to the word-based language models, the morpheme-based ones have also been smoothed using modified Kneser-Ney smoothing technique.

5.3. Performance of Morph-based Speech Recognizers

The morpheme-based speech recognition system has been evaluated on the 5k development test set using the 5k, 20k and 28.4k morph vocabularies. The results are reported in terms of morph recognition accuracy (MRA) and word recognition accuracy (WRA). The word recognition accuracy has been computed after words have been obtained by concatenating the recognized morph sequence. The best performance (see Table 5) has been obtained with the 28.4k morph vocabulary which has an OOV rate of 0.03. Since the OOV rate is very small, an accuracy even higher than the one reported here was expected. The reasons for this disappointing performance (in spite of having a small OOV rate) might be a higher acoustic confusability and the limited language model scope.

Vocabulary	MRA (%)	WRA (%)
5k	55.34	50.04
20k	67.67	62.00
28.4k	68.26	62.78

Table 5. Performance of morph-based speech recognizer

6. COMPARISON OF WORD- AND MORPH-BASED SPEECH RECOGNIZERS

The morph vocabularies have a very low OOV rate compared to the word vocabularies. This has a positive effect on speech recognition accuracy, especially for small vocabularies, namely 5k. The word-based model has a word recognition accuracy of 38.47% when the 5k vocabulary has been used. On the other hand, the morpheme-based system reaches a word recognition accuracy of 50.04% for the 5k morph vocabulary[2], which means an absolute improvement of 11.57%. However, for the 20k the morpheme-based speech recognizer performed slightly worse (62.00%) than the equivalent word-based system which has a word recognition accuracy of 62.51%. The 28.4k vocabulary has morph and word recognition accuracies of 68.26% and 62.78%, respectively. The performance of the recognizer with 28.4k morph vocabulary is only slightly better than the 20k word-based recognizer although it includes all the morphs in the text and has a very low OOV rate. As we have already mentioned, besides the acoustic confusability, the limited scope of the morpheme-based n-gram language model might contribute to the poor performance of the morpheme-based speech recognizer since taking three morphemes might not mean taking three words. This has also been commented by [12] who suggested the use of higher order n-gram models. Thus, higher order morpheme-based language models have been used in our morpheme-based speech recognizers.

We generated lattices using the 20k and 28.4k vocabulary morpheme-based recognizers and rescored the lattices with a quadrogram morpheme-based language model which has been developed in the same manner as the trigram models. The best path transcription decoded from the rescored lattices have morph and word recognition accuracy of 69.70% and 64.46%, respectively for the 28.4k vocabulary and 68.92% and 63.51% for the 20k one (see Table 6). Absolute 1.95% and 1% word recognition accuracy improvement (over the 20k word-based recognizer) have been obtained for the 28.4k and 20k vocabulary morpheme-based recognizers, respectively, as a result of rescoring the lattices with a quadrogram language model.

Vocabulary	MRA (%)	WRA (%)
20k	68.92	63.51
28.4k	69.70	64.46

Table 6. Lattice rescoring with quadrogram morpheme-based language model

As it can be seen from Table 7, rescoring with a pentagram language model did not lead to further improvement. Rather, the morph and word recognition accuracies (for both 20k and 28.4k vocabularies) became worse than the recognizer that used the quadrogram morph-based language model. This might be due to data sparseness. As the language model training corpus is very small many of the pentagrams might not be encountered in the training data and therefore estimated in terms of lower order n-grams. Regarding the language models quality, the pentagram language models did not bring much perplexity improvement (less than 1%) over the quadrogram ones for the 20k and the 28.4k vocabularies. The perplexity gains of the quadrogram language models over the trigram ones are 8.291% and 8.386% for the 20k and 28.4k vocabularies, respectively.

Vocabulary	MRA (%)	WRA (%)
20k	67.69	62.17
28.4k	68.48	63.17

Table 7. Lattice rescoring with pentagram morpheme-based language model

7. CONCLUSIONS AND FURTHER WORK

Speech recognition experiments for Amharic have been conducted to study the effect of OOV words problem for a highly inflectional language and to find out whether the problem can be reduced by using morphemes as lexical and language model units. We did both word-based and morph-based speech recognition experiments. For the word-based systems the OOV rate decreases as the vocabulary size increases and word recognition accuracy increases as the OOV rate decreases. It has also been found that using morphemes as dictionary entries and language model units highly reduces the OOV rate and consequently boosts the word recognition accuracy, especially for small vocabularies (5k). However, as the morph vocabulary grows, the morpheme-based recognizers did not bring notable improvement in word recognition accuracy, which might be due to higher acoustic confusability and a limited language model scope. Rescoring lattices with higher order morpheme-based language model (quadrogram) brought word recognition accuracy improvement.

Although the morpheme-based recognizer benefits from the low OOV rate, it is disadvantaged from the small and acoustically confusable units. Therefore, further improvement can be obtained if care is taken (for instance, using confusion constraints as in [13]) to avoid acoustically confusable units. Moreover, we just concatenated recognized morpheme sequences up to a word boundary marker and no effort has been made to avoid concatenation of illegal morpheme sequences. Attempts in this line may also boost the performance of morpheme-based speech recognizer. For example, rules (such as *ignore the subject marker morph if it comes at the beginning of a morph sequence*) that obstruct the concatenation of illegal morph sequences can be used.

[2] Comparing the morph-based systems directly with the word-based ones may not be fair because they have a higher coverage than word-based systems of the same vocabulary size. On the other hand, the morph-based systems are also dis-favoured by the concatenation of illegal morph-sequences, increasing number of small and acoustically confusable units and a limited language model scope.

8. REFERENCES

[1] P. C. Woodland, C. J. Leggetter, J. J. Odell V. Valtchev, and S. J. Young, “The 1994 HTK large vocabulary speech recognition system,” in *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*, 1995, vol. 1, pp. 73–76.

[2] I. Bazzi, *Modelling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, Massachsetts Institute of Technology, 2002.

[3] M. Gales and P. Woodland, “Recent progress in large vocabulary continuous speech recognition: An htk perspective,” 2006.

[4] P. Geutner, “Using morphology towards better large-vocabulary speech recognition systems,” in *Proceedings of IEEE International on Acoustics, Speech and Signal Processing*, 1995, vol. I, pp. 445–448.

[5] K. Carki, P. Geutner, and T. Schultz, “Turkish lvcsr: towards better speech recognition for agglutinative languages,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1563–1566, 2000.

[6] W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krebc, and J. Psutka, “On large vocabulary continuous speech recognition of highly inflectional language - czech,” in *Proceeding of the European Conference on Speech Communication and Technology*, 2001, pp. 487–489.

[7] E. Whittaker and P. Woodland, “Particle-based language modeling,” in *Proceeding of International Conference on Spoken Language Processing*, 2000, pp. 170–173.

[8] E. W. D. Whittaker, J. M. Van Thong, and P. J. Moreno, “Vocabulary independent speech recognition using particles,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 315–318.

[9] V. Siivola, T. Hirsimki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in *Proceedings of Eurospeech*, 2003, pp. 2293–2296.

[10] T. Hirsimäki, M. Creutz, V. Siivola, and M. Kurimo, “Morphologically motivated language models in speech recognition,” in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005, pp. 121–126.

[11] K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta, “Novel speech recognition models for arabic,” Tech. Rep., Johns-Hopkins University Summer Research Workshop, 2002.

[12] T. Pellegrini and L. Lamel, “Investigating automatic decomposition for asr in less represented languages,” in *Proceedings of INTERSPEECH 2006*, 2006.

[13] T. Pellegrini and L. Lamel, “Using phonetic features in unsupervised word decompounding for asr with application to a less-represented language,” in *Proceedings of INTERSPEECH 2007*, 2007, pp. 1797–1800.

[14] M. Y. Tachbelie, S. T. Abate, and W. Menzel, “Morpheme-based language modeling for amharic speech recognition,” in *Proceedings of the 4th Language and Technology Conference - LTC-09*, 2009, pp. 114–118.

[15] R. M. Voigt, “The classification of central semitic,” *Journal of Semitic Studies*, , no. 32, pp. 1–21, 1987.

[16] Anbessa Teferra and Grover Hudson, *Essentials of Amharic*, Köppe, Köln, 2007.

[17] M.L. Bender, J.D. Bowen, R.L. Cooper, and C.A. Ferguson, *Languages in Ethiopia*, Oxford Univ. Press, London, 1976.

[18] B. Yimam, *yäamarIña säwasäw*, EMPDE, Addis Ababa, 2nd. ed. edition, 2000EC.

[19] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.1,” Tech. Rep. A81, Neural Networks Research Center, Helsinki University of Technology, 2005.

[20] S. T. Abate, W. Menzel, and B. Tafila, “An amharic speech corpus for large vocabulary continuous speech recognition,” in *Proceeding of International Conference on Speech Communication and Technology, Interspeech-2005*, 2005.

[21] A. Stolcke, “SRILM — an extensible language modeling toolkit,” in *Proceedings of International Conference on Spoken Language Processing*, 2002, vol. II, pp. 901–904.

[22] A. Bayou, “Developing automatic word parser for amharic verbs and their derivation,” M.S. thesis, Addis Ababa University, 2000.

[23] T. Bayu, “Automatic morphological analyzer for amharic: An experiment employing unsupervised learning and autosegmental analysis approaches,” M.S. thesis, Addis Ababa University, 2002.

[24] S. Amsalu and D. Gibbon, “Finite state morphology of amharic,” in *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 2005, pp. 47–51.

INITIALIZING ACOUSTIC PHONE MODELS OF UNDER-RESOURCED LANGUAGES: A CASE-STUDY OF LUXEMBOURGISH

Martine Adda-Decker, Lori Lamel & Natalie D. Snoeren

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{madda,lamel,nsnoeren}@limsi.fr

ABSTRACT

The national language of the Grand-Duchy of Luxembourg, Luxembourgish, has often been characterized as one of Europe's under-described and under-resourced languages. In this contribution we report on our ongoing work to take Luxembourgish on board as an e-language : an electronically searchable spoken language. More specifically, we focus on the issue of producing acoustic seed models for Luxembourgish. A phonemic inventory was defined and linked to inventories from major neighboring languages (German, French and English), with the help of the IPA symbol set. Acoustic seed model sets were composed using monolingual German, French or English acoustic model sets and corresponding forced alignment segmentations were compared.

Next a super-set of multilingual acoustic seeds was used putting together the three language-dependent sets. The language-identity of the aligned acoustic models provides information about the overall acoustic adequacy of both the cross-language phonemic correspondances and the acoustic models. Furthermore some information can be gleaned on inter-language distances : the German acoustic models provided the best match with 54.3% of the segments aligned using German seeds, 35.3% using the English ones and only 10.4% using the French acoustic models. Since Luxembourgish is considered a Western Germanic language close to German, this result is in line with its linguistic typology.

Introduction

Luxembourg is a small country in Western Europe, bordered by Belgium, France and Germany (see Figure 1). The national language Luxembourgish ("Lëtzebuergesch") is the language spoken by native Luxembourgers. From a linguistic typological point of view, Luxembourgish belongs to the West central dialects of High German and is therefore part of the Germanic Franconian languages. Just like the English language, Luxembourgish can be considered as a mixed language with strong Romance and Germanic influences. Because of the fact that Luxembourgish is embedded in this multilingual context on the divide between Romance and Germanic cultures, people switch from one language to another fairly easily. Therefore, the linguistic situation in Luxembourg poses a real challenge for researchers concerned with both automatic and human language processing for at least two reasons. First, Luxembourgish is strongly embedded in a multilingual context, entailing frequent code-switching and code-mixing. Luxembourgish hence represent an interesting testbed for multilingual processing [1]. Second, Luxembourgish may be considered as a partially under-resourced language, as the written production remains relatively low. Such languages currently represent a hot topic in the field of automatic speech processing, because of a limited written production of Luxembourgish, a poorly observed writing standardization (as compared to other languages such as English and French), and a large diversity of spoken varieties.

Fig. 1. Geographical situation of Luxembourg in the heart of Western Europe and on the globe.

In the next section we give some more insight into the linguistic situation in Luxemburg, with a focus on the luxembourgophone situation. Section 2 presents the phonemic inventory of Luxembourgish and its link with other major Western languages. Section 3 presents alignment results with sets of monolingual and multilingual seed models. Section 4 summarizes the achieved results and discusses some major future challenges for both speech technologies and linguistic studies of Luxembourgish.

1. LINGUISTIC SITUATION OF LUXEMBOURGISH

1.1. Multilingual context

Luxembourg, a small country of less than 500,000 inhabitants in the center of Western Europe, is composed of about 65% of native inhabitants and 35% of immigrants. The national language Luxembourgish is considered the official language of Luxembourg only since 1984. Luxembourgish is the (Moselle Franconian) language spoken by native Luxembourgers, French and German being easily used for communication among residents [2]. Major languages practiced by immigrants used to be Portuguese and Italian. The immigrated population generally speaks or learns one of Luxembourg's other official languages : French or German. Recently, English has joined the set of prestigious languages of communication, and tends to become a major communication tool in professional environments.

Although the country is often considered a successful example of a multilingual society, the linguistic situation of Luxembourg remains complex. Different reasons contribute to this. First, the small size of the country entails a dependance on neighboring countries (Germany, France, Belgium) with a very high rate of cross-boundary exchanges. Moreover, its historical background and its geographical situation puts Luxembourg at the frontier of the Germanic and Romance worlds. Last but not least, an important proportion of immigrants of different linguistic origins adds to the complex linguistic situation that can be observed in Luxembourg.

1.2. An under-resourced language

As was pointed out by [3] and [4], Luxembourgish should be considered as a partially under-resourced language, mainly because of the fact that the written production remains relatively low and that linguistic knowledge and resources, such as lexica and pronunciation dictionaries, are sparse. Rather surprisingly, written Luxembourgish is not systematically taught to children in primary school : German is usually the first written language learned, followed by French [5]. A number of proposals for standardizing the orthography of Luxembourgish can be traced back to the middle of the 19th century. Recently, a successful standardization eventually emerged from the work of a number of specialists charged with the task of creating a dictionary that was published between 1950 and 1977 [6]. The latest spelling reform [2] has been adopted in 1999, and is being used to create official language resources (Cortina, CPLL dictionary). Nonetheless, up until today, German and French are the most practiced languages for written communication and administrative purposes in Luxembourg, guaranteeing a larger dissemination, whereas Luxembourgish is mainly being used for oral communication. It is precisely because of the strong influence of both German and French that Luxembourgish exhibits a large amount of both pronunciation and derived potential writing variants. Pronunciation variants may give rise to resulting variations in written Luxembourgish, as Luxembourgish orthography strives for phonetic accuracy [2]. The question then arises, in particular for oral transcripts, whether the written form reflects the perceived pronunciation form or whether some sort of normalization process is at work that eliminates part of the variation. With respect to automatic speech recognition, text normalization is an important issue in order to achieve reliable estimates for n-gram based language models. In sum, Luxembourgish is predominantly a spoken language that tends to reproduce the observed variations when written.

The limited production of written material is related to the easy use of French and German as written communication languages. Further, no orthographic standards were clearly established before the end of the 20th century. This implies a high degree of variation in the observed written forms. An exhaustive Luxembourgish dictionary was produced after World War II, and this large scale effort actively contributed to the elaboration of spelling standards settled in 1975 and revised in 1999) [7, 8]. Written Luxembourgish sources, although not very widespread, can yet be found over the last decades and even centuries. It is difficult to estimate the numbers of Romance/Germanic influenced words in Luxembourgish, as proportions greatly depend on communicative settings. Nonetheless, one may note that vernacular Luxembourgish is highly influenced by its Germanic filiation, whereas more technical and administrative jargons include a higher proportion of Romance words. Examples in Table 2 are almost all of Germanic influence, except those illustrating nasal vowels, and the /ʒ/and /ɥ/ consonants.

Beyond written material, the existence of sibling resources, providing similar content in both written and audio modalities are particularly helpful for automatic speech recognition (ASR). Steps to an autonomous ASR system include acoustic modeling, the development of a pronunciation dictionary and language modeling [9]. Most languages make use of broadcast news audio data, together with, as written sources, newspaper texts, news wires and related web pages. In Luxembourg news broadcasts are proposed in Luxembourgish on a daily basis, however newspapers remain for the most part bilingual German/French, with only limited code-switching and code-mixing to Luxembourgish, generally for titles. Yet, it is important to highlight recent efforts that have been made regarding the establishment of word lists and multilingual dictionaries in electronic form [10]. Furthermore concerning the WEB, Luxembourgish actually holds rank 55 in the list of 272 official Wikipedias, published by the Wikimedia Foundation for various languages (http ://meta.wikimedia.org/wiki/List_of_Wikipedias). The number of Luxembourgish native speakers can be estimated to 300,000. The immigrated population and the number of daily cross-boarder commuters has steadily increased over the past decades. A relatively high number of more or less proficient

L2 speakers can be found among them, especially as they express a great interest in learning the basics of the Luxembourgish language.

1.3. Luxembourgish corpora

As was mentioned before, sibling resources, providing both audio and related written material are of major interest for ASR development. The most relevant resource we found here, consists in the *Chamber* (House of Parliament) debates and to some extend in news channels, such as delivered by the Luxembourgish radio and television broadcast company RTL.

The Parliament debates are broadcast and made available on the official web site (www.chd.lu), together with written reports (the *Chamber* reports), which correspond to rather close manual transcripts of the oral debates. Another interesting sibling resource stems from the Luxembourgish radio and television broadcast company RTL, which produces news written in Luxembourgish on its web site (www.rtl.lu), together with the corresponding audio data. However only very limited amounts of written Luxembourgish can be found here, whereas RTL has a profuse audio/video production. Table 1 summarizes the different text and audio resources currently being collected. 12M words have been extracted from the *Chamber* reports (years 2002-2008), which mainly comprise professionally transcribed oral debates. However they also include some written subjects in French. The collected audio data correspond to the debates of the two most recent years, totalling a volume of approximately two hundred hours.

Table 1. Major Luxembourgish text and audio sources for ASR. Collected amounts are given in number of words

	written	**sibling : audio+written**	
Source :	WIKIPEDIA	CHAMBER	RTL
	lb.wikipedia.org	www.chd.lu	www.rtl.lu
Volume :	500k	12M/(300h)	700k/(40h)
Years	2008	2002-2008	2007-2008

2. PHONEMIC INVENTORY

The word lists derived from the written material allow to fix optimal vocabularies for the ASR system. A further step consists in providing pronunciations for each lexical entry. Such pronunciations rely on a phonemic inventory. Hereafter we will give details about the Luxembourgish phonemic inventory, detailing vowels, diphthongs and consonants.

The adopted Luxembourgish phonemic inventory includes a total of 60 phonemic symbols plus 3 extra-phonemic symbols (for silence, breath and hesitations). Table 2 present the selected phonemic inventory together with illustrating examples. Luxembourgish is characterized by a particularly high number of diphthongs. To minimize the phonemic inventory size, we could have chosen to code diphthongs using two consecutive symbols, one for the nucleus and one for the offglide (e.g. the sequence /a/ and /j/ for diphthong a͜ɪ). We prefered the option of coding diphthongs and affricates using specific unique symbols. Given the importance of French imports, nasal vowels, although not required for typical Luxembourgish words, were included into the inventory. Furthermore, the native Luxembourgish makes use of a rather complex set of voiced/unvoiced fricatives.

Concerning linguistic studies [11], many aspects of the Luxembourgish language have been explored on limited spoken material. They still need to be investigated on a larger scale and on fluent speech, in particular for pronunciation variants. The existing phonetic, phonological, prosodic, lexical and morphosyntactic studies are generally carried out using limited objective observations. Large oral corpus-based studies might be carried out, provided Luxembourgish automatic speech alignment and transcription systems were available.

In the following, we raise some issues concerning high-quality pronunciation dictionaries.

2.1. Spelling

Luxembourgish spelling standards aim at minimizing pronunciation ambiguities, even though minor problems remain. For example, the `au` letter sequence is ambiguous with respect to /ɛʊ/ (`Haut`) or /ʌʊ/ (`haut`) pronunciations.

Concerning Romance or Germanic origins of Luxembourgish lexical entries, writing standards may stay more or less close to the language of origin, as discussed in section 1.1. For French words such as `attaquer` (eng. to attack) or `abdiquer` (eng. to abdicate), the corresponding lëtzebuergesch orthographic forms are `attackéieren` and `abdiquéieren` (after the official Luxembourgish CORTINA spellchecker [1]). For Romance items, different pronunciation rule sets need to be developed, that differ from Germanic or Moselle-Franconian pronunciation rules. For instance, depending on the origin, `qu` letter sequence of germanic items such as `quälen, quëtschen, Quetschen` calls for a /kw/ pronunciation, whereas Romance rules generally demand a simple /k/ pronunciation.

3. ALIGNMENT EXPERIMENTS

Alignment experiments are carried out using different initializations for the Luxembourgish acoustic models and different pronunciation dictionaries.

3.1. Acoustic seed models

Many researches have addressed the issue of building acoustic seed models for underresourced languages [12].

[1] More information about the Cortina Luxembourgish spellchecker can be found at http ://cortina.lippmann.lu.

In this work three sets of context-independent and gender-independent acoustic models were built, one for each seed language (i.e., English, French, German). The models were trained on manually transcribed audio data (between 40 and 150 hours) from a variety of sources, using language specific phone sets. The amount of data used to train the acoustic models and the number of phonemes per language are given in Table 3. Each phone model is a tied-state left-to-right, 3-state CDHMM with Gaussian mixture observation densities (typically 32 components). The acoustic features are derived from a PLP-like [13] acoustic parametrization, which has been used in the LIMSI systems since 1996. The speech features consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. For each 30ms frame, the Mel scale power spectrum was computed, and the cubic root taken, followed by an inverse Fourier transform. LPC-based cepstrum coefficients were then computed. These cepstral coefficients were normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Table 3. Characteristics of English, French and German original acoustic model sets.

Language	*#phonemes*	*#training (h)*
English	48	150
French	37	150
German	49	40

Four sets of pseudo-Luxembourgish acoustic models, each including 63 phones, were created from the English, French and German seed models by mapping the Luxembourgish phonemes to a close equivalent in each of the three model sets. Table 2 include the adopted cross-lingual associations, to initialze seed models for Luxembourgish. It can be noted that some symbols are used several times for different Luxembourgish phonemes. In particular, for the diphthongs, which are missing in French, we chose to select the phonemes corresponding to the nucleus vowel. A fourth model set was constructed by concatentating the first three model sets, so that the decoder could chose among the three languages' models (see Table 4). For each word, choose the acoustic models from the language with the best match.

3.2. Multilingual pronunciation variants

For the alignment experiments using the language-dependent phone sets the same pronunciation dictionary was used. We introduced some variants for the most frequent function words, French imports and some variants to account for word-final mobile-n deletion (or insertion) [4]. Example variants are shown in Table 5.

Table 4. Pseudo-Luxembourgish acoustic models using either English, French and German acoustic model sets or a super-set of multilingual acoustic seeds.

Language	*#phonemes*	*#training (h)*
English	63	150
French	63	150
German	63	40
Super-set (E,F,G)	3x63	340

Table 5. Excerpt of the Luxembourgish pronunciation dictionary as used for the proposed alignments. The upper part shows typical examples of variants (frequent words, French loan words, mobile-n deletion). The lower part illustrates the pronunciation dictionary used for alignments with the multilingual acoustic super-set.

lexical entry (English)	*citation form*	variants
déi (those)	de͜ɪ	dɪ
President (president)	pʁəzidɛnt	pʁezidã
Europa	ɔ͜ɪʁopa	øʁopa
an (and, in)	ʌn	ʌ
	Multilingual dictionary	
déi (those)	d$_g$e͜ɪ$_g$ d$_f$e͜ɪ$_f$ d$_e$e͜ɪ$_e$	d$_g$ɪ$_g$ d$_f$ɪ$_f$ d$_e$ɪ$_e$

3.3. Luxembourgish audio alignment

The Luxembourgish audio corpus with corresponding detailed acoustic transcripts comprised 80 minutes of hand transcribed audio data (Chamber (70') and News (10')). We produced these detailed transcripts from scratch for the news data. For the Chamber data, the audio stream was manually segmented into speaker turns, according to the existing *bona fide* report. For each speaker, the *bona fide* transcriptions were changed if necessary to faithfully reflect the speech flow. All uttered audible speech events, including disfluencies and speech errors were manually transcribed. The quality of the manual *verbatim* transcripts were checked via the resulting word lists for typos and orthographic inconsistencies. The transcript quality further needs to be questioned, if significant amounts of data are rejected during alignment. As the same transcripts were used for the different Luxembourgish acoustic seed models, if more data are rejected for a given model set than for the others, this set may be considered as less appropriate, without blaming the transcripts.

The percentage of the audio data aligned with phone segments varies from 77-80%, the lowest figure corresponding

Table 6. Total duration (in seconds) aligned as phones, as extra phonemic segments (silence, breath or hesitation) or rejected due to model/data mismatch.

Language	*phon.dur.*	*#extra dur.*	rejected
English	3910	673	516
French	3933	790	373
German	4043	921	131
Super-set (E,F,G)	4077	814	203

to English, the highest to the multilingual and German configurations. The remaining 20-23% of the acoustic data are either aligned with extra-phonemic symbols or rejected by the alignment system, due to model/data mismatch. It can be noted that English has the highest rate of rejected data : 516 seconds which correspond to 10% of the data. Such a high rejection rate normally would require to check the manual transcripts and/or the pronunciation dictionary. Fortunately, for the other configurations, the rejection rates are much lower, the lowest rates being achieved by the German language (131 seconds, $< 3\%$). German has the highest contribution to the extra-phonemic symbol set.

The average phone segment duration remains almost stable with respect to the different monolingual seed alignments. Variations here stem from variable proportions of the acoustic signal assigned to the extra-phonemic models. The German alignment yields the smallest average phone duration of 0.07 seconds on average (silence, breath and hesitation segments are not considered). For English and French the average segment duration corresponds to 0.08 seconds. We could observe that independently of acoustic-phonetic considerations, the (German) silence (including background noise) model was made use of more frequently during the German monolingual alignment, than was the case for the French or English silence models. This explains the smaller average phone duration. This might be related to the relatively small volume of training data (40h) for the German originated seeds (as opposed to French and English), with a lower capacity to cover various acoustic conditions.

The results presented in Table 6 further suggest that the German acoustic models are globally best at explaining the Luxembourgish data, as the smallest volume of data was rejected.

On a more linguistic level of analysis, the results show that unvoiced segments tend to be longer than their voiced counterparts, and that diphthongs and nasal vowels are about 30% longer than oral vowels. More precise results on the Luxembourgish phonemes will be produced in the future, with acoustic models trained on a larger set of Luxembourgish data.

3.4. Multilingual alignments

The alignment produced by the acoustic super-set model, together with the multilingual pronunciation dictionary achieves the highest proportion of aligned acoustic phone segments. In this configuration, it is interesting to investigate the results on two levels : (i) on the phone segment level, we can measure the proportions of segments aligned using the seeds of a given language. Are there differences in proportions as a function of phonemes ? (ii) on the word level, we may check whether the proportion of aligned French seeds is higher for French loan words than for native Luxembourgish words.

For example, we may expect that for Luxembourgish diphthong segments, the proportion of aligned English seeds may increase, especially for diphthongs not covered by the German language. Conversely the proportion of French and English seeds used for Luxembourgish and German specific sounds (e.g. χ) should remain very low.

Table 7 displays aligned monolingual seed proportions as produced by the multilingual super-set. More than half of the 55873 segments were aligned using the German seeds. About one third corresponds to English seed models and only 10% of the segments were aligned using the French models. Results for some phonemes are shown to illustrate that proportions can notably vary with phoneme identity.

Table 7. Proportions of aligned German, English, French seeds in the multilingual super-set configuration. The number of phone occurrences is provided. Results are given on average and a subset of selected phonemes.

Phone type	*German*	*English*	*French*	# occ.
overall	54.3	35.3	10.4	55873
p	67.05	21.85	11.10	865
t	55.91	35.23	8.86	3588
k	55.15	36.64	8.21	1048
ç	56.80	34.52	8.67	588
χ	80.87	14.29	4.84	413
h	36.05	59.36	4.59	785
ʒ	41.96	25.00	33.04	112
y	25.00	15.62	59.38	32
ʏ	41.03	25.64	33.33	39

4. SUMMARY AND PROSPECTS

The main goal of the present contribution was to draw attention to the complex linguistic situation of Luxembourgish, a partially under-resourced and under-described language. For ASR development, the use of sibling resources that provide similar contents in both written and oral/auditory modalities is extremely useful. Although there are relatively few written resources in Luxembourgish as compared to other European languages, corpus studies in Luxembourgish will substantially add to the current debate on the processing of pronunciation variants in automatic and natural speech processing.

In the present work, we focused on the issue of producing acoustic seed models for Luxembourgish. A phonemic inventory was defined and linked to inventories from major neighboring languages (German, French and English), with the help of the IPA symbol set. For each of these languages, acoustic seed models were composed using either monolingual German, French or English acoustic model sets. During Luxembourgish speech alignments, a super-set of multilingual acoustic seeds was used putting together the three language-dependent sets. The language-identity of the aligned acoustic models provides information about the overall acoustic adequacy of both the cross-language phonemic correspondances and the acoustic models. Furthermore some information can be gleaned on inter-language distances : the German acoustic models provided the best match with 54.3% of the segments aligned using German seeds, 35.3% using the English ones and only 10.4% using the French acoustic models. Since Luxembourgish is considered a Western Germanic language close to German, this result is in line with its linguistic typology.

Computational ASR investigations and corpus-based analyses will not only enhance the development of a more full-fledged ASR system for Luxembourgish, but can also be used to generate more specific predictions about the role of the actual experience that listeners have with pronunciation variants. In turn their predictions can then be tested in other domains such as psycholinguistics. Given the implications of large corpus-based analyses, it is hoped that this line of research on Luxembourgish will sparkle more interest for this language in researchers working in the domains of ASR and linguistics.

Acknowledgements

The authors would like to thank the members of the CPLL (Permanent Council of the Luxembourgish Language) for their help with sibling corpora and fruitful discussions. This work has been partially financed by OSEO under the QUAERO program. Part of the research has been supported by an FNR grant from the Luxembourgish gouvernment awarded to N.D.S.

5. REFERENCES

[1] M. Adda-Decker and L. Lamel, *Multilingual pronunciation dictionaries in Multilingual Speech Processing*, Elsevier, 2006.

[2] F. Schanen, *Parlons Luxembourgeois*, L'Harmattan, 2004.

[3] M. Adda-Decker, T. Pellegrini, E. Bilinski, and G. Adda, "Developments of letzebuergesch resources for automatic speech processing and linguistic studies.," in *LREC*, 2008.

[4] C. Krummes, "Sinn si or si si ? mobile-n deletion in luxembourgish," in *Papers in Linguistics from the University of Manchester : Proceedings of the 15th Postgraduate Conference in Linguistics*, Manchester, 2006.

[5] C. Berg and C. Weis, "Sociologie de l'enseignement des langues dans un environnement multilingue. rapport national en vue de l'élaboration du profil des politiques linguistiques éducatives luxembourgeoises," Tech. Rep., 2005.

[6] P. Linden, *Luxemburger Wörterbuch*, P. Linden, Hofbuchdrucker, 1950.

[7] G. Newton, *Studies from the Germanic Languages - The standardization of Luxembourgish*, John Benjamins Publishing Company, 2002.

[8] F. Schanen and J. Lulling, "Introduction à l'orthographe luxembourgeoise.," in *www.cpll.lu/ortholuxs_l.html*, G.-D. de Luxembourg, 2003.

[9] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–229, 2002.

[10] J. Lulling, *Luxdico - dictionnaire bilingue luxembourgeois - français*, Presses universitaires de Namur, 2005.

[11] C. Moulin, *Perspektiven einer linguistischen Luxemburgistik - Studien zu Diachronie und Synchronie*, Universitätsverlag WINTER Heidelberg, 2005.

[12] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," in *Proceedings of Eurospeech*, Aalborg, 2001.

[13] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech.," vol. 87, no. 4, pp. 1738–1752, 1990.

Table 2. Cross-lingual phone association table. Luxembourgish target phonemes are associated to identical or similar (in grey) phonemes of the different French, German, English source languages.

Carrier word (Eng.)	Lux.	Fre	Ger	Eng
ORAL VOWELS				
l**ii**cht (light)	i	i	i	i
L**i**dd (song)	ɪ	i	ɪ	ɪ
S**ü**den (south)	y	y	y	i
sch**ü**tzen (shelter)	ʏ	y	ʏ	ɪ
L**ee**d (sorrow)	e	e	e	e
z**é**ng (ten)	ᴇ	e	e	ɛ
f**ä**eg (able)	ɛ :	ɛ	ɛ :	ɛ
h**ä**tt (would)	ɛ	ɛ	ɛ	ɛ
F**öh**n (hairdryer)	ø	ø	ø	ɔ
m**ë**ll (soft)	œ	œ	œ	ə
et (it)	ə	ə	ə	ə
h**a**t (had)	a	a	a	ɑ
h**a**tt (she)	ᴀ	a	ᴀ	ɑ
R**o**t (advice)	o	o	o	o
L**o**ft (air)	ɔ	ɔ	ɔ	ɔ
L**uu**cht (lamp)	u	u	u	u
H**u**tt (hat)	ʊ	u	ʊ	ʊ
NASAL VOWELS : French imports				
enf**in**	ɛ̃	ɛ̃	ɛ	æ
enfin	ã	ã	a	ʌ
b**on**	õ	õ	o	o
DIPHTHONGS				
liewen (to live)	ɪ͜ə	i	i	i
léien (to tell lies)	e͜ɪ	e	e	e
läit ((he) lies down)	ɛ͜ɪ	e	e	e
lauschteren (to listen)	ɛ͜ʊ	ɛ	ɛ	æ
leien (to lie down)	a͜ɪ	a	a͜ɪ	a͜ɪ
lauden (to ring)	a͜ʊ	a	a͜ʊ	a͜ʊ
Europa	ɔ͜ɪ	ɔ	ɔ͜ɪ	ɔ͜ɪ
lounen (to hire)	ɔ͜ʊ	o	o	o
luewen (to praise)	ʊ͜ə	u	ʊ	ʊ
SYLLABICS				
Kann**er** (children)	ɐ	ə	ɐ	ʌ
fein**em** (fine)	m̩	m	m̩	m̩
laf**en** (to run)	n̩	n	n̩	n̩
eid**el** (empty)	l̩	l	l̩	l̩

Carrier word (Eng.)	Lux.	Fre	Ger	Eng
PLOSIVES				
paken (to package)	p	p	p	p
taasch**t**en (to touch)	t	t	t	t
kachen (to cook)	k	k	k	k
baken (to bake)	b	b	b	b
droen (to carry)	d	d	d	d
goen (to go)	g	g	g	g
FRICATIVES & AFFRICATES				
Feier (fire)	f	f	f	f
lue**s** (slow)	s	s	s	s
Zuch (train)	ʦ	s	s	s
Schoul (school)	ʃ	ʃ	ʃ	ʃ
Ee**ch**en	ç	ʃ	ç	ʃ
Zu**ch** (train)	χ	k	χ	k
Hand (hand)	h	{br}	h	h
Wieder (weather)	v	v	v	v
Summer (summer)	z	z	z	z
Gilet (vest)	ʒ	ʒ	ʒ	ʒ
Li**g**en (lie)	ʝ	ʒ	ç	ʒ
NASALS & GLIDES				
Mamm (mother)	m	m	m	m
Noper (neighbour)	n	n	n	n
mé**ng** (mine)	ŋ	n	ŋ	ŋ
Leit (people)	l	l	l	l
Rou (rest)	ʁ	ʁ	ʁ	ɹ
Hä**r** (mister)	ɐ̯	ə	ɐ̯	ə
Suite (suite)	ɥ	ɥ	ʏ	w
Juli (July)	j	j	j	j
Quetsch (plum)	w	w	ʊ	w
EXTRA-PHONEMIC SYMBOLS				
silence	{sil}	{sil}	{sil}	{sil}
hesitation	{hes}	{hes}	{hes}	{hes}
breath	{br}	{br}	{br}	{br}

MOTÀMOT PROJECT: BUILDING A MULTILINGUAL LEXICAL SYSTEM VIA BILINGUAL DICTIONARIES

Mathieu MANGEOT, Sereysethy TOUCH

Laboratoire GETALP-LIG 385 rue de la bibliothèque BP 53,
F-38041 GRENOBLE CEDEX 9, France

ABSTRACT

The MotAMot project aims to develop of a multilingual lexical network focused on languages of Southeast Asia and especially Vietnamese and Khmer. The macrostructure is a pivot structure with a monolingual volume for every language and a pivot one connecting each word sense of each monolingual volume. The microstructure is based on the explanatory and combinatorial lexicography. Contributions will be made online on the Jibiki platform by a community of volunteers constituted around serious games lexical. Each entry will be given a level of quality, as well as for each contributor.

Index Terms--- multilingual lexicography, under resourced languages, contributive project, MotÀMot project.

1. INTRODUCTION

Economic issues related to technical processing Information is very important. The development of such technology is a key asset for developing countries such as Cambodia and Laos, or emerging ones such as Vietnam, Malaysia and Thailand.

As indicated by V. Berment in his thesis [1], "Development of personal computers and networks make are now necessary to write and communicate in the same way as paper and printing were before. Word processing, emails, or even more advanced systems such as dictation software or speech synthesis are now widespread tools. It is then necessary to consider that computer programs must be added to the traditional tools otherwise the targeted goals can not be achieved any more. Computerization of a language has and an essential place in this broad context."

However, among the 6,000 languages spoken around the world, only a handful of them reach a satisfactory "Level of computerization". To quantitatively assess the degree of computerization of a language, V. Berment proposes the following protocol: to each service or resource, a group of users representative of the language speakers assign a level of criticality Ck and a score Nk. The average weighted scores - called index - reflects their overall satisfaction. A poorly equipped language can be defined as a language whose index is less than 10/20. For example, the Khmer language, spoken in Cambodia obtains 6.5/20, and the Vietnamese language 10/20.

This is mostly because the services related to the treatment of oral (or speech technologies, ie speech synthesis and word recognition) are not yet available for these two languages. It is also the case for a majority of languages in the world some of which are spoken by several tens of million speakers (for example, Bengali: 189 million, Tamil: 63 million), including within Europe countries (Lithuanian, Latvian, Polish ...)!

2. PRESENT STATE OF BILINGUAL LEXICOGRAPHY

The main difficulty at present for bilingual lexicography is the prohibitive construction cost for large amounts of data. For example, the Electronic Dictionary Research project (EDR) whose aim was to build a Japanese-English dictionary required more than 1,200 men-years of work. Its selling price of approximately € 84,000 is far below the actual costs of construction, costs that will probably never be reached.

Anyway, these costs are too high for an individual. Thus, only institutions can acquire such a resource. Moreover, data provided at this price is used by some machine translation systems based on specific techniques.

Faced with these costs difficult to manage, publishing houses end up living on their laurels and do mainly propose new editions of existing dictionaries. Few publishers have the courage to embark on the implementation of a new high quality bilingual dictionary from scratch.

Moreover, even in the most complete dictionaries, there is almost always a lack of information especially on collocations. The few resources that take them into account do it not systematically.

Despite the advent of the Internet, there are currently few lexical resources available freely online in a good quality. Most are in fact small bilingual lexicons made by volunteers not specialists in lexicography.

Multilingual lexicography as such is still in its infancy. Indeed, there is no really a way to print a true "multilingual dictionary". However, it is possible to find multilingual terminological databases (like Iate) or of small lexicons or multilingual phrases books.

It has also not been sufficiently proven that reusing a dictionary of a language couple A➔B in order to build two other language B➔language C and language A ➔language C was really advantageous. So this is what we would like to tackle in this project.

3. GOALS OF THE PROJECT

With the overall objective to participate in the computerization of under-resourced languages, this project aims to develop a lexical system in multiple languages by simultaneously building several bilingual dictionaries sharing at least one language between them. The construction of the bilingual dictionaries will be online on a Papillon-like site built on the Jibiki platform with a collaborative and volunteer based work like Wikipedia.

The bilingual links created during the edition of the entries are used first to generate bilingual reverse links, and second to create new interlingual links.

The three main objectives of this project are the launching of a new contribution dynamic around the construction of each bilingual dictionary involved - the success of Wikipedia shows that it is possible, provided that you have simple and easy to use tools -; moving laboratory experiments such as the DiCo database [2] or the PARAX system[3] to a large-scale and finally developing a testbed for validation of several assumptions made in previous work:

- Bijectivity of bilingual links and transitivity of interlingual ones;
- Massive contribution on the Web;
- Construction of a multilingual lexical system [4].

4. PROGRESS IN THE CONSTRUCTION OF ONLINE RESOURCES

4.1. On the architecture of multilingual resources: the Papillon project

A perfect solution, the holy grail of lexical resources, would be a multilingual lexical database with a pivot structure, of good quality and wide coverage with rich monolingual entries and interlingual links, used both by humans and machines, editable online and freely available. We launched in 2000 the Papillon project[1] in order to advance in this direction.

The macrostructure consists of a monolingual volume for each language and a pivot volume in the center (see Figure 4.1). When a new entry in a language A is added, it must be connected to the interlingual volume. These links are created either by reusing existing bilingual dictionaries of language A ➔language B, either by adding them manually from a translation. Link language A ➔language B becomes language A ➔pivot ➔language B. If the language B entry is already connected to another entry of language C, then language A entry also benefit from these links.

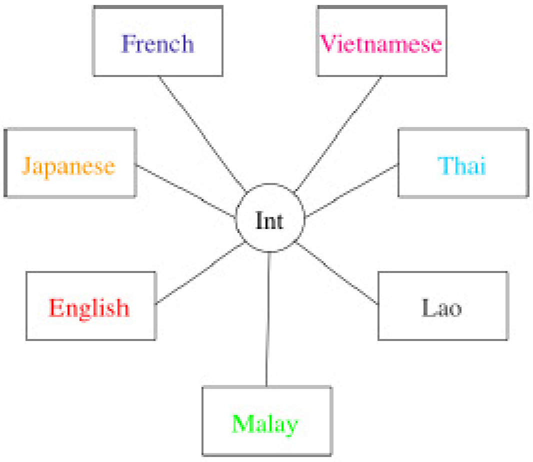


Fig. 1. Papillon Pivot Macrostructure

The microstructure of the monolingual entries is rich and very detailed. It is based on the structure used for the DiCo lexical database[5] from OLST, University of Montreal. The encoding method is borrowed directly from the explanatory and combinatorial lexicology part of the Meaning-Text Theory (MTT). This theory gives the information needed to go from a meaning to its realizations in a given language. The microstructure of the dictionaries is independent of the languages and can theoretically be used by humans and machines.

Each entry or lexical unit is based on the word sense or lexie. It consists of a name, grammatical properties, a semantic formula that can be seen as a formal definition - in the case of a predicative word sense, the formula describes the predicate and its arguments - and then a list of lexico-semantic functions - there are 56 basic functions applicable to any language and they can be combined between them - a list of examples and a list of idiomatic expressions.

The Papillon project specifications are directly inspired by this famous holy grail. But like any ambitious project, it cannot be acheived in one shot. Within time, the project Papillon has become a kind of framework or meta-project [6] with several derivatives projects, each one corresponding to a particular aspect of our initial goal. As we will detail, the tools and systems aspects are covered by the Jibiki project and data collection by the JeuxDeMots project.

4.2. On the contributing aspects: Wikipedia and Wiktionary projects

The online contributory encyclopedia Wikipedia encountered a large and unquestionable success. We could expect a similar success for its little brother Wiktionary but it is yet to go (1.5 million entries for French and only 44,000 for Japanese). Wiktionaryz, who claimed to solve wiktionary misconception problems has not yet achieved its goal. Wiktionary is anyway

[1] http://www.papillon-dictionary.org

not truly a bilingual dictionary even if there are some translation links (for example. indications of the translations context is missing).

One hypothesis that could explain this problem is the motivation. Indeed, when a person contributes to a Wikipedia article, it is rewarded by the fame. It will then be recognized as an expert in its field. It is not possible with a dictionary. The contributions are located on small parts of information and are therefore anonymous.

On the other hand, there is a technical aspect related to the structure. An encyclopedia article has a more or less free structure while a dictionary entry must follow a very specific one (catchword, grammatical information, semantic blocks, translation block, blocks of examples, etc..). It is not possible to reuse a wiki platform for build a dictionary with a well defined structure.

Once accepted the idea that writing entries dictionary is not as fun as working on a Wikipedia article, we must find solutions to motivate a community of volunteers to contribute to a dictionary. Serious lexical games are a first track. We should also highlight contributors suing for example an array of top contributors of the month. And finally, using community networks such as Facebook should also grist to the mill.

4.3. On the data collection via serious games: the JeuxDeMots project

The JeuxDeMots game[7] aims at building a rich and evolving lexical network, that could be compared to a certain extent to the famous WordNet [8] database.

The principle is the following: a game needs two players. When player A initiates a game, an instruction is displayed concerning a type of competency corresponding to a lexical relation (synonym, antonym, domain, intensifier, etc.) and a word W is chosen randomly in the database. Player A has then a limited amount of time for giving propositions that answer the instruction applied to the word W.

The same word W with the same instruction is proposed to another player B and the process is the same. The two half-games of player A and player B are not simultaneous but asynchronous. For each common answer in A and B propositions, the two players earn a certain amount of points and credits. For the word W, the common answers of A and B players are entered into the database. This process participates to the construction of a lexical network linking terms with typed and weighted relations, validated by pairs of players. The relations are typed by the instructions given to the players and weighted with the number of pair players that proposed them. The first version of the French game was launched in July 2007.

4.4. On the technical aspects: the platform Jibiki

Jibiki [9] is a generic online platform for manipulating lexical resources with users and groups management. It is a community website developed initially for the Papillon Project. The platform is programmed entirely in Java, based on the "Enhydra" environment. All the data is stored in XML format in a database (Postgres). This website offers two main services: a unified interface to access simultaneously to many heterogeneous resources (monolingual dictionaries bilingual. multilingual databases, etc..) and an editing interface in order to contribute directly to dictionaries available on the platform.

The editor is based on a HTML template interface instantiated with the entry one wants to edit. The template can be generated automatically from a description of the entry structure using an XML schema. It may then be modified to improve screen rendering. It is possible to edit any type of dictionary provided that it is encoded in XML.

Several construction projects of lexical resources have used or still use this platform with success, like the GDEF project about a Estonian-French bilingual dictionary or the LexALP terminological database. The code for this platform is open source and available for download from the LIG laboratory forge[2].

5. DESCRIPTION OF THE RESOURCE TO BUILD

5.1. Microstructure of entries based on the Meaning-Text Theory

The microstructure of the entries composing the volumes monolingual is a simplification of the Papillon project one. This time, the entry is based on a whole word. A word is either a combination of lexical items (word meanings) or an idiomatic expression. To cope with different skill levels of contributors, the editing interface can adapt itsel and show an adapted granularity of information. For example, a beginner contributor will be invited to give a simple gloss to in order to characterize a word sense, while an expert linguist will describe the entire semantic formula.

5.2. Pivot macrostructure via bilingual interfaces

The macrostructure is also derived from the Papillon Project with a monolingual volume for each language and pivot volume in the center. However, in order not to confuse users, they will contribute via an interface with a classical view of bilingual dictionary.

Each bilingual link language A ➔language B added via this interface will actually be translated in the background by the creating two interlingual links and a pivot entry representing the initial translation link. Finally the following schema will be obtained: language A ➔pivot entry ➔language B.

5.3. Creating bilingual and interlingual links

If a contributor wants to add a translation link between a word Wa in language A and a word Wb in language B, s/he can establish this link at different levels.

[2] http://jibiki.ligforge.imag.fr

The ideal solution is to connect a word sense Sa of the word Wa to another word sense Sb of the word Wb. In this case, the link is bijective and Sb is also connected to Sa (see Figure 5.3).

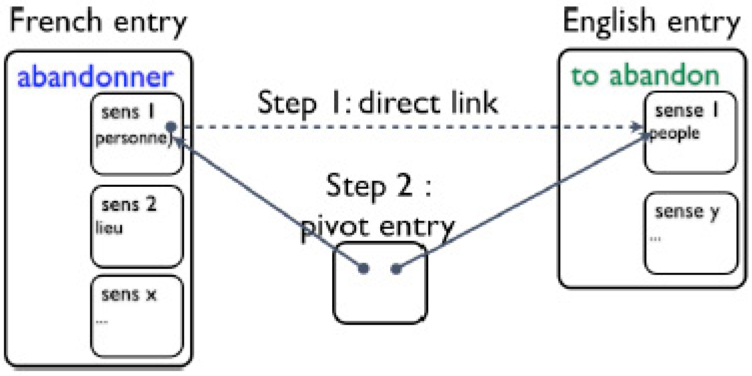


Fig. 2. Word Sense Linking Process

If the word Wb does not have any precise word sense or If the contributor is not able to choose the correct one, it can connect directly to the word Wb. In this case, a new word sense Sb is created with a draft quality level and the link and the word sense are labeled as to refine.

In the case of reusing existing data, it is often impossible to relate information to a specific word sense. In this case, we add at the end of the word Wa, the information that one of Wa word senses may be linked to one of the word senses of Wb, but This information will not be added to Wb. It is be of course tagged as to refine as soon as possible!

With the pivot macrostructure, if two links language A ➔language B and language B ➔language C exist, then it will automatically created a link language A ➔language C which quality level will be marked as draft and revise.

5.4. Data quality and contributors levels

Each part of information for each entry will be assigned a level of quality. The levels range from 1 star for a draft (when the reused data quality is not known) to 5 stars quality certified by an expert (eg, a link translation validated by a certified translator).

Similarly, contributors will be assigned a proficiency level (1 to 5 stars also). 1 star is the level of a beginner unknown in the community and 5 stars being the level of an acknowledged expert.

Then, when a contributor of level 3 reviews an entry of level 2, the entry level rises automatically to level 3. Similarly, if the work of a contributor without corrections is systematically validated by other contributors of higher level, s/he can pass automatically to the next level after a certain threshold (eg 10 contributions). For example, Figure 5.4 shows an entry with a level of 3 stars.

To go further, we plan to analyze the work of contributors. If a person contributes heavily for example on a particular area, the system can automatically send regular contribution proposals in the domain.

6. DATA BUILDING METHODOLOGY

The data building methodology consists in three main steps: retrieving existing data, collecting new data via serious games and finally, online contribution on the Web.

6.1. Retrieving existing data

To encourage contributions, it is preferable to propose a skeleton of existing data (even of bad quality), rather than an empty dictionary (writer's block). For each language involved, a list of words will be collected in order to create an initial list of entries. It is always possible to create a new entry, but the creations will be subject to verifications.

According to the sub-projects and languages involved, several dictionaries can be used:

- Fe* dictionary projects (French - English + other language): FeM (Malay) [10], FET (Thai), Feb (Vietnamese);
- The DiCo database for French;
- The VietDict French-Vietnamese bilingual dictionary.
- The French-Khmer phonetic bilingual dictionary[11].

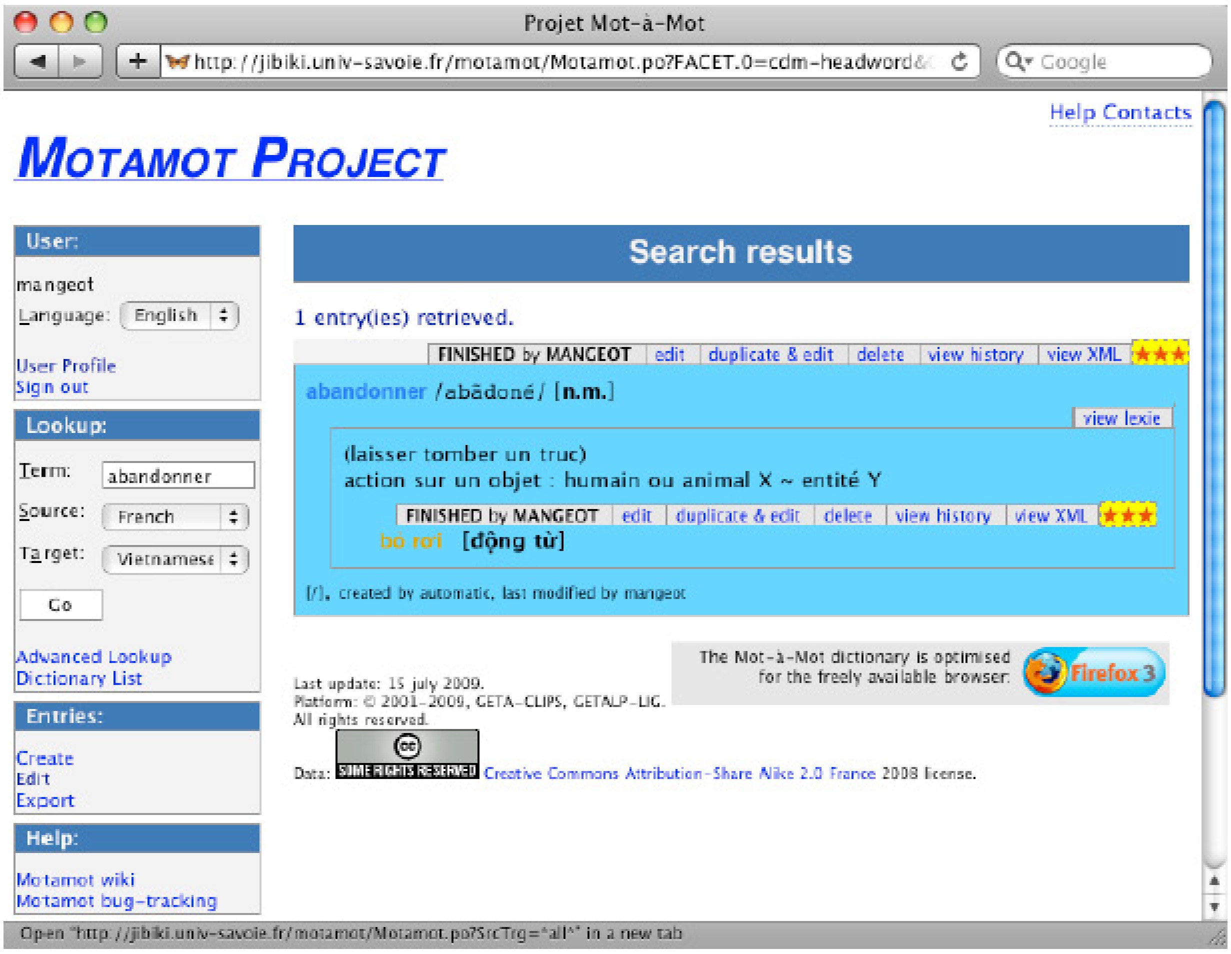
The number of stars of initial items generated from this data set is based on the quality of dictionary and the granularity of data retrieved.

6.1.1. Special handling for Khmer

For the Khmer language, we plan to computerize an existing French-Khmer phonetic dictionary. Its building began in the late 90s, and was completed in 2006 by a small group of researchers and computer scientists gathered in the non profit organisation "Pays perdu" created by Denis Richer, a French ethnolinguist, established in Siem Reap (Cambodia). The first version of dictionary was published in spring 2007 and includes 20,000 entries. Table 1 shows an example of what the dictionary looks like.

Table 1. Excerpt of the French-Khmer dictionary

French	**Khmer**
jarret	kɔ̄nlēak-cə̥ɳ
jars	kəɳān-chmɔ̥l
jasmin	mlih
jauge	▯cɑmnoh-rɔ̄ɳvuəl
— (techn.)	māet-stuəɳ



Projet Mot-à-Mot

http://jibiki.univ-savoie.fr/motamot/Motamot.po?FACET.0=cdm-headword&

Help Contacts

MOTAMOT PROJECT

User:

mangeot

Language: English

User Profile
Sign out

Lookup:

Term: abandonner

Source: French

Target: Vietnamese

Go

Advanced Lookup
Dictionary List

Entries:

Create
Edit
Export

Help:

Motamot wiki
Motamot bug-tracking

Search results

1 entry(ies) retrieved.

FINISHED by MANGEOT | edit | duplicate & edit | delete | view history | view XML | ★★★

abandonner /abãdoné/ [n.m.]

view lexie

(laisser tomber un truc)
action sur un objet : humain ou animal X ~ entité Y

FINISHED by MANGEOT | edit | duplicate & edit | delete | view history | view XML | ★★★

bỏ rơi [động từ]

[/], created by automatic, last modified by mangeot

The Mot-à-Mot dictionary is optimised for the freely available browser: Firefox 3

Last update: 15 july 2009.
Platform: © 2001-2009, GETA-CLIPS, GETALP-LIG.
All rights reserved.

Data: SOME RIGHTS RESERVED Creative Commons Attribution-Share Alike 2.0 France 2008 license.

Open "http://jibiki.univ-savoie.fr/motamot/Motamot.po?SrcTrg='all'" in a new tab

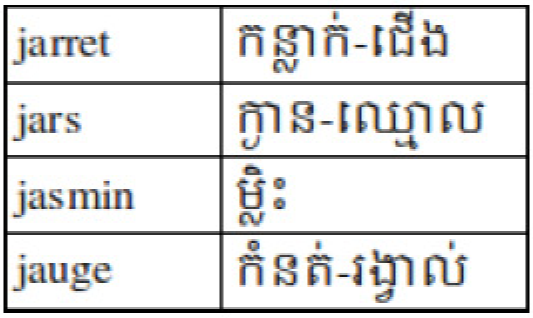
Fig. 3. Example entry in MotÀMot dictionary

The dictionary is in Word format and the Khmer part consists only in a phonetic transcription of the entry written in a special IPA script (SIL Sophia IPA 93) set up by the Summer Institute of Linguistics. Therefore, in order to obtain a completely XML Unicode data, we have to perform the following steps:

- convert the words in SIL IPA script into an Unicode IPA script
- obtain the equivalent words in khmer script. This step might be done semi-automatically, but we will probably need a post-edition step.
- tag the entries Most of the French entries consist in a simple word, but some of them have additional information that is not tagged, eg: a gloss `jambose (fruit)`, a feminine `jardinier, ère`, a domain `jauge - (techn.)`, etc.

Figure 6.1.1 shows the same example in khmer script.

Unfortunately, the Khmer encoded in Unicode cannot yet be read correctly on all the current platforms. On the Apple MacOs, some characters are not rendered correctly. For example, the figure 6.1.1 shows the rendering of the first khmer word. Therefore, we may need to automatically generate an image for each khmer word in order to fix the problem temporarily. A better solution would be to discuss with Apple in order to fix the problem definitely, but this is another story.



jarret	កន្លាក់-ជើង
jars	ក្ងាន-ឈ្មោល
jasmin	ម្លិះ
jauge	កំនត់-រង្វាល់

Fig. 4. Dictionary in Khmer script

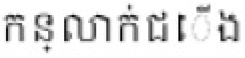


Fig. 5. Rendering problems for Khmer

6.2. Data Collection via serious games

The idea is to launch a JeuxDeMots for each language project[3]. The French JeuxDeMots was launched two years ago. The Vietnamese version was launched in autumn 2009. The khmer version is being translated. We hope to find a similar success than the French JeuxDeMots. Furthermore, we should should think about games allowing bilingual data collection. People interested can contact us if they want to launch a JeuxDeMots game in their language.

6.3. Online contribution on the web

The retrieved data is collected and then merged in order to give birth to a skeleton dictionary. It is then put online for correction and enrichment.

7. CONCLUSION

The project is already fairly well advanced. Most technical aspects concerning the platform and online serious games are solved. It remains to gather and convert existing resources. The major challenge of the project is actually our ability to motivate communities of contributors. We hope that our experience and the attraction of such a project will allow us to make a significant step in these sociological aspects.

The benefits of such a project are numerous and will help to revive the interest in the francophonie in the Southeast Asian countries. Data generated can be used by learners of French in these countries, or francophones wishing to learn a Southeast Asian language. The dictionaries may be used by tourists or businessmen directly online or via PDAs.

The communities of contributors should launch a new dynamics of cooperation around a common humanistic purpose. Moreover, it can arouse interest for expanding the project to other languages of the region.

8. ACKNOWLEDGMENTS

The MotAMot project has been partly funded by the Action de Recherche en Réseau program of the Agence Universitaire de La Francophonie[4].

9. REFERENCES

[1] Vincent Berment, *Méthodes pour informatiser des langues et des groupes de langues "peu dotées"*, Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France, 18 mai, 277 p. 2004.

[2] Igor Mel'čuk and Alain Polguère, ``Dérivations sémantiques et collocations dans le dico/laf.,'' in *Langue française, numéro spécial sur la collocation « Collocations, corpus, dictionnaires »*, vol. 150, pp. 66--83. P. Blumenthal and F. J. Hausmann, June 2006.

[3] Étienne Blanc, ``Parax-unl: a large scale hypertextual multilingual lexical database,'' in *NLPRS'99: the 5th Natural Language Processing Pacific Rim Symposium*, Beijing, China, 1999, p. 4.

[4] Alain Polguère, ``Structural properties of lexical systems: Monolingual and multilingual perspectives,'' in *Workshop on Multilingual Language Resources and Interoperability (COLING/ACL 2006)*, Sydney, 17-21 July 2006, pp. 50--59.

[5] Alain Polguère, ``Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french,'' in *Proceedings of EURALEX'2000*, Stuttgart, Germany, 2000, pp. 517--527.

[6] Mathieu Mangeot, ``Papillon project: Retrospective and perspectives.,'' in *Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine, LREC workshop*, Pierre Zweigenbaum, Ed., Genoa, Italy, 22 May 2006, p. 6.

[7] Mathieu Lafourcade and Alain Joubert, ``Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes,'' in *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, France, 12-14 mars 2008, pp. 657--666.

[8] G. A. Miller, R. Beckwith, C . Fellbaum, D. Gross, and K. J. Miller, ``Introduction to wordnet: an on-line lexical database,'' *International Journal of Lexicography*, vol. 3, no. 4, pp. 235--244, 1990.

[9] Mathieu Mangeot and David Thevenin, ``Online generic editing of heterogeneous dictionary entries in papillon project,'' in *Proc. of the COLING 2004 conference*, Geneva, Switzerland, 26 August 2004, vol. 2, pp. 1029--1035.

[10] Yvan Gut, Puteri Rashida Megat Ramli, Zaharin Yusoff, Kim Choy Chuah, Salina A. Samat, Christian Boitet, Nicolai Nedobejkine, Mathieu Lafourcade, Jean Gaschler, and Dorian Levenbach, *Kamus Perancis-Melayu Dewan, Dictionnaire francais-malais*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1996.

[11] Denis Richer, Keo T, and Vanra Ieng, *Dictionnaire Français-Khmer (en phonétique)*, D.R. Edition, 2007.

[3] `http://www.jeuxdemots.org`

[4] `http://www.ltt.auf.org`

ENGLISH-LATVIAN SMT: THE CHALLENGE OF TRANSLATING INTO A FREE WORD ORDER LANGUAGE

Maxim Khalilov[1], José A. R. Fonollosa[2], Inguna Skadiņa[3], Edgars Brālītis[3], and Lauma Pretkalniņa[3]*

[1] Institute for Logic, Language and Computation
Universiteit van Amsterdam
Amsterdam, The Netherlands

[2] Centre de Recerca TALP
Universitat Politècnica de Catalunya
Barcelona, Spain

[3] Institute of Mathematics and Computer Science
University of Latvia
Riga, Latvia

ABSTRACT

This paper presents a comparative study of two approaches to statistical machine translation (SMT) and their application to a task of English-to-Latvian translation, which is still an open research line in the field of automatic translation.

We consider a state-of-the-art phrase-based SMT and an alternative N-gram-based SMT systems. The major differences between these two approaches lie in the distinct representations of bilingual units, which are the components of the bilingual model driving translation process and in the statistical modeling of the translation context.

Latvian being a rather free word order language implies additional difficulties to the translation process. We contrast different reordering models and investigate how well they deal with the word ordering issue.

Moving beyond automatic scores of translation quality that are classically presented in MT research papers, we contribute presenting a manual error analysis of MT systems output that helps to shed light on advantages and disadvantages of the SMT systems under consideration and identify the most prominent source of errors typical for both SMT systems.

Index Terms**—** Natural languages, finite state machines, language processing, statistical machine translation.

1. INTRODUCTION

Translation into languages with relatively free word order has received a lot less attention than translation into fixed word order languages (English), or into analytical languages (Chinese). Free word order languages differ crucially from the languages that follow a restrictive word order scheme, first of all, in the way how the pragmatic information is conveyed. In fixed word order languages (like, German, English, or Spanish) the order of syntactic constituents varies negligibly (or does not vary at all) and the emotional component of the message is usually transmitted through intonation variation[1]. In contrast to them, the free word order languages (like, Latvian, Russian, or Greek) often rely on the order of constituents to convey the topicalization or focus of the sentence.

Latvian language is the target language in the experiments that we report in this paper. There are about 1.5 million native Latvian speakers around the world: 1.38 million are in Latvia, while others are spread in USA, Russia, Sweden, and some other countries. Also Latvian language is second language for about 0.5 million inhabitants of Latvia and several tens of thousands from neighbor countries, especially Lithuania[2].

Latvian is one of two living Baltic languages and it is characterized by rich morphology, relatively complex pre- and postposition structures and high level of morphosyntactic ambiguity. Despite that it descends from the same ancestor language as Germanic languages, it differs from them significantly and the experience gained from machine translation into German or English can hardly be transferred to the English-to-Latvian translation task.

Nowadays, scientific community is starting to express doubts that the models working pretty well for fixed word order languages are still efficient for free word order languages (for example, construction of an English-to-Czech SMT system taking into consideration very rich morphology

[*] The bulk of the work presented in this paper was done during the first author's Ph.D studies in Centre de Recerca TALP, Universitat Politècnica de Catalunya, Barcelona (Spain).

[1] There are some exceptions to the general rule, for example, when it is necessary to emphasize the object of the sentence (*I agree* ***with you*** -> ***With you*** *I agree*), or in question sentences.

[2] Source: State Language Agency `http://www.valoda.lv/lv/latviesuval`

and relatively free word order of Czech is one of the goals of the Euromatrix(plus) project[3]). A thorough discussion of the appropriate word ordering strategy (using contextual information) for English-to-Turkish rule-based machine translation can be found in [1]; in [2], the authors concentrate on SMT of indigenous Australian languages (one of the two languages under consideration is a prototypical non-configurational language).

However, translation from Latvian into English and vice versa has not received much attention in the SMT community: the first and only study on Latvian-to-English SMT, to our knowledge, was dated to 2007 [3], that is much later than SMT systems for popular language pairs.

In this paper, we study some aspects of English-to-Latvian MT. First, we compare the outputs of two SMT systems following two different approaches to MT and reporting results in terms of automatic evaluation metrics. We consider a "de facto" standard phrase-based Moses[4] system [4] and an N-gram-based SMT system [5]. We then study two alternative word reordering techniques for each translation system and measure how effective they are translating from English into a non-configurational Latvian language.

The paper concludes with human error analysis performed in order to identify the major strengths and weaknesses of the Moses and N-gram-based SMT systems when translating into Latvian.

The rest of this paper is organized as follows. Section 2 briefly describes phrase- and N-gram-based SMT system configurations, Section 3 outlines the experimental setup, Section 4 details the results of automatic translation quality evaluation, along with the results of human evaluation and error analysis, while Section 5 presents the conclusions drawn from the study.

2. TWO APPROACHES TO SMT

SMT is based on the principle of translating a source sentence ($f_1^J = f_1, f_2, ..., f_J$) into a sentence in the target language ($e_1^I = e_1, e_2, ..., e_I$). The problem is formulated in terms of source and target languages; it is defined according to equation (1) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (2):

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} = & (1) \\ &= \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} & (2) \end{aligned}$$

where I and J represent the number of words in the target and source languages, respectively.

Modern state-of-the-art SMT systems operate with the bilingual units extracted from the parallel corpus based on word-to-word alignment. They are enhanced by the *maximum entropy approach* and the posterior probability is calculated as a *log-linear combination* of a set of feature functions [6]. Using this technique, the additional models are combined to determine the translation hypothesis $\hat{e}_1^I$ that maximizes a log-linear combination of these feature models [7], as shown in (3):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \qquad (3)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

There have been a bunch of publications that investigate the source of the possible improvements and degradations in translation quality when using translation systems underlined by different statistical models. For example, in [8], the N-gram-based system is contrasted with a state-of-the-art phrase-based framework, while in [9], the authors seek to estimate the advantages, weakest points, and possible overlap between syntax-augmented MT and N-gram-based SMT. In [10] the comparison of phrase-based, hirearchical, and syntax-based SMT systems is provided.

In this section we discuss the translation models compared in this work.

2.1. Phrase-based SMT

Most of modern state-of-the-art SMT systems follow the phrase-based approach to translation. The basic idea of this approach is to segment the given source word sequence into monolingual phrases, afterwards translate them and compose the target sentence [6].

A phrase-based translation is considered as a three step algorithm: (1) the source sequence of words is segmented in phrases, (2) each phrase is translated into target language using translation table, (3) the target phrases are reordered to be inherent in the target language.

A bilingual phrase (which in the context of SMT do not necessarily coincide with their linguistic analogies) is any aligned pair of m source words and n target words that satisfies two basic constraints: (1) words are consecutive along both sides of the bilingual phrase and (2) no word on either side of the phrase is aligned to a word outside the phrase [11]. The probability of the phrases is estimated by relative frequencies of their appearance in the training corpus.

The system built for the English-to-Latvian translation experiments is implemented within the open-source MOSES toolkit [12]. Standard training and weights tuning procedures which were used to build our system are explained in details on the MOSES web page: `http://www.statmt.org/moses/`. Two word reordering methods are considered: a

[3] `http://www.euromatrix.net/`

[4] `http://www.statmt.org/moses/`

distance-based distortion model (see 2.1.1) and lexicalized MSD block-oriented model (see 2.1.2).

2.1.1. Distance-based

A simple distance-based reordering model default for Moses system is the first reordering technique under consideration. This model provides the decoder with a cost linear to the distance between words that should be reordered.

2.1.2. MSD

A lexicalized block-oriented data-driven MSD reordering model [13] considers three different orientation types: monotone (M), swap (S), and discontinuous(D). MSD model conditions reordering probabilities on the word context of each phrase pair and considers decoding process a block sequence generation process with the possibility of swapping a pair of word blocks. Notice that in the experiments conducted within the framework of this study a MSD model was used together with a distance-based reordering model.

2.2. N-gram-based SMT system

Alternative approach to SMT is the N-gram-based approach [5], which regards translation as a stochastic process that maximizes the joint probability $p(s,t)$, leading to a decomposition based on bilingual n-grams, typically implemented by means of a Finite-State Transducer [14].

The core part of the system constructed in this way is a translation model (TM), which is based on bilingual units, called tuples, that are extracted from a word alignment according to certain constraints. A bilingual TM actually constitutes an n-gram LM of tuples, which approximates the joint probability between the languages under consideration and can be seen here as a LM, where the language is composed of tuples.

The tuple-based approach is considered monotonous because the model is based on the sequential order of tuples during training. However, for a great number of translation tasks, a certain reordering strategy is required. In the framework of this study we consider two reordering models: a nondeterministic reordering method (see 2.2.2) and a deterministic version of the statistical machine reordering (SMR) algorithm (see 2.2.3).

2.2.1. Additional features

The N-gram translation system implements a log-linear combination of five additional models:

- *an n-gram target LM;*
- *a target LM of Part-of-Speech (POS) tags*;
- *a word penalty model* that is used to compensate for the system's preference for short output sentences;
- *source-to-target and target-to-source lexicon models* as shown in [15]).

2.2.2. Extended word reordering

An extended monotone distortion model based on the automatically learned reordering rules was implemented as described in [16]. Based on the word-to-word alignment, tuples were extracted by an *unfolding* technique. As a result, the tuples were broken into smaller tuples, and these were sequenced in the order of the target words.

The reordering strategy is additionally supported by a 4-gram LM of reordered source POS tags. In training, POS tags are reordered according to the extracted reordering patterns and word-to-word links. The resulting sequence of source POS tags is used to train the n-gram LM.

2.2.3. Statistical machine reordering

A SMR technique is described in details in [17]. Here, reordering is thought as a first-pass translation performed on the source corpus, which converts it into an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. A monotone sequence of source words is translated into the reordered sequence using SMT techniques: SMR and SMT are performed using the same modeling tools as N-gram-based systems but using different statistical log-linear models.

Statistical word classes are used to introduce generalization power to the reordering model.

2.2.4. Decoding and optimization

The open-source MARIE[5] decoder was used as a search engine for the translation system. Details can be found in [18]. The decoder implements a beam-search algorithm with pruning capabilities. All the additional feature models were taken into account during the decoding process. Given the development set and references, the log-linear combination of weights was adjusted using a *simplex* optimization method and an n-best re-ranking as described in `http://www.statmt.org/jhuws/`.

3. EXPERIMENTS

3.1. Data

We used JRC Acquis 2.2 parallel corpus [19] of about 270K parallel sentences. Development set contained of 500 sentences randomly extracted from the bilingual corpus, test corpus size was 1,000 lines. Both the datasets were provided

[5] `http://gps-tsc.upc.es/veu/soft/soft/marie/`

with 1 reference translation. Basic statistics of the bilingual corpus can be found in Table 1.

	Latvian	English
Training		
Sentences	269.98 K	269.98 K
Words	5.40 M	6.65 M
Vocabulary	101.25 K	60.47 K
Development		
Sentences	0.50 K	0.50 K
Words	9.90 K	12.36 K
Vocabulary	3.08 K	2.30 K
Test		
Sentences	1.00 K	1.00 K
Words	20.18 K	24.64 K
Vocabulary	4.98 K	3.49 K

Table 1: Basic statistics of the JRC-Acquis corpus.

3.2. Experimental details

Word alignments were estimated with GIZA++ tool[6] assuming 4 iterations of the IBM2 model, 5 HMM model iterations, 4 iterations of the IBM4 model, and 50 statistical word classes (estimated with the mkcls tool[7]).

Phrase-based experiments were conducted following the guidelines provided on the Moses site[4]. We used the 2008 version of Moses decoder. As an alternative to a traditional (unfactored) model (*PB-u*), we considered a factored phrase-based SMT (*PB-f*) that constructed translation/generation models on the basis of the factorized corpus (preface words, POS tags, and lemmas for English and Latvian).

A 4-gram target LM with unmodified Kneser-Ney backoff discounting was generated using the SRI Language Modeling Toolkit [20] and was used in all the experiments.

The following MSD reordering system configuration was used: (*msd-bidirectional-fe* configuration).

The SMR experiments were carried out using 50 classes in the reordering step.

4. RESULTS

4.1. System configurations and evaluation

Two SMT systems (*PB-u* - unfactored and *PB-f* - factored) were contrasted considering the set of experiments carried out on the phrase-based system. Within each system configuration we considered two reordering models: a distance-based model alone (as described in 2.1.1) and a distance-based model operating together with a MSD model (see 2.1.2).

N-gram-based SMT system was enhanced with two alternative reordering models: SMR (see 2.2.3) and an extended input graph model (details can be found in 2.2.2).

We considered four evaluation metrics:

- The BLEU score [21] that accounts for evaluation of the translation quality, by measuring the distance between a given translation and the set of reference translations using an n-gram LM (a 4-gram in this study);
- The NIST score [22] which is based on the BLEU score, but weights n-grams in order to provide less informative n-grams with higher weights;
- The WER score [23] which calculates the minimum word-level Levenshtein distance between a translation system output and a reference translation;
- The PER score [24] which is a variation of WER metric, alleviating the effect of a possibly different word order between an acceptable translation hypothesis and reference translation.

System	Reordering	Dev	Test			
			BLEU	NIST	PER	WER
Phrase-based SMT (Moses)						
PB-u	distance	42.38	43.87	78.80	38.34	51.12
	distance + MSD	42.69	43.95	78.91	38.48	50.47
PB-f	distance	42.11	42.96	78.68	38.71	51.75
	distance + MSD	42.40	43.80	78.63	38.63	50.93
N-gram-based SMT (TALP)						
NB	SMR	43.20	44.64	82.03	35.01	47.98
	Input graph	43.52	45.11	82.40	35.05	47.97

Table 2: English-to-Latvian experimental results.

[6] http://code.google.com/p/giza-pp/

[7] http://www.fjoch.com/mkcls.html

Automatic evaluation was case sensitive and punctuation marks were considered.

4.2. Automatic evaluation

The results of automatic evaluation of translation quality are shown in Table 2. Best scores are placed in cells filled with grey (within phrase-based and N-gram-based experimental sets).

The major conclusion that can be drawn from the results is that the N-gram-based translation model significantly outperforms the phrase-base system for the English-Latvian language pair. The absolute difference in BLEU score of the best ranked *NB* (namely, *NB* with input graph reordering model) and *PB* (namely, *PB-u* with distance-based and MSD reordering models) systems is about 1.15 BLEU points (that accounts for ≈2.6% in a relative scale). This difference is statistically significant for a 95% confidence interval and 1000 resamples [25][8].

Another important observation is that both "distance+MSD" *PB* models (factored and unfactored) are comparable in terms of automatically evaluated accuracy and both outperform their "distance-based only" versions. The difference between *PB-u* and *PB-f* "distance+MSD" systems is not statistically significant. We speculate that a reordering model plays more important role than a translation model factorization when translating into free word order languages.

The *NB* system enhanced with an input graph POS reordering model achieves better MT performance than the SMR version of this system and this difference is statistically significant.

The difference between "distance-based only" and "distance+MSD" versions of the phrase-based SMT systems is not statistically significant in case of the unfactored TM and it is significant in case of the factored model.

According to the PER metric, the introduction of the MSD model does not introduce any significant improvement. At the same time, the performance of the "distance+MSD" configurations expressed in the WER score is about 0.6-0.8 points better[9] than the performance shown by the distance-based reordering models. As a rough approximation, these results can be interpreted as that the MSD model implies an important improvement in word ordering within a sentence and outperforms the distance-based model applied alone.

4.3. Human evaluation and error analysis

Human analysis of translation output allows going beyond automatic scores and, in the general case, provides a comprehensive comparison of multiple translation systems.

Two best systems according to automatic scores were chosen from the phrase-based and N-gram-based experiment sets for human evaluation (*PB-u* with distance-based and MSD reordering models, and *NB* with input word graph model). Every non-repetitive test line from the output of these systems was presented to the judge, who was instructed to decide that the two translations were of equal quality, or that one translation was better than the other. The results of the standard systems comparison can be found in Table 3 and demonstrate that the *NB* system outperforms the *PB* one.

	PB-u +distance +MSD	NB +input graph	Equal
Preference	58	193	539

Table 3: Human evaluation results (standard systems).

In addition, we performed error analysis on 100 first sentences from the test data. The analysis of typical errors generated by each system was done following the error classification scheme proposed in [26] by contrasting the systems output with the reference translation. Table 4 presents the comparative statistics of errors generated by the *PB-u* system enhanced with distance-based and MSD reorderings and the *NB* system with input graph reordering model.

Evaluation of the word order correctness for free word order languages is not a trivial task. We considered equally all admissible word order combinations for the Latvian translations. The clumps are marked erroneous only if the word order is not acceptable in Latvian. In this sence, error analysis gives a more complete and fair view of translation quality than automatic scores which just compare a translation output with a reference translation.

The most prominent source of errors generated by the *PB-u* system, in comparison wit the *NB* system, is related to missing words found in the translation output. We explain it by a high analytical inflection of the Balto-Slavic languages that is modeled better by the N-gram-based system since it involves surrounding context not only for phrase reordering, but conditions translation decisions on previous translation decisions.

However, the aforementioned feature of the N-gram-based architecture turns to be a weakness when dealing with local word reordering, that is reflected in the high number of reordering errors produced by the *NB* system. Experimental results show that internal phrase-based reordering enhanced with the distance-based and MSD block-oriented reordering models (viewing translation as a monotone block sequence generation process) outperforms the POS-based word graph reordering model used in N-gram-based experiments (22 local word/phrase order errors coming from the *Pb-u* system vs. 37 errors of this type produced by the *NB* system).

At the same time, long-range word dependencies are modeled by *PB-u* and *NB* with comparable performance. For clar-

[8] Hereafter, statistical significance test is carried out on the BLEU score measured on the test dataset.

[9] For the WER and PER metrics the lower the score, the better the performance of a SMT system.

Type	Sub-type	PB-u + MSD	NB + input graph
Missing words		**64**	**16**
	Content words	52	10
	Filler words	12	6
Word order		**35**	**58**
	Local word order	11	23
	Local phrase order	11	14
	Long range word order	6	7
	Long range phrase order	7	14
Incorrect words		**128**	**82**
	Wrong lexical choice	25	20
	Incorrect disambiguation	10	4
	Incorrect form	51	46
	Extra words	34	9
	Style	8	2
	Idioms	0	1
Unknown words		**4**	**8**
Punctuation		**20**	**18**
Total		**250**	**182**

Table 4: Error statistics for a 100-line representative test set.

ity's sake, it is important to notice that the English-to-Latvian translation task is not characterized by the high number of long-range reordering dependencies.

Other important sources of errors of the *PB-u* system are extra words embedded into translated sentences (34 for the *PB-u* vs. 9 for the *NB*). We explain it by the key difference in internal representation of translation units between phrase-based and N-gram-based SMT systems.

5. CONCLUSIONS AND FUTURE WORK

In this paper two alternative SMT systems are compared: the standard phrase-based and the N-gram-based SMT systems. Both translation systems include modern reordering models in final configuration. The comparison was created to be as fair as possible, using the same training material and the same tools on the preprocessing, word-to-word alignment, and language modeling steps.

The results shows that the N-gram-based SMT outperforms Moses-based translation system for the English-to-Latvian translation task in terms of automatic scores (the difference is $\approx$1.15 BLEU points) and human "best/worse" evaluation (the output of the N-gram-based system was ranked higher than the one of the phrase-based system in 193 sentences, while the opposite occurred in 58 cases).

Human error analysis clarifies advantages and disadvantages of the systems under consideration and reveals the most important sources of errors for both systems. The phrase-based system suffers from the missing words problem, while, in case of N-gram-based SMT, the most frequent errors are caused by weak word reordering on the local level.

Findings of this study, along with the robust error analysis of the SMT system outputs can be a very important step on the way of the translation quality improvement when dealing with free word order languages.

A study on introducing of a feature intending to reflect a free word order scheme of the Latvian language is an interesting research topic to be done in the future. Another appealing research topic can be to modify the standard evaluation metrics used for the automatic assessment of translation quality such that they could consider multiple addmisible word permutations within a sentence to express the same message typical for the non-configurational languages.

6. ACKNOWLEDGEMENTS

Work partially supported by the Spanish Ministerio de Educación y Ciencia (TIN2006-12767), by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project), and by the Latvian Council of Science (project "Application of Factorized methods in English-Latvian Statistical Machine Translation System"). The authors want to thank Khalil Sima'an (Universiteit van Amsterdam) for his valuable discussions and suggestions.

7. REFERENCES

[1] B. Hoffman, “Translating into free word order languages,” in *Proceedings of COLING’96*, Copenhagen, Denmark, August 1996, pp. 556–561.

[2] S. Zwarts and M. Dras, “Statistical machine translation of australian aboriginal languages: Morphological analysis with languages of differing morphological richness,” in *Proceedings of the Australasian Language Technology Workshop*, Melbourne, Australia, December 2007, pp. 134–142.

[3] I. Skadiņa and E. Brālītis, “Experimental statistical machine translation system for Latvian,” in *Proceedings of the 3rd Baltic Conference on HLT*, 2008, pp. 281–286.

[4] Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: open-source toolkit for statistical machine translation,” in *Proceedings of ACL 2007*, 2007, pp. 177–180.

[5] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà, “N-gram based machine translation,” *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, December 2006.

[6] F. Och and H. Ney, “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation,” in *Proceedings of ACL 2002*, 2002, pp. 295–302.

[7] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin, “A statistical approach to machine translation,” *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[8] J. M. Crego, M. R. Costa-jussà, J. B. Mariño, and J. A. R. Fonollosa, “Ngram-based versus phrase-based statistical machine translation,” in *Proceedings of the 2nd Int. Workshop on Spoken Language Translation (IWSLT’05)*, 2005, pp. 177–184.

[9] M. Khalilov M. and J. A. R. Fonollosa, “N-gram-based statistical machine translation versus syntax augmented machine translation: comparison and system combination,” in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL’09)*, Athens, Greece, April 2009, pp. 424–432.

[10] A. Zollmann, A. Venugopal, F. Och, and J. Ponte, “A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT,” in *Proceedings of Coling 2008*, Manchester, August 2008, pp. 1145–1152.

[11] F. Och and H. Ney, “The alignment template approach to statistical machine translation,” *Computational Linguistics*, vol. 3, no. 4, pp. 417–449, December 2004.

[12] Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: open-source toolkit for statistical machine translation,” in *Proceedings of ACL 2007*, 2007, pp. 177–180.

[13] C. Tillman, “A unigram orientation model for statistical machine translation,” in *Proceedings of HLT-NAACL’04*, 2004.

[14] F. Casacuberta, E. Vidal, and J. M. Vilar, “Architectures for speech-to-speech translation using finite-state models,” in *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, 2002, pp. 39–44.

[15] F. Och and H. Ney, “The alignment template approach to statistical machine translation,” *Computational Linguistics*, 2004.

[16] J. M. Crego and J. B. Mariño, “Improving statistical MT by coupling reordering and decoding,” *Machine Translation*, vol. 20, no. 3, pp. 199–215, 2006.

[17] M. R. Costa-jussà and J. A. R. Fonollosa, “Statistical machine reordering,” in *Proceedings of the HLT/EMNLP 2006*, Sydney, Australia, 2006, pp. 70–76.

[18] J. M. Crego, J. B. Mariño, and A. de Gispert, “An ngram-based statistical machine translation decoder,” in *Proceedings of INTERSPEECH05*, 2005.

[19] S. Ralf, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga, “The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages,” in *Proceedings of LREC’2006*, Genoa, Italy, May 2006.

[20] A. Stolcke, “SRILM: an extensible language modeling toolkit,” in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of ACL 2002*, 2002, pp. 311–318.

[22] G. Doddington, “Automatic evaluation of machine translation quality using n-grams co-occurrence statistics,” in *Proceedings of HLT 2002*, 2002, pp. 128–132.

[23] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, “On the use of information retrieval measures for speech recognition evaluation,” IDIAP-RR 73, IDIAP, Martigny, Switzerland, 2004.

[24] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, "Accelerated DP based search for statistical translation," in *Proceedings of EUROSPEECH 1997*, Rhodes, Greece, September 1997, pp. 2667–2670.

[25] Ph. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP 2004*, 2004, pp. 388–395.

[26] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney, "Error Analysis of Machine Translation Output," in *Proceedings of LREC 2006*, 2006, pp. 697–702.

ANALYSIS OF NOORI NASTA'LEEQ FOR MAJOR PAKISTANI LANGUAGES

*M. G. Abbas Malik** *Christian Boitet** *Pushpak Bhattachariyya+*

Abbas.Malik@imag.fr Christian.Boitet@imag.fr pb@cse.iitb.ac.in

*GETALP-LIG, Université de Grenoble (Ex. Université Joseph Fourier), France.
+CSE, IIT Bombay, India.

ABSTRACT

Nasta'leeq is a bidirectional, diagonal, non-monotonic, cursive, highly context-sensitive and very complex writing style for languages like Urdu, Punjabi, Balochi and Kashmiri. Each is written in a variant of the Perso-Arabic script. The style is characterized by well-formed orthographic rules that are passed down from generation to generation of calligraphers and old manuscripts. It is present in calligraphic arts and printed materials of the present, but orthographic rules have not been quantitatively analyzed in detail for the above-mentioned languages. This paper first presents the salient features of the Perso-Arabic script and briefly introduces its different writing styles. It also briefly discusses alphabets of major Pakistani languages. Finally, it gives the quantitative analysis of Nasta'leeq and explains its context-sensitive behavior with respect to Pakistani languages, knowing that it is equally true for Arabic, Persian and other languages written in derivations of the Perso-Arabic script. Finally, it discusses the Context-Sensitive Substitution Grammar of Nasta'leeq, a computational model of Nasta'leeq.

Index Terms— Nasta'leeq, script, Arabic, Persian, Urdu, Punjabi, Sindhi, Balochi, Kashmiri

1. INTRODUCTION

Pakistan is a country with at least six major languages and 58 minor ones [1]. Urdu, the national language, has over 11 million (7.57%) native speakers while those who use it as a second language are more than 105 million [2]. Punjabi, the mother tongue of 44.15% of the population, is the biggest language of Pakistan. Other major languages are Pashto, Sindhi, Balochi and Kashmiri. The size of these languages and Urdu is shown in Table 1.

The benefits from the Information Technology (IT) revolution cannot be reaped unless masses use it, which is not possible unless computing is possible in the languages that are understood by the masses [3]. Information has become such an integral part of our global society that access to it is considered as a basic human right. Internet is believed to be the dominant carrier of information across the globe. Currently, English is the lingua franca for Internet and most of the information is available in it, but that makes information practically inaccessible to the vast majority of the world. This is applicable especially to countries like Pakistan where those who may be considered barely literate in Urdu represent only 43.92% population (66 millions according the 1998 census). That is rather a large number compared to the nearly 26 millions (17.29%) who, having passed the ten-year school system (matriculation), can presumably read and understand a little English. Internet and computer programs function in English in Pakistan and not even in Urdu let alone in the other languages. This means that most Pakistanis are either excluded from the digital world or function in it as handicapped aliens. In other words, Pakistani languages are under-resourced. Indeed, knowledge of English of most matriculates from Urdu and Sindhi medium schools is so rudimentary that they cannot carry out any meaningful interaction, especially those that would increase their knowledge or analytical skills, with the digital world. Perhaps only 4.38% graduates (about 6.5 millions) could do so [1].

Language	**Number of Speakers**
Urdu*	164,290,000
Punjabi	66,225,000
Pashto	23,130,000
Sindhi	21,150,000
Balochi	5,355,000
Kashmiri	4,496,000

* We include native and 2nd language speakers of Urdu. Source: [1]

Table 1: Speakers of Pakistani languages

2. ARABIC SCRIPT AND ITS WRITING STYLES

The Arabic script is a cursive writing system. It has many writing styles, including Naskh, Kufi, Sulus, Riqah, Deevani, etc. Some of them are shown in figure 1. The Nasta'leeq writing style was developed in Iran during the 14th and 15th centuries by combining Naskh and Taleeq (an old obsolete style)[1]. It is one of the main genres of the Islamic calligraphy. It is rich in calligraphic content. Owing to complexities of orthographic rendering, the basic shapes identified in this section are unable to render a language in an acceptable form in any Nasta'leeq style. A detailed quantitative analysis of Nasta'leeq with respect to Pakistani languages is given in section 4.

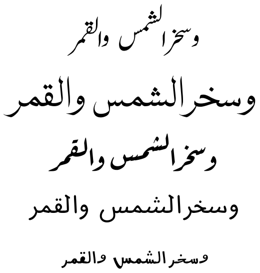


Figure 1: Different writing styles for Arabic

The distinguishing characteristics of Perso-Arabic script are discussed for the benefit of the unacquainted reader. It is read from right-to-left. Figure 2 shows some sample characters of Pakistani languages. Unlike English, characters do not have upper and lower case.

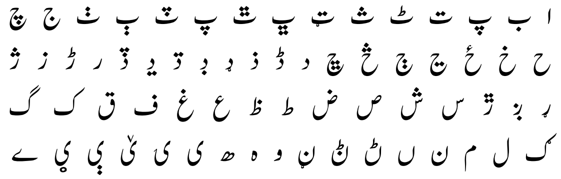


Figure 2: Sample characters of Pakistani languages

The shape assumed by a character in a word is context-sensitive, i.e. the shape is different depending on whether the position of the character is at the beginning, in the middle or at the end of the constituent word. This generates three shapes, the fourth being the independent shape of the character [4,5]. Figure 3 shows these four shapes of the character Beh in Naskh writing style.

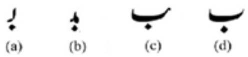


Figure 3: Context-sensitive shapes of BEH [4]

To be precise, the above is true for all except certain characters that only have the independent and the terminating shape when they occur at the beginning and the middle or end of a word respectively [4,5]. Some of these characters are shown in Figure 4.

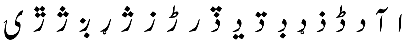


Figure 4: Sample characters having only two shapes

Hamza appears at the beginning of a word [4], but it could come at the beginning of a ligature. Also it takes the independent shape instead of the final shape when it comes at the end of the word. Thus, it has initial, middle and independent shapes [4,5], as illustrated in figure 5.

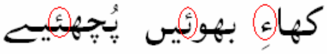


Figure 5: Shapes of Hamza (circled) [5]

The Arabic, Persian and Pakistani languages have a large set of diacritical marks that are necessary for the correct articulation of a word. The diacritical marks appear above or below a character to define a vowel or to geminate a character [4,5]. They are the foundation of the vowel system in these scripts. The most common diacritical marks with the character Beh are shown in Figure 6.

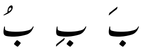


Figure 6: BEH with Diacritical Marks

Diacritics, though part of the writing system, are sparingly used [4]. They are essential for ambiguities removal, natural language processing and speech synthesis [4,5,6,7].

3. PAKISTANI LANGUAGES

Pakistani languages are written in an alphabet that is derived from the Perso-Arabic alphabet. It is not possible to discuss all Pakistani languages here. This paper only discusses the six languages given in Table 1 because the last five represent the major geographical divisions of Pakistan, and Urdu is the National language of Pakistan. All of these languages belong to the Indo-European language family. Their family tree is given in Figure 7.

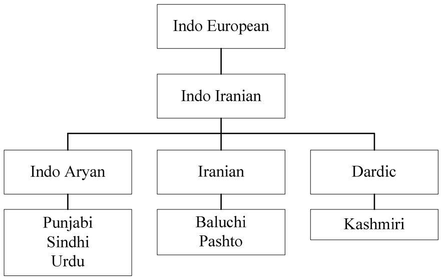


Figure 7: Language tree of 6 major Pakistani languages

The character sets of each of these languages are discussed separately here with their Unicode values. In Unicode, Arabic and its associated languages like Urdu, Punjabi, Pashto, Sindhi, etc. have been allocated the code points 0600h – 06FFh, 0750h – 077Fh and FB50h – FEFFh.

3.1. Urdu

Urdu is the National language of Pakistan and one of the state languages of India with more than 60 million native speakers. It is one of the largest languages of the world, if one considers Hindi/Urdu as dialects of the same language called Hindustani by Platts [8]. Table 2 gives the size of Hindi/Urdu.

[1] http://en.wikipedia.org/wiki/Nasta%27liq_script

Speakers	**Native**	**2nd Language**	**Total**
Hindi	366,000,000	487,000,000	853,000,000
Urdu	60,290,000	104,000,000	164,290,000
Total	426,290,000	591,000,000	**1,017,000,000**

Table 2: Hindi and Urdu speakers [7]

Urdu is written in Nasta'leeq style. It has 35 consonant characters representing 27 consonant sounds as some consonant sounds are represented by two or more consonant characters, e.g. the sound 's' is represented by three different characters Seh (ث), Seen (س) and Sad (ص) [7]. Out of 35 consonant characters, 32 are adopted from Persian. 3 retroflex consonants are added to accommodate the indigenous sounds of the Indian sub-continent. These characters are Tteh (ٹ) [ʈ], Ddal (ڈ) [ɖ] and Rreh (ڑ) [ɽ]. Non-aspirated consonants of Urdu are given in Table 3.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ب [b]	0628	19	ص [s]	0635
2	پ [p]	067E	20	ض [z]	0636
3	ت [t̪]	062A	21	ط [t̪]	0637
4	ٹ [ʈ]	0679	22	ظ [z]	0638
5	ث [s]	06B2	23	ع [ʔ]	0639
6	ج [ʤ]	062C	24	غ [ɣ]	063A
7	چ [ʧ]	0686	25	ف [f]	0641
8	ح [h]	062D	26	ق [q]	0642
9	خ [x]	062E	27	ک [k]	06A9
10	د [d̪]	062F	28	گ [g]	06AF
11	ڈ [ɖ]	0688	29	ل [l]	0644
12	ذ [z]	0630	30	م [m]	0645
13	ر [r]	0631	31	ن [n]	0646
14	ڑ [ɽ]	0691	32	و [v]	0648
15	ز [z]	0632	33	ہ [h]	06C1
16	ژ [ʒ]	0698	34	ی [j]	06CC
17	س [s]	0633	35	ة [t̪]	0629
18	ش [ʃ]	0634			

Table 3: Non-aspirated Urdu consonants

The phenomenon of aspiration does not exist in Persian or Arabic but it exists in languages of the region e.g. Hindi, Urdu, Punjabi, etc. In Urdu, the special character Heh Doachashmee (ھ) is used to mark the aspiration. Thus aspirated consonants are represented by the combination of the consonant to be aspirated and Heh Doachashmee (ھ) e.g. ب [b] + ھ [h] = بھ [b^h], ج [ʤ] + ھ [h] = جھ [ʤ^h], etc. Urdu has 15 aspirated consonants [7]. Aspirated Urdu consonants are given in Table 4.

Sr.	Symbol	Sr.	Symbol	Sr.	Urdu
1	بھ [b^h]	6	چھ [ʧ^h]	11	کھ [k^h]
2	پھ [p^h]	7	دھ [d̪^{h}]	12	گھ [g^h]
3	تھ [t̪^{h}]	8	ڈھ [ɖ^h]	13	لھ [l^h]
4	ٹھ [ʈ^h]	9	رھ [r^h]	14	مھ [m^h]
5	جھ [ʤ^h]	10	ڑھ [ɽ^h]	15	نھ [n^h]

Table 4: Aspirated Urdu consonants

In addition to consonants, Urdu has 10 vowels and 7 of them also have nasalized forms [9]. They are represented with the help of four long vowels (Alef Madda (آ), Alef (ا), Waw (و) and Yeh (ی)) and three short vowels (Arabic Fatha ◌َ, Damma ◌ُ and Kasra ◌ِ). The representation of a vowel is context-sensitive, i.e. a vowel may be written in two or more ways according to the context in a word, e.g. the vowel sound [ə] is represented by Alef (ا) + Zabar (◌َ) at the start of a word and by Zabar (◌َ) in the middle of a word. The vowel sound [ə] never comes at the end of a word. Nasalization of a vowel is marked with Noon-ghunna (ں) and with Noon (ن) at the end and in the middle of a word respectively [7]. For more details, see [7].

Urdu contains 15 diacritical marks. They represent vowel sounds, except Hamza-e-Izafat (◌ٔ) and Kasr-e-Izafat (◌ِ) that are used to build compound words, *e.g.* ادارۂ سائنس [ɪɖɑrəhɪsɑɪns] (Institute of Science), تاریخِ پیدائش [tɑrixɪpedɑɪʃ] (date of birth), *etc.* Shadda (◌ّ) is used to geminate a consonant *e.g.* ربّ [rəbb] (God), اچّھا [əʧʧʰɑ] (good), *etc.* Sukun (◌ْ) is used to mark the absence of a vowel after the base consonant [7,8].

Pakistani languages also share the Perso-Arabic punctuation and special symbols. These punctuation marks and symbols are given in Table 5.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	،	060C	10	؏	060F
2	؛	061B	11	◌ؐ	0610
3	؟	061F	12	◌ؑ	0611
4	۔	06D4	13	◌ؒ	0612
5	؀	0600	14	◌ؓ	0613
6	؁	0601	15	◌ؔ	0614
7	؂	0602	16	◌ؕ	0615
8	؃	0603	17	٪	066A
9	؎	060E			

Table 5: Punctuation marks and other symbols

Urdu has a numeral system that is derived from Persian. It assigns the same Unicode values as Persian ranging 06F0 – 06F9 but employs different shapes for number 4, 5 and 7. They are shown in Table 6.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	۰	06F0	6	۵	06F5
2	۱	06F1	7	۶	06F6
3	۲	06F2	8	۷	06F7
4	۳	06F3	9	۸	06F8
5	۴	06F4	10	۹	06F9

Table 6: Urdu numerals

3.2. Punjabi

Punjabi is written in two mutually incomprehensible scripts. One is the derivation of Perso-Arabic script (called Shahmukhi) used in Pakistan and the other is Gurmukhi, used in India. The Punjabi (Shahmukhi) alphabet is a superset of the Urdu alphabet and has one additional non-aspirated consonant, Rnoon (ݨ) [ɳ] [5,6]. The rest is the same as Urdu. Punjabi is also traditionally written in Nasta'leeq style. For more details on the Punjabi (Shahmukhi) alphabet see [5,6].

3.3. Pashto

Like Persian, Pashto does not have the aspiration. Heh Gol (ہ) takes the shape of Heh Doachashmee (ھ) when it comes at the start or middle of a ligature. Retroflex sounds also exist in Pashto like in Urdu and Punjabi, but Pashto employs different graphemes for them. Table 7 gives a shape comparison of retroflex consonants in six major Pakistani languages.

IPA	Urdu, Balochi, Kashmiri	Punjabi	Pashto	Sindhi
ʈ	ٹ	ٹ	ټ	ٽ
ɖ	ڈ	ڈ	ډ	ڊ
ɽ	ڑ	ڑ	ړ	ڙ
ɳ	-	ݨ	ڼ	ڻ

Table 7: Comparison of retroflex consonants

In Pashto, there exist five different kinds of Yeh. One is employed as a consonant and the others represent different vowel sounds. They are shown in Figure 8.

ی [j], ي [i], ې [e], ی [əy], ئ [ə]

Figure 8: Five Yehs of Pashto

Pashto has 39 consonants and uses the same Persian number system without any change. The vowel system of the Pashto script is also context-sensitive and is represented with the help of long vowels and diacritical marks. Pashto is traditionally written in Naskh style. Table 8 shows remaining Pashto characters that are not present in Urdu or have different shapes than in Urdu.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ځ [dz]	0681	4	ښ [ʂ]	069A
2	څ [ts]	0685	5	ګ [g]	06AB
3	ږ [ʐ]	0696			

Table 8: Pashto characters

3.4. Sindhi

Sindhi has 40 non-aspirated consonants and 11 aspirated consonants. In Sindhi, aspiration is expressed in different ways. For example, the aspiration of Jeem (ج) is indicated by Heh Doachashmee (ھ) like in Urdu and Punjabi, and the aspiration of Beh (ب) is expressed by a separate new character with four dots below ڀ. Sindhi aspirated and non-aspirated consonants that are not present in Urdu or have different shapes from those in Urdu are given in Table 9.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ٻ [ɓ]	067B	12	ڍ [ɖʰ]	068D
2	ڀ [bʰ]	0680	13	ڙ [ɽ]	0699
3	ٿ [t̪ʰ]	067F	14	ڙھ [ɽʰ]	-
4	ٽ [ʈ]	067D	15	ڦ [pʰ]	06A6
5	ٺ [ʈʰ]	067A	16	ڪ [k]	06AA
6	ڄ []	0684	17	ک [kʰ]	06A9
7	ڃ [ɲ]	0683	18	ڳ [ɠ]	06B3
8	ڇ [ʧʰ]	0687	19	ڱ [ŋ]	06B1
9	ڌ [d̪ʰ]	068C	20	ڻ [ɳ]	06BB
10	ڊ [ɖ]	068A	21	ي [j]	064A
11	ڏ [ɗ]	068F			

Table 9: Aspirated and non-aspirated Sindhi consonants

Sindhi has 16 vowels that are also context-sensitive.

Pashto and Sindhi are both traditionally written in Naskh and their analysis for a Nasta'leeq style has never been done before. We have done it because they could also be written in Nasta'leeq just like Arabic[2]. Thus it is worthwhile to provide an analysis of Nasta'leeq for Pashto and Sindhi and provide an opportunity to the Pashto and Sindhi communities to write their languages in Nasta'leeq.

3.5. Balochi

Balochi uses a modified alphabet of Urdu and is written in Nasta'leeq style. Balochi has removed the redundant characters for the same sound, e.g. for the sound of [s], it keeps the character Seen (س) and discards the others (ص ،ث). Thus Balochi has 22 consonants. Like Persian and Pashto, it

[2] Arabic is also traditionally written in Naskh but there are very beautiful manuscripts of Arabic and Qur'an in the Indian sub-continent that are written in Nasta'leeq style. The first author has seen one in Pakistan.

also has no aspiration. It has two additional diacritics; one is the Hamza mark (◌ٔ) above and the other is similar to the inverted Damma (◌ٗ), but is horizontally reversed and much flatter (◌ٝ). Some native speakers also write Balochi using the Urdu script.

3.6. Kashmiri

Kashmiri employs the Urdu alphabet with a few additions to represent its specific vowels. Kashmiri has two additional Yehs (ی), one with an oval below (ؠ) and the other with a 'v' mark above (ێ). It also has two additional Waws (و), one with a circle at the ending tail (ۄ) and the other with a 'v' mark above (ۆ). In diacritical marks, it adds two diacritical marks, a slightly modified Hamza (ٔ) written above and below the character. The extra characters of Kashmiri are shown in Table 10. It is also traditionally written in Nasta'leeq style.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ؠ []	-	4	ێ [e]	06CE
2	ۄ [ɔ]	06C4	5	◌ٔ [ə]	-
3	ۆ [o:]	06C6		◌ٕ []	-

Table 10: Kashmiri characters

4. ANALYSIS OF NASTA'LEEQ

The rendering of Pakistani languages in Nasta'leeq is very complex because the shape of a character not only depends on its position (at the start, in the middle or at the end) in the word but also depends on surrounding characters in the word. The fundamental shapes of the analysis of Section 2 are not sufficient to produce orthographic rendering of major Pakistani languages in Nasta'leeq, because Nasta'leeq is inherently context-sensitive. Figure 9 shows different context-sensitive shapes of the character Beh.

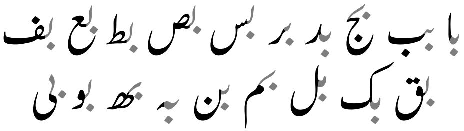


Figure 9: Context-sensitive shapes of Beh

Wali and Hussain [10] have given a quantitative analysis of Nasta'leeq (Nafees style) for Urdu. In this study, we give a quantitative analysis of the Noori style of Nasta'leeq for the six major Pakistani languages of Table 1.

For analysis purposes, we can divide our discussion in four parts, concerning independent shapes, two, three and four characters-joining. After the analysis of four characters long ligatures, the joining is recursive for ligatures longer than four, thus no further analysis and no new shapes are required to represent a text in Nasta'leeq style. This is shown in Figure 10.

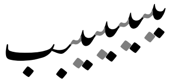


Figure 10: Recursive nature of Nasta'leeq

To ease the analysis, we can divide characters into different groups on the basis of similarity in shapes. For example, the set of characters shown in Figure 11 can be grouped under the name Beh_Family.

ب پ ت ٹ ث ټ ٻ ڀ ٿ ٽ ٺ

Figure 11: Beh_Family members

The basic shape of each character of Figure 11 is exactly the same except their Noktas (dots or marks) above or below. Similarly, we can divide all other characters into different groups. All different groups of characters are given in Table 11.

Sr.	Name	Members
1	Alef	ا آ ٱ ٲ ٳ أ إ
2	Beh	ب پ ت ٹ ث ټ ٻ ڀ ٿ ٽ ٺ
3	Jeem	ج چ ح خ ځ څ ڄ ڃ ڇ
4	Dal	د ڈ ذ ډ ڌ ڊ ڏ ڍ
5	Reh	ر ڑ ز ژ ړ ږ ڙ
6	Seen	س ش ښ
7	Sad	ص ض
8	Toain	ط ظ
9	Ain	ع غ
10	Feh	ف ڦ
11	Qaf	ق
12	Kaf	ک ک گ ګ ک ڳ ڱ
13	Lam	ل
14	Meem	م
15	Noon	ن ں ڻ ڼ ڻ
16	Waw	و ۄ ۆ
17	Heh	ہ
18	Heh-Doachashmee	ھ
19	Hamza	ء
20	Choti-Yeh	ی ي ی ی ې ێ ؠ
21	Bari-Yeh	ے

Table 11: Character families

In addition to all characters of Table 11, there exist certain ligatures that are treated like independent characters in Nasta'leeq. They are given in Figure 12. They act like independent characters that do not join with the following character in the ligature and have only two (independent and final) shapes.

ل + ا = لا، ک + ا = کا، گ + ا = گا

ک + ا = کا، ڭ + ا = ڭا، ڳ + ا = ڳا

Figure 12: Ligatures 1

4.1. Independent Shapes

All characters of Table 11, the ligatures of Figure 12, the punctuation marks and the special symbols of Table 6, the Urdu Numerals of Table 5 and the Arabic numerals are independent characters. In addition to the punctuation marks of Table 6, other English punctuation marks like single quotes, double quotes, colon, etc. are also included in Nasta'leeq.

There are certain special ligatures that are included in Nasta'leeq, *e.g.* Allah ligature (ﷲ), Muhammad ligature (محمد), etc. 23 other two character ligatures are also included in Nasta'leeq. In addition to all the above characters, Nasta'leeq also has a large set of diacritical marks that contains the diacritical marks of Arabic, Persian, Urdu, Punjabi, Pashto, Sindhi, Balochi, and Kashmiri. All these ligatures and diacritical marks are given in Table 12.

Sr.	Symbol	Sr.	Symbol	Sr.	Symbol
1	بسم اللہ الرحمن الرحیم	18	ڼے	35	◌ْ
2	ﷲ	19	ټے	36	◌َّ
3	محمد	20	یے	37	◌ٔ
4	ﷺ	21	ېے	38	◌ٕ
5	بے	22	ێے	39	◌ٖ
6	پے	23	ۑے	40	◌ؕ
7	تے	24	ئے	41	◌ٚ
8	ٹے	25	ۂ	42	◌ٓ
9	ثے	26	ؤ	43	◌ٛ
10	ٺے	27	ئ	44	◌٘
11	ٻے	28	◌ٗ	45	◌ٜ
12	ڀے	29	◌ؒ	46	◌ؐ
13	ٿے	30	◌ٍ	47	◌ؑ
14	ٽے	31	◌ٙ	48	◌ٰ
15	ڽے	32	◌ُ	49	◌ؓ
16	نے	33	◌ِ	50	◌ؔ
17	ڻے	34	◌ّ	51	◌ٟ

Table 12: Ligatures and diacritical marks

4.2. Two Characters Joining

We do the analysis of two characters joining in reverse order, i.e. first we identify the final shape for an initial shape, a context before. The group having only two shapes consists of Alef, Dal, Reh, Waw, two characters from Choti-Yeh_Family, ى (Alef Maskura) and ۍ (Pashto yeh with tail), Bari-Yeh, La and Ka families. Some of these characters have two final shapes depending on their joining behavior with different families, *e.g.* Reh_Family has two final shapes, one shape has only two (independent and final) shapes for Beh, Jeem, Kaf, Lam, Noon, Hamza and choti-yeh families and the other for the rest. The final shapes of 2-shapes families are given in Table 13.

Sr.	Shape	Examples
1	ا	با جا سا صا نا یا
2	د	بد جد سد صد ند ید
3	ر ر	بر جر سر صر نر یر
4	و و	بو جو سو صو نو یو
5	ی ی	بی جی سی صی نی یی
6	ے	جے سے صے ے طے
7	لا	بلا جلا سلا صلا نلا یلا
8	کا گا	بکا جکا سکا صکا نکا یکا

Table 13: Final shapes of Alef, Dal, Reh, Waw, Bari-yeh, La, Ka and two Choti-Yehs

Final shapes of the other families are given in Table 14.

Sr.	Shape	Examples
1	ٮ ٮ	بٹ جٹ سٹ صٹ نٹ یٹ
2	ح	بچ جچ سچ صچ نچ یچ
3	س س	بس جس سس صس نس یس
4	ص ص	بص جص سص صص نص یص
5	ط	بط جط سط صط نط یط
6	ع ع	بع جع سع صع نع یع
7	ڨ ف	بف جف سف صف نف یف
8	ق	بق جق سق صق نق یق
9	گ ک	بک جک سک صک نک یک
10	ل	بل جل سل صل نل یل
11	م	بم جم سم صم نم یم
12	ں	بں جں سں صں نں یں
13	ہ	بہ جہ سہ صہ نہ یہ
14	ھ	بھ جھ سھ صھ نھ یھ

Table 14: Final shapes

Hamza (ء) does not have a final shape. Thus there are 22 final families depending upon their final shapes, given in Table 13 and 14.

The above two tables not only give us the final shapes of all the families of Table 11 and of the ligatures of Figure 12 (La لا and Ka family ڳا، ڭا، کا، گا، کا), they also give us the analysis of initial shapes of the Beh, Jeem, Seen, Sad, Noon and Choti-yeh families. The analysis of initial shapes of

Beh, Noon, Hamza and Choti-yeh family in the above examples shows that they have the same base form for the initial shape with variations in Noktas. It is also clear that the initial form for final shapes of the Sad and Ain families are the same. Thus the Behinit family (including initial forms of Beh, Noon, Hamza and Choti-yeh families) has 21 initial shapes. The initial shapes of the Behinit and Jeeminit families are given in Table 15.

Sr.	Behinit Shape	Jeeminit Shapes	Final Families
1			Alef_Final
2			Beh_Final
3			Jeem_Final
4			Dal_Final
5			Reh_Final
6			Seen_Final
7			Sad_Ain_Final
8			Tah_Final
9			Feh_Final
10			Qaf_Final
11			Kaf_Final
12			Lam_Final
13			Meem_Final
14			Noon_Final
15			Waw_Final
16			Hehgol_Final
17			Heh-doachashmee_Final
18			Choti-Yeh_Final
19	*		Bari-yeh_Final
20			La_Final
21			Ka_Final
* Behinit family with Bari-yeh is stored as ligatures			

Table 15: Initial shapes of Beh and Jeem families

With 21 initial shapes of all families, all possible two character ligatures can be represented in Nasta'leeq. The Kaf and Lam families do not have an initial shape for Alef because these pairs are stored as ligatures, as shown in Figure 12.

4.3. Three Characters Joining

The final shapes have already been identified in the previous section. Similar to the initial shapes, 21 medial shapes are identified for the final shape families. The medial shapes of Behmedi and Jeemmedi families for final families are given in Table 16.

Sr.	Behmedi Shape	Jeemedi Shapes	Final Families
1			Alef_Final
2			Beh_Final
3			Jeem_Final
4			Dal_Final
5			Reh_Final
6			Seen_Final
7			Sad_Ain_Final
8			Tah_Final
9			Feh_Final
10			Qaf_Final
11			Kaf_Final, Gaf_Final
12			Lam_Final
13			Meem_Final
14			Noon_Final
15			Waw_Final
16			Hehgol_Final
17			Heh-doachashmee_Final
18			Choti-Yeh_Final
19			Bari-yeh_Final
20			La_Final
21			Ka_Final

Table 16: Medial shapes of Beh and Jeem families

The Behmedi shapes can be grouped into four different families according to their joining behavior with the previous character. This is shown in Table 17.

Name of Family	Shape	Members
Behmedi1		1, 2, 4, 7, 8, 9, 10, 11, 12, 15, 16, 19, 20, 21, 24, 25, 28, 29
Behmedi2		3, 13, 17, 18, 26, 30
Behmedi3		6, 14, 22, 23, 27
Behmedi4		5

Table 17: Behmedi families

For the families of Table 17, we need four initial shapes of each family having an initial shape. Thus the Behinit family has four new shapes for the Behmedi family, one shape for the Jeemmedi family and so on. All additional initial shapes of the Behinit and Jeeminit families, identified for medial shapes, are given in Table 18.

Sr.	Behinit Shape	Jeeminit Shapes	Medial Families
22			Behmedi1
23			Behmedi2
24			Behmedi3
25			Behmedi4
26			Jeemmedi
27			Seenmedi
28			Sadmedi, Tahmedi, Ainmedi, Fehmedi
29			Kafmedi, Gafmedi, Lammedi
30			Meemmedi, Hehgolmedi, Heh-doachashmeemedi

Table 18: More initial shapes of Beh and Jeem families

We have thus 30 initial shapes and 21 medial shapes that represent all possible ligatures of length three of the six major Pakistani languages when written in the Noori Nasta'leeq style. It is not possible to list all shapes of all characters due to space shortage.

4.4. Four Characters Joining

We do our analysis in the reverse direction, i.e. from left-to-right. In the analysis of three characters joining, we have already identified the shapes of the last two characters of our ligatures of length 4 that are final shapes and medial shapes for our final shapes. Now first we need to identify the medial shapes that will join with the already identified medial shapes. Secondly, we need to identify the initial shapes that will join with newly identified medial shapes in the previous step and this will complete our joining analysis.

Sr.	Behinit Shape	Jeeminit Shapes	Medial Families
22			Behmedi1
23			Behmedi2
24			Behmedi3
25			Behmedi4
26			Jeemmedi
27			Seenmedi
28			Sadmedi, Tahmedi, Ainmedi, Fehmedi
29			Kafmedi, Gafmedi, Lammedi
30			Meemmedi, Hehgolmedi, Heh-doachashmeemedi

Table 19: More medial shapes of Beh and Jeem families

The process of identifying the new medial shapes is the same as that used to identify the initial shapes for the first 21 medial shapes. Similar to the Behinit family, the Behmedi family also has four new shapes for its first 21 members, one shape for the Jeemmedi family and so on. All additional medial shapes of the Behmedi and Jeemmedi families, identified for medial shapes, are given in Table 19.

Table 17 shows that the Behmedi2 family includes the medial shapes # 26 and 30. Thus, fortunately, we do not have new initial shapes for these newly identified medial shapes of Table 19. Hence, our analysis for Noori Nasta'leeq style is complete.

Ligatures longer than 4 can be built using recursively the shapes already identified. That is shown in Figure 10. We have 1 or 2 final shapes, 30 initial shapes and 30 medial shapes for the characters of major Pakistani languages. Thus we need more than 1300 glyphs to represent the scripts of major Pakistani languages in the Noori Nasta'leeq style or build a good looking font for these languages.

5. CONTEXT-SENSITIVE SUBSTITUTION GRAMMAR

The analysis given in Section 4 can be represented in the *Context-Sensitive Substitution Grammar*. Figure 13 shows some rules of the contextual substitution grammar of Nasta'leeq.

Initial Rule
beh → behinit1 aiknoktabelow
jeem → jeeminit1 aiknoktabelow
No Context (Before | After)
Medial Rule
Beh → behmedi1 aiknoktabelow
Jeem → jeemmedi1 aiknoktabelow
No Context (Before | After)
Final Rule
beh → behfina1
jeem → jeemfina
No Context (Before | After)
Contextual Substitution Rule for Behfina1
behinit1 → behinit2
jeeminit1 → jeeminit2
behmedi1 → behmedi2
jeemmedi1 → jeemmedi2
Context (| behfina1)
Contextual Substitution Rule for Jeemfina1
behinit1 → behinit3
jeeminit1 → jeeminit3
behmedi1 → behmedi3
jeemmedi1 → jeemmedi3
Context (| jeemfina)
Contextual Substitution Rule for Behmedi1 Family
behinit1 → behinit22
jeeminit1 → jeeminit22
behmedi1 → behmedi22
jeemmedi1 → jeemmedi22
Context (| <behmedi1 Family>)

Figure 13: Context-Sensitive Substitution Grammar

The Initial Rule tells that Beh (ب) and Jeem (ج) are substituted by behinit1 (ﺑ) and jeeminit1 (ﺟ) respectively with appropriate Nokta on them whenever they come at the initial position of a ligature. Medial and Final rules also have the same kind of interpretation for the medial and final positions respectively. The Contextual Substitution Rule for Behfina1 tells that default initial shapes behinit1 (ﺑ) and jeeminit1 (ﺟ) at the initial position are substituted by behinit2 (ﺑ) and jeeminit2 (ﺟ) when they are followed by a glyph of the Behfina1 family. It also tells that default medial shapes behmedi1 (ﺒ) and jeemmedi1 (ﺠ) at the medial position are substituted with behmedi2 (ﺒ) and jeeminit2 (ﺠ) when they are followed by a character of the Behfina1 family. The other rules have the same kind of interpretations. Figure 13 shows a very small part of the *Context-Sensitive Substitution Grammar* of Noori Nasta'leeq. This shows the contextual nature and complexity of the Noori Nasta'leeq style. Theoretically, the *Context-Sensitive Substitution Grammar* is a computational model of the Noori Nasta'leeq contextual complexity.

6. CONCLUSION

Nasta'leeq is a bidirectional, diagonal, non-monotonic, cursive, highly context-sensitive and very complex writing system for languages written in the Arabic or in extended Arabic scripts like those of Urdu, Punjabi, Pashto, Sindhi, Balochi, Kashmiri, *etc*. The analysis of Nasta'leeq for major Pakistani languages applies equally to Arabic, Persian and other languages written in extended Arabic scripts. The analysis of Nasta'leeq and the *Context-Sensitive Substitution Grammar*, discussed in this paper, can be used to build a good quality and high speed font for the Arabic, Persian, Urdu, Punjabi, Pashto, Sindhi, Balochi and Kashmiri languages to write them in the Noori Nasta'leeq style.

The practical implementation of a character-based Nasta'leeq font for Arabic, Persian and Pakistani languages is a much more complex process than its theoretical analysis. A practical development of a Nasta'leeq font not only needs the *Context-Sensitive Substitution Grammar*, but it also requires other important and vital positioning information to correctly position glyphs and Noktas considering their contextual glyphs and Noktas, as shown in Figure 10. An implementation for Urdu and Punjabi has been produced by the first author in 2004 and is available as a freeware on the Web at *www.puran.info*[3]. Practical details cannot be discussed here due to shortage of space. We plan to discuss it in a future paper. Digital graphical representation in a computer is vital for under-resourced languages, so that native people can understand their native languages and can contribute to the development of computational linguistic resources for their languages.

7. REFERENCES

[1] Rahman, T. "*Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift*", in proc. Crossing the Digital Divide, SCALLA Conference on Computational Linguistics, pp. 5-7 January, 2004.

[2] Grimes, B. F. "*Pakistan*". Ethnologue: Languages of the World. 14th Edition Dallas, Texas; Summer Institute of Linguistics, 2000.

[3] Afzal, M. and Hussain, S. "*Urdu Computing Standards: Development of Urdu Zabta Takhti (UZT) 1.01*". in proc. INMIC-2001, Lahore, 2001.

[4] Khaver, Z. "*Standard Code Table for Urdu*", in proc. 4th Symposium on Multilingual Information Processing (MLIT-4), Yangon, Myanmar, CICC, japan, 1999.

[5] Malik, M. G. Abbas; "*Towards a Unicode Compatible Punjabi Character Set*". In proc. 27th Internationalization and Unicode Conference, Berlin, Germany, 2005.

[6] Malik, M. G. Abbas; "*Punjabi Machine Transliteration*". In proc. 21st International Conference on Computational Linguisitcs COLING-06 and 44th Annual Meeting of ACL, Sydney, Australia, 2006.

[7] Malik, M. G. Abbas; Boitet, Christian; and Bhattcharyya, Pushpak; "*Hindi Urdu Machine Transliteration using Finite-state Transducers*". In proc. 22nd International Conference on Computational Linguistics COLING-08, Manchester, UK, 2008.

[8] Platts, J. T. *A Grammar of the Hindustani or Urdu Language*. Crosby Lockwood and Son, 7 Stationers Hall Court, Ludgate hill, London. E.C., 1909.

[9] Hussain, S. "*Letter to Sound Rules for Urdu Text to Speech System*", in Proc. of Workshop on "Computational Approaches to Arabic Script-based Languages", COLING-04, Geneva, Switzerland, 2004.

[10] Wali, A., Hussain, S., "*Context Sensitive Shape-Substitution in Nastaliq Writing System: an analysis and fomulation*". In Proc. of International Joint Conferences on Computer, Information and Systems Sciences and Engeenering, 2006.

[3] www.puran.info

OPTIMIZATION ON VIETNAMESE LARGE VOCABULARY SPEECH RECOGNITION

Ngoc Thang Vu, Tanja Schultz

Cognitive Systems Lab (CSL), Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)

ABSTRACT

This paper summarizes our latest efforts toward a large vocabulary speech recognition system for Vietnamese. We describe the Vietnamese text and speech database which we collected as part of our GlobalPhone corpus. Based on these data we improve our initial Vietnamese recognition system [1] by applying various state-of-the art techniques such as semi-tied covariance and discriminative training. Furthermore, we achieve significant improvements by building two systems based on different tone modeling approaches and then apply system cross-adaptation and confusion networks combination. The best Vietnamese speech recognition system employs a 3-pass decoding strategy and achieves a syllable-based error rate of 7.9% on read newspaper speech. In addition, we perform initial experiments on the Voice of Vietnam (VOV) speech corpus [2] and achieve a syllable error rate of 16.5%.

Index Terms— Vietnamese speech recognition, data collection, discriminative training, system combination

1. INTRODUCTION

The performance of speech and language processing technologies has improved dramatically and an increasing number of systems are being deployed in a large variety of applications. To date, most efforts were focused on a very small number of languages spoken by a large number of speakers in countries of great economic potential, and a population with immediate information technology needs. With more than 6900 languages in the world and the need to support multiple input and output languages, the most important challenge today is to port or adapt speech processing systems to unsupported languages rapidly and at reasonable costs. Despite the fact that the Vietnamese language is spoken by more than 80 Million people and thus is listed among the top-25 languages, there is a surprisingly small number of groups investigating Vietnamese speech and language processing technologies and applications, with notable exceptions like IOIT [2] and MICA [3].

Last year we started to applying our Rapid Language Adaptation Tools (RLAT) [4] to Vietnamese. In [1] we reported on our development and optimization of a Vietnamese large vocabulary speech recognition system and described particular characteristics of the Vietnamese language, such as the monosyllabic structure and tonality of the sound system. Our best system achieved a syllable error rate (SyllER) of 12.6 % on the development and 11.7% on the evaluation set. However, this initial system did not employ the full range of state-of-the-art techniques, which have shown to be very effective for high-resource languages. In this paper we apply these techniques to our initial Vietnamese system and study to what extend the reported performance improvements on languages like English and Chinese apply to Vietnamese. Among the state-of-the-art techniques we applied are semi-tied covariances [5], discriminative training [6], system cross adaptation, and confusion network combination [7].

The paper is organized as follows. In Section II we describe our Vietnamese resources, which consist of a audio data and corresponding transcriptions in the newspaper domain, and a large text corpus harvested from the internet on the same domain. Section III introduces our baseline recognition system which was presented in [1]. In Section IV we give a detailed description of the optimization steps and report recognition results on the development and evaluation set. The study is concluded in Section VI with a summary and an outlook to future steps.

2. VIETNAMESE LANGUAGE PECULIARITIES

Vietnamese is a language with very interesting characteristics, three of which are particularly challenging for automatic speech recognition. The first peculiarity is the monosyllabic nature of Vietnamese. For example the sentence ”Xin chao Viet Nam“ (in English: hello Vietnam) contains 4 word units, each consisting of a single syllable. This monosyllabic nature poses two problems to speech recognition, i.e. due to the shortness, the word units are acoustically confusable and the short units limit the language model history. In [1] we compensated the restricted language model history by concatenating monosyllabic words to multisyllabic words. After concatenation, the example sentence from above looks like ”Xin_chao Viet_Nam“. The sentence has now 2 multisyllabic words. Multisyllabic words achieve significant improvements ranging from 10% to 20% relative, depending on the tone modeling approaches.

The second peculiarity of the Vietnamese language is the tonality of the sound system. Vietnamese has six different tones, which can discriminate the meaning of words. So, it is advisable to use tone information in the acoustic model. In [1] we extracted pitch information using the Cepstrum and gained about 6% to 9% relative improvement depending on the tone modeling approaches.

The third important characteristic of Vietnamese results from the large amount of diphthongs and triphthongs in the phoneme set. In total, Vietnamese has 22 consonants, 11 vowels, 21 diphthongs and 3 triphthongs. So, compared to languages like English or French, the number of diphthongs and triphthongs is pretty high. In addition to the large number, some of these phonemes are very rare, and thus may lead to poorly estimated acoustic models. While it is possible to collapse the phone set by subsuming the rare phonemes under their closest match, or by splitting the rare diphthongs and triphthongs into their respective monophthongs parts, both approaches have disadvantages. Collapsing the phoneme set results in an increased confusability, and splitting up diphthongs and triphthongs overestimates the phoneme duration. Therefore, we decided in our study to collect additional data to cover rare diphthongs and triphthongs. As reported in [1] we achieved about 8% relative improvement. These gains suggest that for Vietnamese speech recognition care needs to be taken to collect a corpus such that it covers all phonemes.

3. VIETNAMESE LANGUAGE RESOURCES

The development of a state-of-the-art speech recognition system starts with collecting speech data and corresponding transcriptions, as well as written text resources for vocabulary selection and language modeling. Data collection is an extremely time and cost consuming task but its careful excecution is crucial to the performance of the final system. We applied our Rapid Language Adaptation Tools (RLAT) [4], which allow us to collect massiv amounts of text data from the web and to record speech data over the Internet using a web-based recorder. In the following subsections we describe the collected corpus for Vietnamese language that was collected in 2009 as part of our GlobalPhone project [8].

3.1. Text Corpus

For the text corpus of Vietnamese words we used RLAT to collect text from fifteen different websites, covering main Vietnamese newspaper sources. RLAT enables the user to crawl text from a given webpage with different link depths. The websites were crawled with a link depth of 5 or 10, i.e. we captured the content of the given webpage, then followed all links of that page to crawl the content of the successor pages (link level 2) and so forth until we reached the specified link depth. After collecting the Vietnamese text content of all pages, the text was cleaned and normalized with four different steps: (1) Remove all HTML-Tags and codes, (2) Remove special characters and empty lines, (3) Delete lines with less than 75% tonal words (identification of Vietnamese language) and (4) Delete line which appear repeatedly. The first twelve websites of Table 1 were used to build the language model (see below). The text from the remaining three websites was used to select prompts for recording speech data for the development and evaluation set. In total we collected roughly 40 Million Vietnamese word tokens (see 4 below).

Table 1. List of all 15 Vietnamese websites

Websites	Link depth
www.tintuconline.vn	10
www.nhandan.org.vn	10
www.tuoitre.org.vn	10
www.tinmoi.com.vn	5
www.laodong.com.vn	5
www.tet.tintuconline.com.vn	5
www.anninhthudo.vn	5
www.thanhnien.com.vn	5
www.baomoi.com	5
www.ca.cand.com.vn	5
www.vnn.vn	5
www.tinthethao.com.vn	5
www.thethaovanhoa.vn	5
www.vnexpress.net	5
www.dantri.com	5

3.2. Speech Corpus

3.2.1. GlobalPhone Data

To collect Vietnamese speech data in a very short time, the author spent one month in Vietname and recruited friends and relatives to donate their voice for research. The web-based recording tool turned out to be difficult as many sites in Vietname did not provide Internet connection, so we used an offline version of the same recording tools. In order to control the quality of recordings and to avoid the amount of transcription work, we collected Vietnamese speech data in GlobalPhone style [8], i.e. we asked native speakers of Vietnamese to read prompted sentences of newspaper articles. The resulting corpus consists of 25 hours of speech data spoken by 140 native speakers, from the cities of Hanoi and Ho Chi Minh City in Vietnam as well as 20 native speakers living in Karlsruhe, Germany. Each speaker read between 50 and 200 utterances which were collected from the above listed 15 different Vietnamese websites. In total the corpus contains 22.112 utterances spoken by 90 male and 70 female speakers. All speech data was recorded with a headset microphone in clean environmental conditions. The data is sampled at 16 kHz with a resolution of 16 bits and stored at PCM encoding. The Vietnamese portion of the GlobalPhone database is listed in Table 2.

Table 2. Vietnamese GlobalPhone Speech corpus

Set	#Speakers		#Utterances	Duration
	Male	Female		
Training	78	62	19596	22h 15min
Development	6	4	1291	1h 40min
Evaluation	6	4	1225	1h 30min
Total	90	70	22112	25h 25min

3.2.2. *Voice of Vietnam Data*

The Voice of Vietnam (VOV) speech corpus was collected in 2005 by IOIT and kindly provided to us for research purposes [2]. The VOV data is a collection of story reading, VOV mailbag, news report and colloqiums from the radio program ”The Voice of Vietnam”. The database consists of are 22549 audio files with transcriptions from 30 male and female broadcasters and visitors. The number of distinct syllables with tone is 4923 and the number of distinct syllables without tone is 2101 [2]. The VOV corpus covers all Vietnamese phonemes and most Vietnamese syllables. The data is provided in wav format, using a sampling rate of 16kHz and A/D conversion precision of 16 bits. We splitted the VOV data in a training and testing part. Table 3 shows the relevant information about the VOV corpus for the training and the test set.

Table 3. The Voice of Vietnam Speech corpus

Set	#Utterances	Duration
Training	20990	19h 31min
Testing	1459	1h 18min
Total	22549	20h 49min

3.3. Language Model

Based on the crawled text corpus (see above), we built a statistical n-gram language model using the SRI language model toolkit [9]. We trained a 5-gram language model on the cleaned and normalized text data from the 12 first websites listed in Table 1. Table 4 gives the characteristics of the language models calculated on the GlobalPhone development set, evaluation set, and VOV test set.

3.4. Pronounciation Dictionary

Next to the speech and text data, the pronounciation dictionary is a very important part of an automatic speech recognition system. The dictionary guides the decoder and ensures proper training alignment. We used the RLAT tools to generate the dictionary. In RLAT an interactiv rule-based lexlearner is implemented which enable the user to learn pronunciation rules by providing initial letter-to-sound mappings and interactively confirming or correcting pronunciation examples as proposed by the lexlearner. We took the RLAT dictionary and performed some manual corrections. More particularly, we wanted to model the impact of dialectal variations by using pronunciation variants. The data were intentionally collected in the North and South of Vietnam and many words are spoken different between the Northern and Southern dialect. Table 5 shows some examples from our pronunciation dictionary applying pronunciation variants.

Table 4. Performance of LM in development and evaluation set

Criteria	GP-Dev	GP-Eval	VOV-Test
# word tokens	39043284		
# vocabulary	29967		
OOV-Rate (%)	0	0.067	0.11
Perplexity	282	277	392
Coverage (%):			
1-gram	100	99.94	99.89
2-gram	93.4	92.60	92.99
3-gram	60	54.02	54.84
4-gram	32.6	24.2	20.01
5-gram	21.3	12.1	5.8

Table 5. Pronunciation dictionary with different variants for Northern and Southern dialect in Vietnamese

Words	Pronunciation
xin_chao	{x i11 n ch ao2}
vo	{v o36}
vo(1)	{j o36}
ra	{r a11}
ra(1)	{d1 a11}

4. BASELINE RECOGNITION

To model the tonal structure of Vietnamese we explored two different acoustic modeling schemes. In the so-called ”‘Explicite tone modeling”’ (ETM) scheme all tonal phonemes (vowels, diphthongs, and triphthongs) are modeled with 6 different models, one per tone. For example, the vowel ’a’ is represented by the models ’a1’, ’a2’, ..., ’a6’, where the numerals identify the tones. In the so-called ”‘Data-driven tone modeling”’ (DDTM) we used only one model for all tonal variants of a phoneme, i.e. vowel ’a’ is represented by only one model ’a’. However, the information about the tone was added to the dictionary in form of a tone tag. The Janus Recognition Toolkit (JRTk) [10] allows using these tags as questions to be asked in the context decision tree when building context dependent acoustic models. This way, the data will decide during model clustering if two tones have a similar impact on the basic phoneme. If so, the two tonal variants of that basic phoneme would share one common model. In case the tone is distinctive (of that phoneme and/or its context), the question about the tone may result in a decision tree split, such that different tonal variants of the same basic phonemes would end up being represented by different models. For context dependent acoustic modeling we stopped the decision tree splitting process at 2500 quintphones for both schemes, the explicite and the data-driven tone modeling. Table 6 describes the phoneme set and the relevant characteristics of the two different tone modeling schemes as used in the experiments reported below. While the number of basic model units is quite different for the two modeling schemes, the number of context dependent models was controlled to be the same for both schemes for better comparison. After context clustering, a merge&split training was applied, which selects the number of Gaussians according to the amount of data. Please note that the "'Explicite tone modeling"' uses about 16% fewer Gaussians than the "'Data-driven tone modeling"'. This is a result from the fact that many tonal variants, particularly diphthongs and triphthongs are very rare and are thus modeled with a small number of Gaussians. The preprocessing consists of feature extraction applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 164 dimensions containing two main parts. The first part has 143 dimensions which were extracted by stacking 11 adjacent frames of 13 coefficient MFCC frames. The second part describes the tone information. We computed the Cepstrum with a window length of 40ms and detected the position of the maximum of all cepstral coefficients starting with the 30th coefficient. Furthermore, we considered the position of the three left and right neighbors, and their first and second derivatives. This resulted in 21 additional coefficients (1 maximum, 3 left neighbors, 3 right neighbors plus the first and second order derivatives). With an LDA transformation we finally reduced this set to 42 dimensions. The acoustic model uses a semi-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. The language model and the pronunciation dictionary are based on bisyllable words. Table 7 shows the Syllabic Error Rate (SyllER) performance of the resulting baseline Vietnamese recognizer on the development set after merge-and-split training and 6 iterations of Viterbi training.

Table 6. Phoneme set and model size

	Explicite tone modeling	**Data-driven tone modeling**
# Consonants	22	22
# Vowels	66	11
# Diphthongs	126	21
# Triphthongs	24	4
$\sum$ Phonemes	238	58
# CI Acoustic Models	715	175
# CD Acoustic Models	2500	2500
# Gaussians (Merge-&-Split)	111421	130263

Table 7. SyllER of the baseline system on development set

Systems	GP Dev-Set
Explicite tone modeling	12.8%
Data-driven tone modeling	12.6%

5. SYSTEM OPTIMIZATION

In this section we describe the steps and techniques taken to optimize the performance of the recognition system. As a first step we applied semi-tied covariances [5] to make the system more robust, for example if training data and test data were recorded in different environments. Second, we ran discriminative training [6] and describe the effect on our speech recognizer. Third, we used cross-adaptation, one of the multi-pass decoding strategies, to combine the advantages of the two different tone modeling approaches, which were implemented as described above. Finally, to minimize the syllabic error rate we used confusion network combination [7] which allows to extract better hypothesis from a combination of two or more systems.

5.1. Semi-tied Covariance Matrices

There is normally a simple choice made in form of the covariance matrix to be used with continuous-density HMMs. Either a diagonal covariance matrix is used, with the underlying assumption that elements of the feature vector are independent, or a full or block-diagonal matrix is used, where all or some of the correlations are explicitly modeled. Unfortunately, full or block-diagonal covariance matrices come with a dramatic increase in the number of parameters per Gaussian component, and thus limiting the number of components which may be estimated robustly. Semi-tied covariance matrices (STC) [5] are a form of covariance matrix which allows a few full covariance matrices to be shared over many distributions, whereas each distribution contains its own diagonal covariance matrix. Furthermore, this technique fits well within the standard maximum-likelihood criterion used for HMM training. Table 8 shows the SyllER performance of the Vietnamese recognizer on the development set after applying semi-tied covariance matrices.

Table 8. SyllER after using Semi-tied Covariance Matrices

Systems	Dev-Set
Explicite tone modeling	11.9%
Data-driven tone modeling	11.8%

After this step we retuned the language model weights and word insertion penalties by rescoring the word lattices on the development set. This gave another improvement of about 4% relative in SyllER. Table 9 shows our results on the development set.

Table 9. SyllER after Language Model Retuning

Systems	Dev-Set
Explicite tone modeling	11.7%
Data-driven tone modeling	11.4%

5.2. Discriminative training (DT)

Discriminative training is an essential technique that consistently leads to significant improvements in speech recognition accuracy. Maximum mutual information estimation (MMIE) [11] and boosted MMIE [6] are common techniques for discriminative training. We applied this technique to our Vietnamese speech recognizer system. Starting with the speaker-independent model using maximum likelihood estimation, we decoded the complete set of training utterances in order to generate word lattices.

MMIE aims at maximizing the posterior probability of a reference compared to the competing hypotheses in a word lattice. The objective function of MMIE is:

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} log \frac{P_\lambda(X_r|M_{s_r})P(s_r)}{\sum_s P_\lambda(X_r|M_s)P(s)}$$

where λ represents model parameters to be optimized; X_r is the r-th training utterance; s_r is the reference and M_s represents the corresponding HMM state sequence of sentence s. Maximizing F_{MMI} improves the posterior probability of the reference in the lattice.

Intuitively, some paths may contain more error than other parts in a word lattice. Boosted MMIE boosts the importance of competitors that make large errors and aims to improve the confusable parts. Table 10 shows our results on the development set after applying the discriminative training. So far, we do not have a good explanation why the gains are smaller than expected.

Table 10. SyllER after applying discriminative training

Systems	Dev-Set
Explicite tone modeling	11.56%
Data-driven tone modeling	11.15%

5.3. Multi-pass decoding: Cross Adaptation

State-of-the-art speech recognition systems commonly use multi-pass decoding with an adaptation of the acoustic model between passes. Adaptation aims at better fitting the system to the speakers and/or acoustic environments found in the test data. The two most popular adaptation methods, which can be found in many systems, are Maxmum Likelihood Linear Regression MLLR, a model transformation and Feature Space Adaptation FSA, a feature transformation. Adaptation is performed in an unsupervised manner, so that the hypotheses obtained from the previous decoding pass are taken as the necessary reference for adaptation. Generally, the word error rates of the hypotheses obtained from the adapted systems are lower than without adaptation. This sequences of adaptation and decoding make it possible to incrementally improve the system, but not always lead to significant improvements. Often, after two or three stages of adapting a system on its own output, no more gains can be obtained. This problem can be solved by adapting a system on the output of a different system, a process called cross-system adaption. In this paper we developed distinct systems with two different approaches for tone modeling. Therefore, it is possible to apply cross-system adaptation. Furthermore, for each tone modeling approach we had two different systems: a Speaker Independent (SI) and a Speaker Adaptive (SA) using FSA and MLLR. So we experimented with various possible system combination to find the best performing decoding strategy. As first pass we always apply the SI system. The second and third pass systems are speaker adaptive system. Furthermore, the third pass system could apply the discriminative training. Table 11 shows the results on the development set after applying the various options of cross-system adaptation.

Table 11. SyllER after using Cross Adaptation

Systems	Dev-Set
ETM x DDTM x ETM (S1)	8.7%
ETM x DDT x ETM+DT (S2)	8.4%
ETM x ETM x DDTM (S3)	8.6%
ETM x ETM x DDTM+DT (S4)	8.6%
DDTM x ETM x DDTM (S5)	8.7%
DDTM x ETM x DDTM+DT (S6)	8.6%
DDTM x DDTM x ETM (S7)	8.7%
DDTM x DDTM x ETM+DT (S8)	8.5%

5.4. Confusion Network Combination

After applying the cross adaptation techniques we got different word lattices which contain alternative hypotheses. Consequently, we applied the confusion network combination technique [7] to combine these lattices and subsequently extract the best hypothesis. We experimented with different lattice combinations. The best combination gave 0.2% absolute improvement. Table 12 shows the all results on the development set after applying confusion network combination.

5.5. Decoding strategy

After the optimization steps on the development set we obtained the best decoding strategy. Two parallel systems decode the audio data and write the word lattices. After that we used confusion networks (CN) to combine these lattices and extract the best hypothesis. The first system (S1) contains 3

Table 12. SyllER after using Confusion Network Combination

Systems	Dev-Set
S2 x S6	8.2%
S2 x S8	8.4%
S2 x S4	8.3%
S6 x S8	8.3%
S4 x S6	8.5%
S2 x S4 x S6	8.4%
S2 x S4 x S8	8.4%
S4 x S6 x S8	8.5%
S2 x S4 x S6 x S8	8.3%

passes: ETM-SI, DDTM-SA, and ETM-SA using DT. The second system (S2) contains also 3 passes: DDTM-SI, ETM-SA and DDTM-SA using DT. We tested our system on the unseen evaluation set using this decoding strategy. Table 13 illustrates the results on the evaluation set.

Table 13. SyllER on the evaluation set using the best decoding strategy

	1.Pass	2.Pass	3.Pass	CN
S1	11.4%	8.7%	8.1%	7.9%
S2	10.8%	8.8%	8.2%	7.9%

5.6. Experiments and Optimization on VOV Data

5.6.1. Experiments with VOV data

The VOV corpus was collected from the audio program ”Voice of Vietnam”. It has substantially different characteristics compared to the GlobalPhone data. As a result the VOV data provide us with a good test case to explore how well our Vietnamese speech recognizer generates. The first experiment applied the “Explicite-tone modeling system“ (ETM) to decode the VOV test set and gave 24.1% SyllER. In the second experiment we trained the speech recognition system on the VOV training data and tested on the VOV test data. We used the ETM system to write the initial alignments for the complete VOV training set. We used these initial alignments to train the system. For system training we applied the same parameter settings as we used to train our best GlobalPhone system. The performance on the VOV test set slightly improves to 23.5% but gets drastically worse on the GlobalPhone development set with 33.4% SyllER. According to our analysis, we believe that the reason for the degradation is that the VOV corpus contains only Northern dialect data, while the GlobalPhone data set covers Northern and Southern dialect. The breakdown for dialects shows that the GlobalPhone part with Northern dialect achieved a performance of 19.6% SyllER, while the Southern dialect part significantly dropped in performance to 51.7% SyllER. So, training on Northern-only VOV data significantly harms the performance on the part of GlobalPhone spoken by Southern Vietnamese speakers. In our last experiment we trained the acoustic model with a combination of GlobalPhone and VOV training data. The results are given in Table 14 and show improvements of about 25% relative on the VOV test set, but 5% degradation on the GlobalPhone development set. A subsequent error analyis of these results indicate that the majority of errors stem from the following issues: (1) large number of proper names, sometimes even a sequence of several proper names, (2) interruptions, unfinished utterances (3) Foreign proper names, most particular English, such as Canada, Vovnews and Singapore. In the following section we describe how the language model was trained to better handle proper names and compensate for the above described issues.

Table 14. SyllER on the VOV test set and GP development set using the speaker independent system

Training-Set	VOV Test	GP dev
GP Daten	24.1%	11.9%
VOV Daten	23.5%	33.4%
VOV+GP Daten	17.8%	12.5%

5.6.2. System Optimization on VOV data

In order to adapt our language model to the VOV test set, we used the RLAT system to crawl the VOV mailbag from 22-12-2008 to 22-12-2009 and built a 3-gram language model ”VOVmail“. Linear interpolation [9] was applied to combine the background and VOVmail language model (LM). The best mixture weight is 0.57 for the background LM and 0.43 for the VOVmail language model. To solve the problem with proper names, we randomly generated 1 million full names and built a 3-gram language model called ”FullName“. A Vietnamese proper name contains usually three parts: surname, middle name, and firstname. In our work we used the 20 most common surnames, the 35 most common middle names, and 65 of the most common first names and combined them randomly. After that we interpolated the three language models and decoded the VOV test set. Table 15 compares the performance of the baseline language model (background), the interpolation with the VOVmail-based language model (+VOVmails), and the interpolation with the VOVmail data and the automatically generated corpus of full names. The results show that the new language model shows significant perplexity reduction on the VOV test data. Our currently best system gives a SyllER of 16.5% on the VOV test set using the interpolation of all three corpora. This is a gain of 7% relative over the baseline language model.

6. CONCLUSION

In this paper we describes our latest improvements to our Vietnamese speech recognition system for large vocabulary.

Table 15. Optimizing LM on VOV dev set

Criteria	Background	+VOVmails	+FullName
OOV-Rate (%)	0.11	0.04	0.04
Perplexity	392	250.4	245.9
Coverage (%):			
1-gram	99.89	99.96	99.96
2-gram	92.99	94.2	94.26
3-gram	54.84	57.05	57.6
4-gram	20.01		20.01
5-gram	5.8	5.8	5.8

The speech corpus as a part of GlobalPhone was used with 25 hours audio data from 160 Vietnamese speakers reading newspaper articles. Applying our Rapid Language Adaptation Tools, we collected about 40 Mio words from 15 different websites for language model training and prompt selection. We subsequently applied state-of-the-art techniques, such as semi-tied covariance matrices, discriminative training, cross adaptation, and confusion network combination to study the impact on Vietnamese speech recognition and to improve our system. Starting from a baseline system with 12.6 % SyllER, we improved the system to 8.2% on the development set, and reduced the error from 11.7% to 7.9% on the evaluation set. The impact of the various optimization steps and the best decoding strategy are summerized in Table 16 and Table 17. Future steps will include further improvements of tone modeling, language modeling, and a more detailed investigation of the effects of dialects.

Table 16. System Optimization

System (SI)	**Explicite** **tone modeling**	**Data-driven** **tone modeling**
Baseline	12.8%	12.6%
Optimal Feature	11.9%	11.8%
LM Tuning	11.7%	11.4%
Discriminative Training	11.56%	11.15%

Table 17. SyllER on development set using the best decoding strategy

Decoding-Pass	**S1**	**S2**
1.Pass	11.7%	11.6%
2.Pass	9.0%	9.0%
3.Pass	8.4%	8.6%
Confusion Network	8.2%	8.2%

7. ACKNOWLEDGMENT

The authors are very grateful to Dr. Luong Chi Mai from IOIT for providing us the VOV speech corpus. We also would like to thank Roger Hsiao for his support and useful discussions.

8. REFERENCES

[1] Ngoc Thang Vu and Tanja Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In: ASRU, Italy 2009.

[2] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong and John-Paul Hosom. Vietnamese Large Vocabulary Continuous Speech Recognition. In: 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005.

[3] Nguyen Hong Quang, Pascal Nocera, Eric Castelli and Trinh Van Loan. A Novel Approach in Continous Speech Recognition for Vietnamese, an isolating tonal language. In: SLTU, Hanoi, Vietnam, 2008.

[4] Tanja Schultz and Alan Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc. ICASSP Las Vegas, NV 2008.

[5] Mark Gales, Semi-tied covariance matrices for hidden Markov models. In: IEEE Transactions Speech and Audio Processing, vol. 7, pp. 272-281, 1999.

[6] Dan Povey, D. Kanevsky, Brian Kingsbury, B. Ramabhadran, George Saon and K. Visweswariah. Boosted MMI for model and feature-space discriminative training. In: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing, 2008.

[7] Lidia Mangu, Eric Brill and Andreas Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In Proc. of EUROSPEECH'99, Budapest, Hungary.

[8] Tanja Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[9] Andreas Stolcke. SRILM - an extensible language modeling toolkit, in Proceedings of ICSLP, 2002.

[10] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," International Conference on Acoustics, Speech, and Signal Processing 1997, ICASSP, Munich; Germany.

[11] V. Valchev, J. Odell, Phil Woodland and Steve Young. MMIE training of large vocabulary speech recognition systems. In: Speech Communication, vol. 22, pp. 303-314, 1997.

VOICE AIDED INPUT FOR PHRASE SELECTION USING A LOW LEVEL ASR APPROACH – APPLICATION TO FRENCH AND KHMER PHRASEBOOKS

Sereysethy Touch', Laurent Besacier*, Eric Castelli', Christian Boitet**

* LIG - GETALP, rue de la Bibliothèque, B.P. 53 38041 Grenoble Cedex 9 France
' Centre MICA – HUT – CNRS/UMI2954 – Grenoble INP - 1 Dai Co Viet, Hanoi, Vietnam
E-mail: Sereysethy.Touch@imag.fr

ABSTRACT

We report the ongoing results of an effort to embed a "light" ASR for a future smart-phone-based multimodal multilingual phrase-book which allows users to look for a sentence by simply pronouncing it. We compared a phoneme-based *low level* approach with a conventional word-based *high level* approach. The former approach has been found promising in terms of accuracy and performance in a restricted task-oriented domain suitable for handheld devices with low-resources. The experiments have been performed on both high- and under-resourced languages: French and Khmer.

***Index Terms*—** Phrasebook, ASR, embedded system

1. INTRODUCTION

With the increase of people movement, in the context of globalization and international exchange, it is essential for visitors to be able to speak and communicate with local people although they do not speak their language. With the trend toward ubiquitous computing, smartphones start to take place in everyday life of people. It is therefore interesting if we could use them as a survival linguistic kit enabling us to translate instantly our speech into a target language and this is what we will address in this paper.

We present some results of an ongoing work in an effort to overcome problems related to the design and development of a multimodal phrasebook system running on smartphones. It allows user to choose a sentence from a collection of sentences by voice input. The selected sentence is then automatically translated into a target language. Allowed voice input to be directly computed on a smartphone raises some great challenges since they have limited resources (CPU, RAM, etc.) and the system has to be able to answer instantly for obvious usability reasons. Consequently, we investigate and experiment different approaches used in automatic speech recognition system (ASR): 1) high level (HL) and 2) low level (LL) which correspond respectively to 1) conventional word-based ASR (approach using an n-gram word language model) and 2) phoneme-based approach (without any language model) in order to choose a suitable approach which can be further deployed on smartphones.

In the next section, we describe different approaches used in ASR. Then, in Section 3 we will detail the experimentation protocol. Discussion of the obtained results and future work are presented in Section 4. Related work will be also briefly described in Section 5, with a conclusion in Section 6.

2. SENTENCE SELECTION AND VOICE AIDED INPUT

A phrase-book is a collection of ready-made sentences usually for a foreign language along with a translation and often in the form of questions and answers. Our aim is to embed a system of phrasebook with multimodality in smartphones. The sentences are organized according to a hierarchy of specific domains or situations such as *transportation, restaurant, shopping, etc.* The basic idea is to search and look for a sentence from that collection of sentences by simply pronouncing it.

2.1. Existing methods

Looking for a sentence may be done by the method which is commonly used in keyword spotting [1] which identifies keywords in utterance and in information retrieval in order to satisfy the users' query. The idea can be extended into our context by using a fuzzy matching search which will return the closest found sentences.

2.2. High level approach

A conventional automatic speech recognition system relies on two models: an HMM acoustic model (AM) in which each state is a Gaussian mixture and an *n-gram* language model (LM). The most commonly used acoustic unit for sound modeling is a phoneme and a LM most commonly used is an n-gram (for instance *3-gram)* of word units. Such a model requires a large amount of training data. Between the AM and a LM, a pronunciation dictionary is used to map a sequence of acoustic units into words present in the LM. At this high level, the hypothesis produced by the recognizer is a sequence of words. The speech decoding is costly in computing time and memory consumption, which is not necessarily adapted for a portable system with low resource constraints, especially if we want to handle multiple languages using the same device.

2.3. Low level approach

Contrarily to the above approach, by using only an AM and a flat LM which is only made up of a phone loop grammar, we can have hypothesis as a sequence of phonemes. With this solution, we clearly see the absence of LM which puts a burden on a computing time during the decoding process for a HL approach. It also avoids the problems of out-of-vocabulary (OOV) and vocabulary dependence which are faced by the word-level approach while it has been shown to maintain a good accuracy in certain cases [2].

3. EXPERIMENTATION PROTOCOL

The experimentation has been done on two different languages: French and Khmer. The aim is to measure the precision of correctly identified sentences for a set of utterances given a collection of sentences in a specific domain for both HL and LL approaches. Formally, for a collection of N known sentences, let T be the total number of sentences to be retrieved and H be the number of sentences effectively retrieved. Hence, the precision P is calculated by $P = H/T$.

In our experimentation for speech to text, we use Sphinx3.0.8 [4] ASR toolkit. The acoustic model is a 3-state HMM. The vector of parameters contains 13 MFCCs, as well as its first and second derivatives.

3.1. French language

The initial corpus was taken from a SurviTra CIFLI [3] project – a web service aimed at building resource and tools for survival language kit for French visitors to communicate with Indian helper when English is not an option. A sentence is either a completely fixed sentence or an instance of a sentence which has fixed parts and variable parts. We choose the domain of *restaurants* for our experimental work. The sentences are short, easy and simple; and they are useful for communication. Example: *Puis-je réserver demain pour [$C_twoToTwelve] personnes?* (Can I book a table for [$C_twoToTwelve] persons tomorrow). In this example, the variable *$C_twoToTwelve* can take an integer value between *2* and *12*. 329 complete test sentences were created and recorded at 16 kHz by using a normal headphone in an office-like environment by 12 native speakers – an average of 2 seconds per sentence, the longest is 4 seconds. The total of recorded sounds of all speakers is around 1h50. The sentences were then divided into two groups namely ***Fr.Test1*** & ***Fr.Test***2 to do different test scenarios. Fr.Test1 contains all the odd sentences (165 sentences), while Fr.Test2 contains all the even sentences (164 sentences); the sentences from the latter (Fr.Test2) will be included in the training set used to train LM for the HL approach. The overall phrase-book corpus contains 1064 sentences including the 329 recorded sentences; thus the other phrase-book sentences are here to bring confusion during the search process as depicted in Figure 1. The choice of this latter number of sentences is done to simulate the real phrasebook sentence retrieval.

Acoustic model: For all experiments, the AM was trained by SphinxTrain [4] using the French BREF120[1] which contains over 100h of 120 speakers of speech corpus. The model topology is 3 states left-to-right, 16 Gaussian mixtures. We use context-dependent acoustic models.

HL Approach - LM training: we used different data to train our 3-gram LMs from (a) in-domain and (b) close-to-domain vocabulary:

- In-domain: the 164 sentences of Fr.Test2 whose vocabulary size is 219 words (consequently. Fr.Test2 can be considered as a “cheating” experiment for which the sentences to be recognized are included in the training data[2])
- Interpolation with a close-to-domain vocabulary: two LMs are extracted from existing data collections: PVE [5] and Nespole![3], and then interpolated to create a background LM. It is then interpolated with another LM trained from Fr.Test2 to create the final LM of 3903 words.
- A pronunciation dictionary of more than 62K words; each pronunciation was either present in our source lexicon, or automatically generated using LIA_PHON [6] when necessary.

LL Approach – Flat LM: To get to the phoneme level a flat LM is used: it is simply made up of a list of all phonemes presented in the AM, all equiprobable (phone loop grammar).

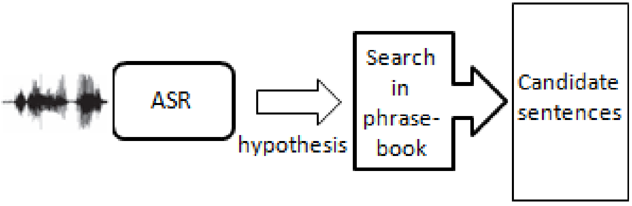


Figure 1. Overall process of voice input for phrase selection

[1] http://www.elda.fr/catalogue/en/speech/S0067.html
[2] This is however a realistic scenario in the case of phrase-books
[3] http://nespole.itc.it/

Multiple references: To increase the performance of sentence retrieval in the LL approach, we have taken the advantage of (a) similarities in pronunciations of French words and (b) the phoneme groups generated by recognition engine, which allow us to combine all pronunciation into a single confusion network. Table 1 shows an example of a first person plural future form of a verb *"bouger - bougerons"* (move) where we can have two similar pronunciations.

Table 1. *Example of two similar pronunciations of a first person plural future form of a verb "bouger"*

bougerons	b	u	ʒ	ə	R	O~
bougerons(2)	b	u	ʒ	-	R	O~

The two pronunciations are combined into a single confusion network with an introduction of an *epsilon* in a place of a dash. With this approach, we can lead to a pronunciation which does not exist in the original ones due to the presence of *epsilon*.

Phrase retrieval: From the output of the recognition engine the edit-distance (Levenshtein distance) between the hypothesis and references is calculated. The candidate sentences are then returned based on their shortest edit-distance. The search can be done at word level or at phoneme level. If a first sentence candidate matches the searched sentence, this will be considered as a hit. Otherwise we continue to select the next candidate that appears in a 2^{nd}, 3^{rd} positions so on and so forth. We do not optimize the runtime of this retrieval mechanism and its computational cost depends on the complexity of the references.

Experimental results: In our experiments, we conducted different test scenarios for two test data Fr.Test1 and Fr.Test2, but only three important results are shown here. The precision is calculated in a window of 6 first candidates; since only these results could be shown on the small-size screen of smartphones for the ergonomic reason for the future phrase-book application. In Figure 2, the first two HL approaches are used as our baselines for comparisons. The lowest curve represents a HL approach, in which we used an *in-domain* small-size LM (trained on the transcriptions of Fr.Test2 only). This shows that among 50% of the cases, a correct sentence appears at the first position of the retrieved candidates. The precision augments when the selection window increases. The curve above the lowest curve also represents a HL approach but with a larger-sized interpolated LM (described earlier). It shows a slightly better result. The top curve is a LL approach with flat LM in which multiple references for each sentence are used to calculate the edit-distance. They are made of similarities of phonemes and phoneme groups. It clearly suggests that this method shows the best results among all. It gains a significant precision over the two HL approaches. This can be explained by some missed recognized words in hypothesis for HL approach but not at the low level approach, the fuzzy matching (similarities) between phonemes proves to be more effective.
In the test data Fr.Test2 (Figure 3), it is not surprising to see that precisions of the HL approaches are significantly better than the ones of LL approaches, since LM contains the transcription Fr.Test2. But interestingly, the LL approach also yields a good precision. And they are well correlated to the results of the Fr.Test1 experiment, the precision increases when the list of candidates grows.

The experiment results are promising for the low-level approach with the usage of multiple references. We will discuss their performances at the end of this section.

3.2. Khmer language

We replicated the same experiment on the Khmer language as we did for French. Khmer is the official language of Cambodia and it is still an under-resourced language.
We started by collecting a test corpus. The sentences were obtained from invented sentences corresponding to restaurant situations. We got a corpus of 268 sentences. All sentences were then checked and verified by native speakers. These sentences were recorded with a normal headphone in the same condition as French by a total of 13 speakers. Each sentence has an average duration of 2 seconds and we obtained a total of 1h40 of recording signal for all speakers. The overall phrase-book corpus contains 1033 sentences including the 268 recorded sentences plus other sentences added to bring confusion during the search process and they have been contributed by a group of Cambodian speakers.

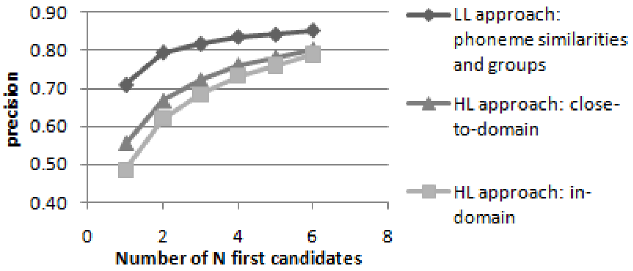


Figure 2. Precision of correctly retrieved phrases Fr.Test1

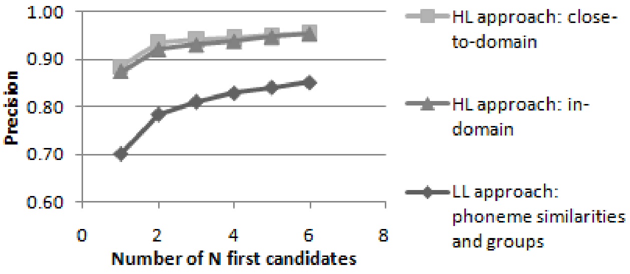


Figure 3. Precision of correctly retrieved phrases Fr.Test2

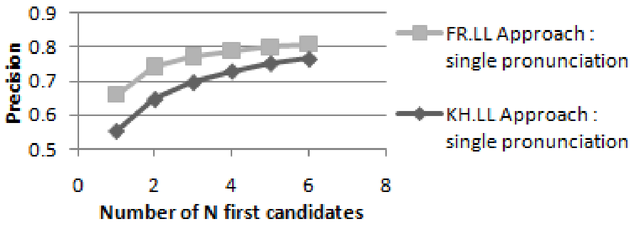


Figure 4. Precision of correctly retrieved phrases for Khmer and French for single pronunciation reference

Due to the limitation of Khmer language resources available for LM training, we were able to experiment only on the low-level (LL) approach, as far as a quick development of Khmer voice input is concerned. If the order of magnitude of the results is similar to the French LL case, we would conclude that our LL approach is useful in the case of low-resourced language where no LM data is available.

Acoustic Model: We use the same tools to train Khmer AM using a Khmer Broadcast News type containing 6h30 sounds of speech by 8 speakers (3 females) [7]. The topology is the same as that of the French model, except it has only 8 Gaussian mixtures per state. We use a context-dependent acoustic model. To train the model, we use a dictionary which is based on the SEALANG[4] project.

For the sentence references, it is not yet possible to use multiple references, due to the limitation of our pronunciation dictionary. The sentence retrieval process is done exactly as the one described above for French.

Experimental results: In Figure 4, we obtained a good precision for Khmer comparable to that of the LL approach for French, where the reference contains only a single pronunciation of each word[5]. The curve tends to increase as the list of candidate sentences is augmented. It is therefore interesting to notice that even a fairly small acoustic model can give rise to a fairly good precision despite the different nature of training speech data and the nature of test data.

3.3. HL vers. LL complexity and performance

Each experiment was executed 40 times consecutively on our Linux server Quad-core in normal load-balancing to measure their decoding time. In Table 2, it shows the execution time expressed in Real Time Ratio (RTR) of each approach for each language and the size of LM. For French, the decoding time of the HL large-sized LM approach is slightly above the half of the duration of the original signal in average. The 2nd column corresponds to the time required to decode each utterance for a small-sized LM, it is less compared to the large-sized approach; hence the size of LM matters. The last column is a LL approach runtime; it is also slightly better comparing to the two previous ones. Despite no clearly significant gain over decoding time among these approaches, the LL approach is still more advantageous in terms of memory consumption because no LM is required. For the Khmer language, the decoding time is very rapid, less than its signal duration. This is due to the small size of its acoustic model.

Table 2. *Decoding runtime in Real Time Ratio (RTR)*

	HL large-sized LM (423K) RTR	**HL small-sized LM (15K) RTR**	**LL RTR**
FR.Test1	0.58	0.47	0.45
FR.Test2	0.55	0.44	0.42
KH			0.06

4. DISCUSSIONS AND FUTURE WORKS

The results of experiments suggest that the LL approach is more efficient and fast enough. It shows a great potential that can be further adapted and optimized for smartphones. More tuning (beam width etc.) is needed in order to reduce the decoding time. The small-sized acoustic model for Khmer also proves to be very efficient in our context of under-resourced languages (and no language model data available) and the small-size *in-domain* language model can also be considered in terms of accuracy. We need also to consider the time required to retrieve sentences; this will affect the overall performance and it is a trade-off between the accuracy and performance.

For the future works, it is also interesting to consider other possibilities of using a single multilingual acoustic model which can be applied on several closely related languages to Khmer such as Thai, Laos and Vietnamese. If it yields similar or better results, it can lead to the use of a generic model that can cover a group of languages sharing acoustically similar characteristics. Another important factor to mention is the size of acoustic model which also has an impact on the performance of the system considering the future applications embedded on small hand-held devices.

5. RELATED WORK

In the literature, we can find some similar work which tried to address the problem of speech-to-speech translation on portable devices, which is different from our case – we focus on the realization of a multimodal multilingual phrasebook running on smartphones but not on speech translation. Among those systems, we can categorize them into two groups: (1) the systems using a portable device as a terminal, where the recognition is done on the server side: Med-SLT [8], Google Mobile App [9] and (2) systems with onboard speech recognition (MASTOR [10], Speechalator [11]).

[4] http://sealang.net/khmer/dictionary.htm

[5] We used a single pronunciation for French in order to perform a fair comparison between French and Khmer

6. CONCLUSION

We have presented in this paper a high and a low level approach used in ASR. The low level approach is promising despite more optimization and adaptation needed in order to embed this technology into our future phrasebook systems running on smartphones.

Acknowledgement
I would like to thank to all my collegues working in the Institute of Technology of Cambodia and GETALP-LIG, Grenoble for helping me to collect and verify the Khmer and French corpora.

7. REFERENCES

[1] J. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, and M. Siu, “Phonetic training and language modeling for word spotting,” in *Proceedings of Int. Conf. Acoustics, Speech, and Signal Processing*, 1993, vol. II, pp.59–462.

[2] F. Seide and P. Yu, “Vocabulary-independent search in spontaneous speech”. In *Proceedings of ICASSP’2004,* Montreal, Canada.

[3] C. Boitet, P. Bhattacharyya, E. Blanc, S. Meena, S. Boudhh, G. Fafiotte, A. Falaise, V. Vacchani, “Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction”. *Actes de SNLP 2007*, Pattaya, Thaïlande.

[4] http://cmusphinx.sourceforge.net/html/cmusphinx.php

[5] Y. Fouquet, “Le magicien d’Oz pour du dialogue oral: expérience avec un assistant virtuel en enterprise”. *Actes de RJCP,* Grenoble, septembre 2003.

[6] F. Bechet, “LIA_PHON - Un système complet de phonétisation de texts”, *revue T.A.L. vol. 42, num. 1/2001, édition Hermes.*

[7] S. Seng, S. Sam, L. Besacier, B. Bigi and E. Castelli, “First Broadcast News Transcription System for Khmer Language”, In *Proceedings. of the Sixth International Language Resources and Evaluation (LREC'08).*

[8] P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, and H. Isahara. 2005. “A generic multilingual open source platform for limited-domain medical speech translation”. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, Budapest, Hungary.

[9] http://www.google.com/mobile/

[10] Gao Y, Zhou B, Sarikaya R, Afify M, Kuo H, Zhu W, Deng Y, Prosser C, Zhang W and Besacier L, “IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator”, In *Proceedings of First International Workshop on Medical Speech Translation*, pp. 53-56, USA, 2006.

[11] A. Waibel, A. Badran, A. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Mayfield Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang, “Speechalator: two-way speech-to-speech translation on a consumer PDA,” in *Proceedings of the European Conference on Speech Communication and Technology*, 2003.

AUTONOMOUS ACOUSTIC MODEL ADAPTATION FOR MULTILINGUAL MEETING TRANSCRIPTION INVOLVING HIGH- AND LOW-RESOURCED LANGUAGES

Sethserey Sam[1,2], *Laurent Besacier*[1], *Eric Castelli*[2], *Bin Ma*[3], *Cheung-Chi Leung*[3], *Haizhou Li*[3]

[1]LIG Laboratory, UMR CNRS 5524 BP 53, 38041 Grenoble Cedex 9, France
[2]MICA research center, UMI CNRS 2954, HUT, Hanoi, Vietnam
[3]Institute for Infocomm Research, Singapore

ABSTRACT

In speech technology, we found several challenges in automatic speech transcription system for multilingual conferences or meetings. Firstly, the dialog occurs between native and non-native speakers. Secondly, the non-native speakers come from different parts of the world (e.g., English spoken by native French speakers or English spoken by native Vietnamese speakers, etc.). Thirdly, no data or a limited amount of data is available to bootstrap the acoustic modeling. This paper presents some autonomous online and offline acoustic model adaptation approaches, which required no additional data in the adaptation process, to deal with above challenges as well as to improve the performance of the phone recognizers used for automatic transcription purpose. Experiments show that our adaptation approach (online interpolation with MLLR based on PR-VSM) can provide about 4% absolute gain in Phone Accuracy Rate (PAR) compared to the multilingual baseline system and it is even better than the performance of the supervised monolingual systems.

Index Terms— ASR**,** multilingual acoustic modeling, language label voting, PR-VSM, MLLR.

1. INTRODUCTION

With maturing speech technology and the need of global communication, automatic transcription of multilingual conference speech is becoming a topic of interest. In multilingual meeting transcription, we find many interesting challenges: 1) How can we improve the system performance while both native and non-native utterances are involved? 2) Non-native speaker with different speaking styles and accents is another concern. For example, English spoken by French speakers is different from English spoken by Vietnamese speakers. According to [1], speakers borrow acoustic features from their native languages in their non-native speech. 3) It is difficult to find enough data with the same nature of multilingual meeting to bootstrap the acoustic modeling. So in this case, what kind of adaptation should we use to improve the system performance?

In this paper, we investigate on multilingual acoustic model (Mult-AM) adaptation for multilingual meeting transcription in which 3 languages are involved: English (EN), French (FR) and Vietnamese (VN). We focus our Mult-AM adaptation in an autonomous fashion. Here, “autonomous” means that the acoustic model is automatically readapted itself before the final decoding for an utterance or a group of utterances. Two reasons for using the autonomous adaptation process are: 1) no external data is available for adaptation; 2) the adaptation is made automatically during the decoding process based on what we call a language observer. The goal of the language observer is to assign a likelihood to each language candidate and to use this information during the adaptation process. To the best of our knowledge, the observer-based approach that we propose has not been studied yet for acoustic model adaptation in multilingual ASR. In the adaptation process, three online and two offline adaptation acoustic modeling approaches are studied. Online adaptation means that only the current utterance is available for the current adaptation process. On the other hand, offline adaptation can use both current utterance and all the data in the document history for the current utterance considered.

This paper is organized as follows. In Section 2 we present the multilingual meeting corpus setup from which we extract the test data. The baseline systems, the autonomous adaptation process, the language observer and the acoustic model adaptation are detailed in Section 3, 4, 5 and 6 respectively. In Section 7 we provide some experimental results and conclude in Section 8.

2. MULTILINGUAL MEETING CORPUS SETUP

We extract the test data from the "MICA meeting speech corpus" that was recorded at the meeting room of MICA[1] research center. This corpus contains around 3h30mn of

[1] www.mica.edu.vn

transcribed speech in 4 languages EN (English), FR (French), VN (Vietnamese) and KH (Khmer: Cambodia's language). This multilingual meeting corpus involves the speech from 9 speakers (3 French, 3 Vietnamese and 3 Cambodian). In the corpus dialog, we discover that each speaker can use their native or non-native languages to communicate according to whom they speak with. Table 1 presents the distribution of the languages spoken by speakers with different native languages.

	Lang-KH	**Lang-VN**	**Lang-FR**	**Lang-EN**
Spk-KH	570	452	1822	3452
Spk-VN	0	1147	577	255
Spk-FR	0	0	2797	1370

Table 1. Duration coverage matrix (in second) of languages spoken by different native speakers.

In Table 1, we notice that non-native speech represents 64% of total speech in the corpus. Moreover, a speaker generally keeps speaking a language unless another speaker starts a new language in the dialog. It means that, in MICA meeting speech corpus, language switching probably occurs when the active speaker in the dialog changes. So we segmented (manually) the speech data based on speaker turns so that each speech segment contains one language only (native or non-native but no code-switching).

In this paper, we extract only the native and non-native speech data of EN, FR and VN from MICA speech corpus and we select only the utterances longer than 3 seconds for our experiments. Table 2 presents the quantity of testing data that will be used in the experiments.

Language	**Native/Non-native speech**	**Test Data**
EN	EN_fr	715
	EN_vn	56
FR	FR_fr	251
	FR_vn	279
VN	VN_vn	219
	TOTAL	**1520**

Table 2. Quantity of testing data (value in seconds) used in our experiments.

Note that, in the content of Table 2, the term in capital letters denotes the language spoken by the speakers in the speech segments and the term in small letters denotes the dominant language of these speakers (for example, EN_fr means English spoken by native French speakers).

So finally, we have a test data set of around 26 minutes where the non-native speech represents 69% of total test speech.

3. BASELINE SYSTEM

All recognition experiments described in this paper use the Sphinx3 decoder [2]. Our baseline system is a multilingual acoustic-phonetic recognizer (Mult-PR) of the three languages (EN, FR, and VN). The multilingual acoustic modeling (Mult-AM) is created by combining the existing acoustic models of EN, FR and VN trained respectively on WSJ corpus [3], BREF120 corpus [4], and VNSpeechCorpus [5]. The combination of acoustic models is simply made based on the *ML-sep* combination method [6]. It means that there is no data to share across language among the three monolingual acoustic models. Moreover, our Mult-AM is a context independent acoustic model that contains 124 acoustic units: EN (40 phonemes), FR (43 phonemes) and VN (41 phonemes). Each acoustic unit is represented by a HMM of 3 states with 16 Gaussian components per state. In this article, we focus definitely on the acoustic model adaptation to improve the performance of acoustic-phonetic speech transcription system. So, for multilingual language modeling and lexical modeling we simply create respectively a flat LM of phones (phone loop grammar) and a phone list, for the 124 phonemes.

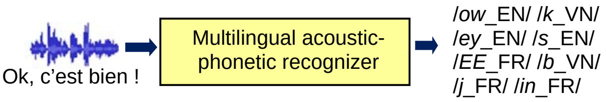


Fig. 1. An example of baseline system output.

In Fig.1, each phoneme, in the baseline system output, is presented in SAMPA format proposed by John Wells [10] and is appended with the label of language that the phoneme belongs to. With this output observation, we believe that we can attempt to identify the spoken language as well as the native language of the speaker in the test utterance. Our autonomous adaptation concept also starts from this initial observation.

4. AUTONOMOUS ADAPTATION PROCESS

In Fig.2, $U_1,\ldots,U_n$ are the speech utterances extracted from MICA multilingual meeting corpus mentioned in Section 2 (each utterance contains only one spoken language). The language observation module (language observer) provides the language information of each utterance by generating the likelihood of every language based on the first pass hypothesis of the utterance. Then the adaptation process is made by using the language information (language scores) generated by previous module (language observation). Finally, the second pass decoding uses the adapted acoustic model to decode the utterance.

So, the language observation is the key module in the autonomous acoustic model adaptation process.

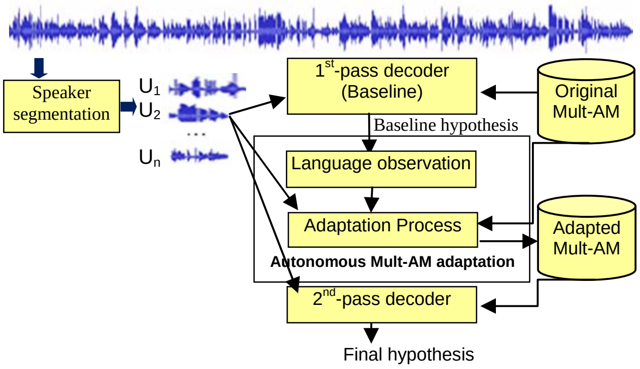


Fig. 2. Autonomous Mult-AM adaptation process.

5. LANGUAGE OBSERVATION

It is important to emphasize the fact that language observation is more than just a spoken language identification module. Similar to spoken language identification, language observation assigns a set of language likelihood scores for each test speech segment. All the language likelihood scores will be further considered during the acoustic model adaptation process. In our previous work [13], we believe that language observer gives not only the information about the spoken language in the speech segment but also the native language of the speaker. For example, if the language observer gives: *P(EN)* = 0.5, *P(FR)* = 0.4 and *P(VN)* = 0.1; then the speech segment may be in English spoken by a French speaker (or vice-versa).

We propose here a simple language observer approach called Language Label Voting (LLV) and compare its language classification performance with a phonotactic language recognition approach called Phone Recognizer followed by Vector Space Modeling (PR-VSM) [7][8].

5.1. Language Label Voting (LLV)

Our language label voting provides language likelihood scores for an utterance based on the estimate of phoneme sequence in the first-pass decoding (Fig.2) by the following formula:

$$P(L_i) = \frac{n(L_i)}{N} \qquad (1)$$

where *P* is the language likelihood score, L_i is one of the three languages (EN, FR or VN), *n* is the number of phonemes labeled by language L_i found in the phoneme sequence which is the result of the first-pass decoding and *N* is the total number of phonemes found in that phoneme sequence.

For example, if "*h*_EN *e*_EN *l*_FR *o*_EN" (“_EN” and “_FR” denote the language that the phoneme belong to) is the phoneme sequence estimated in the first-pass decoding when decoding a speech utterance "hello". With the formula (1), LLV produces the language classification as the following: *P(EN)*=3/4, *P(FR)*=1/4 and *P(VN)*=0.

5.2. Phone Recognizer following by Vector Space Modeling (PR-VSM)

As an alternative to our simple LLV approach, a phonotactic language recognition approach called Phone Recognizer followed by Vector Space Modeling (PR-VSM) [7][8] was also used. The multilingual acoustic-phonetic recognizer (baseline system) is used as the phone recognition frontend in the PR-VSM system. Each vector space model (VSM) representing a language is trained using 2 hours (per language) of the following corpora: WSJ (EN), BREF120 (FR) and VNSpeechCorpus (VN).

6. MULTI-LINGUAL ACOUSTIC MODEL (MULT-AM) ADAPTATION

We recall that the objective of the autonomous adaptation is to automatically readapt the acoustic model based on the language observer without using any external data. So, in the Mult-AM adaptation module, we study 3 online adaptation approaches and compare their performances with 2 offline adaptation approaches. We compared two online autonomous acoustic model adaptation approaches (INTER and INTER-MLLR), which are using the language observer not only as a language identification system, with a conventional online unsupervised acoustic model adaptation approach: MLLR (Section 6.1.1). Moreover, our objective to study autonomous online and offline adaptation is to compare the performance of the online adaptation based on language observer (INTER and INTER-MLLR) with the offline adaptation based on language identification (PM-MLLR and SLI-MLLR).

6.1. Online acoustic model adaptation

6.1.1. Maximum likelihood linear regression (MLLR)
MLLR is simple and known to be robust for unsupervised adaptation as well as effective for small amount of adaptation data [9]. The first pass hypothesis of the current decoding utterance is used to generate the transformation matrices so that it can produce a new adapted mean of the original acoustic modeling. In this paper all MLLR adaptations are based on a global mean-only transformation.

6.1.2. Acoustic model interpolation (INTER)
Because the acoustic model interpolation is useful for non-native ASR and non-native speech often exists in multilingual meeting data (see Section 2), we also investigate adaptation using non-native speaker cross-lingual acoustic model merging and interpolation (hybrid-

interpolation) [1], which is one of the most useful adaptation techniques in non-native ASR to readapt the acoustic model before decoding each speech segment. In this case, the acoustic model of the most likely language generated by language observation is considered as the target AM and the language likelihoods are used as the interpolation weights. Because we have two source models and only one target model, we propose to do the hybrid interpolation in two times successively where 2 acoustic models (the target AM and one of the source AMs) are interpolated at each time. Finally the adapted multilingual acoustic model is made by combining the two interpolated acoustic models based on the *LM-sep* combination method [6] as explained in Section 3.

In each hybrid interpolation process, when the Euclidean distance between a Gaussian in certain state of the target model (referred to as target Gaussian) and the associated Gaussian in certain state of the source model (referred to as source Gaussian), is below a threshold, their means, variances and mixture weights will be interpolated (Equation 2). Otherwise, merging is performed: for the source Gaussians that are far from their associated target Gaussians (Equation 3) or for those target Gaussians without any associated source Gaussian (Equation 4). In merging cases, their mixture weights will be reduced by the interpolation weight. The distance threshold can be calculated for example by measuring the average distance among the Gaussians, and then multiplying it with a constant. We finally formulate the hybrid interpolation of two acoustic models as follows:

$$g_{new,sn} = (1 - P(L_i)).g_{tg,sn} + P(L_i).g_{sc,sn},\ g_{sc,sn} \neq \varnothing,\quad d(g_{tg,sn}, g_{sc,sn}) \leq dist \qquad (2)$$

$$g_{new,sn} = g_{sc,sn},\ \omega_{new,sn} = P(L_i).\omega_{sc,sn},\ g_{sc,sn} \neq \varnothing,\quad d(g_{tg,sn}, g_{sc,sn}) > dist \qquad (3)$$

$$g_{new,sn} = g_{tg,sn},\ \omega_{new,sn} = (1 - P(L_i)).\omega_{tg,sn},\ g_{sc,sn} = \varnothing \qquad (4)$$

where $g_{new,sn}$ represents the interpolated/ merged Gaussian, $g_{tg,sn}$ is the target Gaussian, and $g_{sc,sn}$ is the source Gaussian. $P(L_i)$ is the interpolation weight (one of the source language likelihoods (Equation 1)), ω is the mixture weight for the Gaussian. *d(.)* is a distance function and *dist* is a threshold distance.

We also study the acoustic model interpolation followed by MLLR (INTER-MLLR) by simply applying the MLLR adaptation to the adapted mult-AM based on the above hybrid interpolation approach.

6.2. Offline acoustic model adaptation

6.2.1. Same language identification MLLR (SLI-MLLR)

The adaptation process is made in 3 successive steps: 1) for all utterances already decoded, we group the utterances according to the language identification tag provided by the language observation module (totally 3 groups: EN, FR and VN; 2) we use the speech segments and the first-pass decoding hypothesis of each language group for estimating its MLLR transform; 3) finally, we decode the current utterance by using the adapted Mult-AM.

6.2.2. Phone mapping MLLR (PM-MLLR)

The only difference between PM-MLLR and SLI-MLLR is that, PM-MLLR maps every phoneme from different languages in the estimated phone sequence to the similar phoneme of the most likely language identified by the language observation module. So we need to create 6 phoneme substitution tables to map between three languages (EN, FR and VN). Because the phoneme substitution result based on statistical phoneme confusion matrix [11] is from around 20% to 30% of wrong classification, we create the phoneme substitution tables based on the IPA chart and other studies [12]. For phonemes without their mapping in the IPA chart, phoneme confusion matrix results are used.

For example, if the first-pass decoding hypothesis of an utterance is "*h*_EN *e*_EN *l*_FR *o*_EN", the language information of that utterance provided by language observer is English, so the FR-to-EN phone substitution is called. It means that PM-MLLR maps the phone /*l*_FR/ to the similar phone in English /*l*_EN/. After the mapping process, PM-MLLR performs the same 3-step adaptation process as in SLI-MLLR.

7. EXPERIMENTAL RESULTS

7.1. Baseline multilingual system Vs. monolingual phone recognizers

Table 3 compares the Phone Accuracy Rate (PAR) between the baseline system and the other three monolingual phone recognizers (Mono-PR) applied on the right corresponding language. The comparison is made by using the testing data mentioned in Table 2 (assuming a perfect spoken language identification result is available in the case of monolingual phone recognizers). In all the experiments of this paper, the language label of the phone outputs is removed before evaluating the system performance. The language label of the phonemes is used only in the language observation module.

Native/Non-native speech	Baseline	Mono-PR
EN_fr	39.8	44.1
EN_vn	38.5	40.7
FR	41.8	48.7
FR_vn	40.8	44
VN	43.3	50.3
AVERAGE	**40.84**	**44.56**

Table 3. Baseline Vs. Mono-PR based on PAR [%].

It is important to recall that the test data contains 69% (as shown in Table 2) of non-native speech which explains why lower PARs are achieved in the monolingual phone recognizers as well as the baseline system. Moreover, the overall difference in PAR between Mult-PR (baseline system) and Mono-PR is not too big (< 4%).

7.2. Language observation: LLV versus PR-VSM

Native/Non-native speech	LLV	PR-VSM
EN_fr	**94.34**	89.68
EN_vn	**66.67**	61.9
FR	89.28	**96.43**
FR_vn	0	**17.39**
VN	**58.62**	50.28
AVERAGE	68.83	**69.18**

Table 4. LLV Vs. PR-VSM based on the spoken language identification accuracy [%].

As shown in Table 4, the performance of PR-VSM and LLV approaches are very comparable in term of spoken language identification. Moreover, English spoken by French speaker outperforms significantly English spoken by Vietnamese. On the other hand, the identification performance of French spoken by Vietnamese is poorer compared to other languages because the involved speakers have poor pronunciation knowledge of French language. So we could probably conclude that the identification of language is based not only on the spoken language but also on the speaker origin as well as on the speaker knowledge of the spoken language.

7.3. Acoustic model adaptation: Online Vs. Offline

Non-native/ native speech	Baseline	ONLINE			OFFLINE	
		MLLR	INTER	INTER-MLLR	SLI-MLLR	PM-MLLR
EN_fr	39.8	39.6	45.7	**45.7**	41.8	42.7
EN_vn	38.5	38.2	**43.9**	43.7	35.2	41.65
FR	41.8	43.1	40.7	41.3	43.4	**44.3**
FR_vn	40.8	**41.3**	38.3	39.6	39.5	41.2
VN	43.3	**43.5**	42.85	43.15	41.1	37.3
AVERAGE	40.84	41.2	44.22	**44.68**	41.96	42.38

Table 5.a. PAR [%] of various adaptation techniques (using PR-VSM observer)

Non-native/ native speech	Baseline	ONLINE			OFFLINE	
		MLLR	INTER	INTER-MLLR	SLI-MLLR	PM-MLLR
EN_fr	39.8	39.6	47.1	**47.6**	42.1	44.8
EN_vn	38.5	38.2	**46.3**	46.3	39.5	39.4
FR	41.8	43.1	40.6	41.3	44.8	**45.7**
FR_vn	40.8	41.3	44.95	**44.95**	43.15	43.7
VN	43.3	43.5	42.1	43.1	44.9	**46.3**
AVERAGE	40.84	41.2	45.76	**45.96**	44.2	44.94

Table 5.b. PAR [%] of various adaptation techniques (oracle case of language observer)

Table 5.a presents the system performance by using the PR-VSM approach as the language observation module.

On the other hand, Table 5.b presents the system performance by using perfect language identification (oracle case). In the oracle case, the interpolation is always made based on the language likelihood generated by the PR-VSM observer except that the target language is not the most likely language identified by the language observation module For example, if the utterance is English language but PR-VSM produces the language classification as *P(FR)=0.5*, *P(EN)=0.4* and *P(VN)=0.1*; in the oracle case, English is the target language (not French language) while the others are considered as source languages. Finally the interpolation is made by using the source language likelihood (*P(FR)* and *P(VN)*) as the source language weights in the acoustic model adaptation process (Equation 2).

With the adaptation performance presented in Table 5.a and 5.b, we can make the following comments:

- The system performance depends on the performance accuracy of the language observer. For instance, in Table 5.a, most of the adapted acoustic models for the non-native French spoken by Vietnamese degrade the system performance comparing to the baseline. But in the oracle case of language observation (Table 5.b), all the systems that use the adapted acoustic models for the non-native speech FR_vn outperform significantly the baseline system;
- Meanwhile, online interpolation-based adaptation improves significantly the system performance with non-native speech utterances but degrades the performance with native speech; this result shows, however, that the concept of autonomous adaptation, has a strong potential for decoding non-native speech of unpredictable origin.
- INTER-MLLR adaptation with PR-VSM based language observation (Table 5.a) provides better average performance than the monolingual system, in which perfect spoken language identification is considered on all utterances before decoding. This confirms that for non-native speech, making a hard decision on the language of the utterance, in order to choose the corresponding acoustic model, is not the best approach. Alternatively, the method we proposed suggests that soft decisions based on language observer outputs are useful for online multilingual acoustic model adaptation.

8. CONCLUSION

In this paper, we explored an autonomous approach for acoustic model adaptation in the context of meeting transcription. The advantage of this approach is that it automatically adapts the multilingual acoustic models without using any external data. Moreover, our adaptation approach (Interpolation followed by MLLR based on PR-VSM) can provide more than 4% absolute improvement of the PAR compared to the baseline system and it is even better than the performance obtained by supervised monolingual systems.

9. REFERENCES

[1] T.P. Tan, and L. Besacier, “Modeling context and language variation for non-native speech recognition", INTERSPEECH, 1429-1432, 2007.
[2] http://cmusphinx.sourceforge.net/html/cmusphinx.php
[3] W.M. Fisher, G.R. Doddington and K.M. Goudie-Marshall, “The DARPA Speech Recognition Research Database: Specifications and Status”, In Proc. of DARPA Workshop on Speech Recognition, pp. 93-99, 1986.
[4] L.F. Lamel, J.L. Gauvain and M. Eskénazi. BREF, “A Large Vocabulary Spoken Corpus for French”. In Proc. Eurospeech, pp. 505-508 1991.
[5] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, J-F.Serignat , “Spoken and written language resources for Vietnamese”, LREC, Lisbon, May 2004
V. B. Le, L. Besacier, “Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR”, ICSLP, 2006.
[6] T. Schultz, K. Kirchhoff, ”Multilingual Speech Processing”, Academic Press,2006.
[7] M. Campbell et al., “Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation,” in Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop, pp.1-8, 2006.
[8] H. Li, B. Ma, and C.H Lee, “A Vector Space Modeling Approach to Spoken Language Identification”, in IEEE Transactions on Audio, Speech and Language Processing, 2007.
[9] M. J. F. Gales, “Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition”, Technical Report, Cambridge University Engineering Department, May,1997.
[10] http://www.phon.ucl.ac.uk/home/sampa/
[11] J.J. Humpries and P. Woodland, “The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training”,*ICASSP-98*, Seattle, vol. 1, pp. 317-320, 1998.
[12] J. Flege, “The Production of ‘New’ and ‘Similar’ Phones in a Foreign Language: Evidence for the Effect of Equivalence Classification”, *Journal of Phonetics* vol. 15, pp. 47-65, 1987.
[13] S. Sam, “Vers des modèles autonomes pour la reconnaissance automatique de la parole multilingue ”, RJCP, Avignon, 2009.

MODELING OF GEMINATE DURATION IN AN AMHARIC TEXT-TO-SPEECH SYNTHESIS SYSTEM

Tadesse Anberbir[1]*, Tomio Takara*[2] *and Dong Yoon Kim*[1]

[1]Department of Computer Engineering, Ajou University, Suwon, Korea
[2] Department of Information Engineering, University of the Ryukyus, Okinawa, Japan
e-mail: tadanberbir@gmail.com

ABSTRACT

This paper presents analysis and modeling of geminate duration in Amharic Text-to-Speech (AmhTTS) synthesis system. AmhTTS is a parametric and rule-based system that employs a cepstral method. The system uses a source filter model for speech production and a Log Magnitude Approximation (LMA) filter as the vocal tract filter. Fundamental speech units of the system are syllables. Gemination in Amharic is one of the distinctive features of the language which plays a crucial role for the naturalness of synthesized speech sound. Therefore, in our study we mainly consider geminates and models the duration in AmhTTS system. The effectiveness of the durational model employed in our system was evaluated using 200 words (of which 40% of words containing one or more geminated syllables and 75% of the words containing sixth order syllables) and 5 sentences (with one or more words with geminated syllables) and we found promising results. The listening test results showed that accurate estimation of geminates duration is crucial for intelligibility and natural sounding of AmhTTS system. Our modeling greatly improved the intelligibility and naturalness of the system.

Index Terms— Amharic, geminates, speech synthesis, duration, cepstrum.

1. INTRODUCTION

Text-to-Speech (TTS) synthesis is a process which artificially produces synthetic speech for various applications. In TTS synthesis, naturalness is the main goal and it can be achieved mainly by incorporating prosodic features which include duration of segments, intonation patterns and stress. Particularly modeling the segmental duration based on the context is crucial. The goal of our research is also to predict and model the duration, mainly the duration of geminates, in Amharic Text-to-Speech synthesis system so as to improve the naturalness.

In TTS systems, accurate estimation of segmental duration is one of the most important factors that determine the naturalness of synthesized speech. So far, many researches have been conducted in this area and interesting results are obtained for various languages [1-3]. However, until now the task of duration modeling is still challenging mainly because the features considered for modeling duration are limited to those features that can be automatically derived from the input text only. Moreover, it is highly language dependent. For instance, in Amharic language geminates are unpredictable and cannot be driven from input text and this makes the automatic duration modeling in Amharic TTS system challenging. In our study we located the geminates by looking up Amharic-English dictionary which shows geminates by doubling Latin letters. For example, the word **ገና** is transcribed as /genna/ where the doubling shows the consonant /**ና**/ is geminated.

Amharic is the official language of Ethiopia. In Amharic language duration of geminates plays critical role for naturalness of synthetic speech. Unlike English language in which the rhythm of the speech is mainly characterized by stress (loudness), rhythm in Amharic is mainly marked by longer and shorter syllables depending on gemination of consonants, and by certain features of phrasing [4]. Gemination plays a key role in distinguishing words from one another, in the grammar of verbs and for proper pronunciation (naturalness) of speech.

However, so far, no research has been conducted on the acoustic of Amharic geminates. In general, Amharic is one of the least supported and least researched languages in the world. Although, recently, the development of different natural language processing (NLP) tools for analyzing Amharic text has begun, it is often very far comparing with other languages [5]. Particularly, researches conducted on the language technologies, such as speech synthesis is very limited. To our knowledge, so far there is only one published work [6], and one commercially available system [7] in the area of speech synthesis but no attempts have been made in analysis and modeling of prosody of Amharic in general and particularly in the area of modeling duration of geminates.

This paper reports the preliminary results of analysis and modeling of geminates in Amharic TTS system which is the first published work. The study is part of our ongoing work on prosody modeling (mainly on automatic prediction of geminates duration) for Amharic TTS synthesis system.

2. AMHARIC LANGUAGE'S OVERVIEW

Amharic (**አማርኛ**), the official language of Ethiopia is the Semitic language with the greatest number of speakers after Arabic and it has its own non-Latin based syllabic script called "Fidel" or "Abugida". The orthographic representation of the language is organized into orders (derivatives) as shown in Fig.1. Six of them are CV (C is a consonant, V is a vowel) combinations while the sixth orders are consonants only. In total there are 32 consonants, 7 vowels with 7x32= 224 syllables and 28 phonemes. The phonemes are reduced to 28 (see the APENDIX) because of the redundant graphemes that represent the same sound.

	1st order	2nd order	3rd order	4th order	5th order	6th order	7th order
Vow. Con.	ə	u	i	a	e	i	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

Figure 1: Amharic orthographic representation (7 orders)

Amharic has its own characterizing phonetic, phonological and morphological properties. Some of the striking features of Amharic phonology that gives the language its characteristic sound when one hears it spoken are the following: A weak indeterminate stress; the presence of glottalic, palatal, and labialized consonants; the frequent gemination of consonants; the frequency of central vowels; and the use of an automatic helping vowel [4].

Among these, in our study [8] we found geminates to be very critical for naturalness of synthesized speech. Gemination in Amharic is one of the most distinctive characteristics of the cadence of the speech, and also caries a very heavy semantic and syntactic functional weight [9]. The rhythm of speech in Amharic is mainly marked by longer and shorter syllables depending on gemination of consonants. Amharic is one of the typical examples of the world languages for having many consonants except consonant /h/ as geminate. A study of a survey of 45 languages with geminates also reported Amharic as one of the few languages having fricatives, liquids, and glides geminates which are less likely to occur as geminate in most languages [10]. In Amharic there are few comparative words which differ only by presence or absence of geminate consonants (see Table 1) and there are also many non-comparative words with one or more geminate consonants.

Table 1: Minimal pair of words with singleton vs. geminate consonants

Am	Eng	meaning	Am	Eng	meaning
ገና	/gena/	(still/yet)	ከፋ	/kefa/	(place name)
	/ge'na/	(christmas)		/ke'fa/	(worse)
ለጋ	/lega/	(fresh)	ሰፊ	/sefii/	(tailor)
	/le'ga/	(hit)		/se'fii/	(wide)
ዋና	/wana/	(swimming	ሽፍታ	/sxixfixta/	(rebel)
	/wa'na/	(main/core)		/sxix'fixta/	(rash)

Am: Amharic orthography **Eng**: English transcription

Gemination in Amharic is either lexical or morphological. As a lexical feature, it cannot be predicted. For instance, **ገና** may be read as /gena/ meaning 'still/yet', or /genna/ meaning 'christmas' as shown in Table-1. Native speakers easily perceive the difference of such words from context, but in speech synthesis it is very difficult to identify such words from the input text. As a morphological feature gemination is more predictable in the verb than in the noun [9]. However, the complex morphology of the languages makes the prediction very challenging. And so far no research has been conducted in this issue. In general, the lack of the orthography of Amharic to show geminates is the main problem in speech synthesis. In our study we employed a manual gemination insertion mechanism using apostrophe (') marks as shown in Table 1, and modeled the duration of geminates. In our system the gemination and other marks are defined externally and can be easily changed.

3. THRESHOLD DURATION OF CONSONANTS BETWEEN SINGLETON AND GEMINATE

It has been shown that the durational difference between singletons and geminates varies widely from language to language.

From phonological point of view, it has been shown that Amharic gemination (doubling of consonants) occurs when the consonants production takes longer time than the non-geminated ('single') consonants [4, 9, and 12]. However, the actual prolongation of geminates for different groups of consonants (stops, fricatives, nasals etc.) has not been determined and no research has been conducted from acoustical point of view. Therefore, in order to properly model duration of geminates, it is very important to study threshold duration between singletons and geminates of different consonant groups.

In this section, we discuss the results of an experiment performed to determine the threshold duration of consonants between the singletons and geminate consonants.

3.1. Stimuli

Using six pair of comparative words shown in Table 1, we performed two experiments to determine the threshold duration between singleton and geminates of voiced and unvoiced consonants. Three words with voiced stops /g/, /n/, and three with unvoiced fricative consonant /f/ were used. We considered only pair of words with continuant (voiced and unvoiced) consonants because we could not find comparative words with non-continuant (voiceless stops and glottallized) consonants.

For both experiments, we prepared 16 types of data for each word, among which, twelve were synthesized by repeating the parameters of the singletons in unstressed words as shown in Table 2. And two were analysis-synthesis words with singleton and geminate consonants and two original speech words were also added for comparison purpose. In the synthesized words, the parameters of the singleton consonant of unstressed word is repeated one frame per data. The conditions of repetition are shown in Table 2. The duration of each data from DATA(c) - DATA(n) has increased by an interval of 10ms. These are: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 [ms] plus the duration of singleton consonants in unstressed words.

Table 2: Duration settings of the files tested for determining the threshold duration

Stimulus	Duration Setting (ms)
DATA (c)	10
DATA (d)	20
DATA (e)	30
DATA (f)	40
DATA (g)	50
DATA (h)	60
DATA (i)	70
DATA (j)	80
DATA (k)	90
DATA (l)	100
DATA (m)	110
DATA (n)	120

3.2. Procedure of the listening test

Three native speakers, two male and one female ranged in age from 25-40, performed the listening test in a sound proof room using a head phone. All listeners have normal hearing ability and they are not aware of the difference between stress and geminates. Each sound was played once to each listener randomly in 2-s interval and the listener listened to the sound and select what he/she perceived among the list of three pairs of words using a listening test program. Each listener performed the listening test ten times and each word data was presented 3x10=30 times.

3.3. Results and discussion

Upon completion of experiments, we tried to determine the threshold duration between singletons and geminate consonants in unstressed and stressed words respectively. The graphs in Fig.2 and 3 show the average identification perception and response time for voiced and unvoiced consonants. Each point represents a mean of responses over the three listeners. The filled squares and diamond show the percentage of /Stressed/ or /unstressed/ responses to each of the twelve stimuli in the interval with duration. The filled triangles represent the corresponding latency of identification response to each stimulus.

Examination of both figures indicates that the identification is quite consistent. In both graphs listeners partitioned the stimulus into two groups (stressed or unstressed). The categorical boundary or crossover point in identification is about 45 ms for voiced words and 55 ms for unvoiced words. Inspection of the response time (RTs) for identification shows that listeners are slowest for stimuli 4 in the case of voiced words and slowest for stimuli 5 in the case of unvoiced words, and fastest for the other stimuli, which are within phonetic categories. We observed that the perception of durational is categorical between stressed and unstressed words. Listeners regularly perceived the different stimuli as being instances of either of the two words (stressed or unstressed). In both figures, we can see that the more the duration increase to the right, the more the words perceived as stressed. Fig.2 shows that the average threshold duration of consonants for voiced stops is 50ms, and Fig.3 shows that the average threshold duration of consonants for unvoiced fricatives is 70ms.

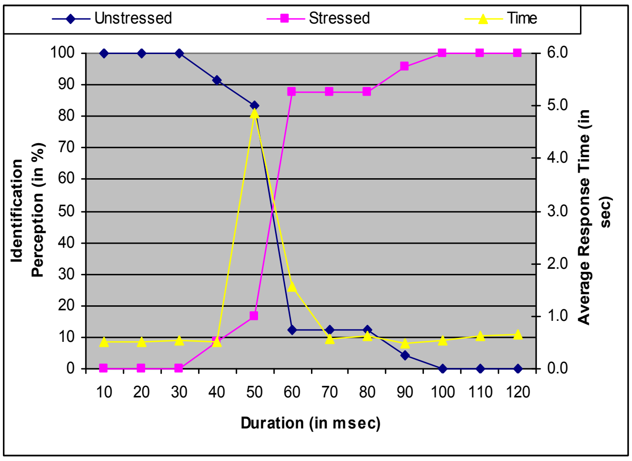


Figure 2: Average identification perception for the interval duration increase with average response time during identification for unvoiced consonants

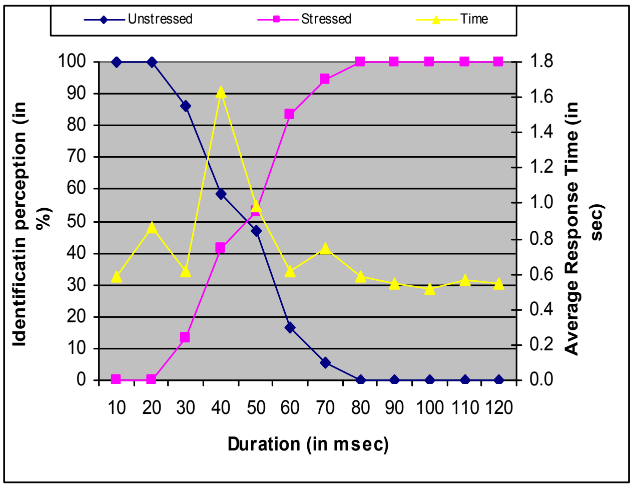


Figure 3: Average identification perception for the interval duration increase with average response time during identification for voiced consonants

4. AMHARIC TTS SYSTEM

In TTS systems, the process of converting written text into speech contains a number of steps. In general, TTS system contains two components: the Natural Language Processing (NLP) and the Digital Signal Processing (DSP). Similarly AmhTTS has two components. AmhTTS is a parametric and rule-based system that employs a Cepstral method and uses a Log Magnitude Approximation (LMA) filter. The system is designed based on the general speech synthesis system [11]. The input is Amharic text, and the output is synthetic speech. The text analysis subsystem converts Amharic text into a sequence of mapped characters, and then this sequence is used to get information for synthesis. The speech synthesis subsystem generates speech from pre-stored parameters under the control of systems rules. The database contains data for rules and syllable parameters with suitable formats. Fig. 4 shows the design of Amharic speech synthesis system.

4.1. Text Analysis

The text analysis subsystem extracts the linguistic and prosodic information from the input text. The program iterates through the input text and extracts the gemination and other marks (work interval and end marks) and then converts into a sequence of syllables using the syllabification rule. The letter-to-sound conversion has simple one-to-one mapping between orthography and phonetic transcription. As defined by Baye [12] and others, Amharic can be considered as a phonetic language with relatively simple relationship between orthography and phonology.

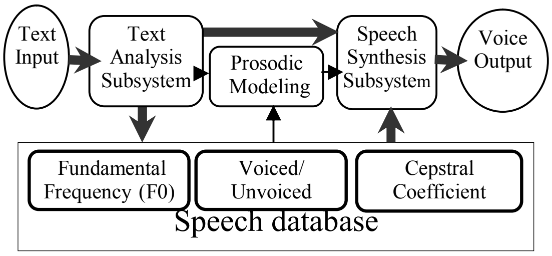


Figure 4. Amharic Speech Synthesis System

4.2. Speech Analysis and Synthesis systems

As a speech database, 196 Amharic syllables are collected and their sounds are prepared by recording on digital audio tape (DAT) at a 48 kHz sampling rate and 16-bit resolution. After that, they are down-sampled to 10 kHz for analyzing. All speech units are recorded in isolation with a natural reading. We used syllables as a basic unit because syllables are better in modeling the co-articulation effects than smaller units. Moreover the writing system is also syllabic.

Then, the recorded speech sounds were analyzed by the analysis system. The analysis system adopts short-time cepstral analysis with frame length 25.6 ms and frame shifting time of 10 ms. A time-domain Hamming window with a length of 25.6 ms is used in analysis. The cepstrum is defined as the inverse Fourier transform of the short-time logarithm amplitude spectrum. Cepstral analysis has the advantage that it could separate the spectral envelope part and the excitation part. The resulting parameters of speech unit include the number of frames and, for each frame, voiced/unvoiced (V/UV) decision, pitch period and cepstral coefficients c[m], $0 \leq m \leq 29$. The speech database contains these parameters as shown in fig.4.

Finally, the speech synthesis subsystem generates speech from pre-stored parameters under the control of the prosodic rules. For speech synthesis, the general source-filter model is used as a speech production model as shown in fig.5. The database contains data for rules and syllable parameters with suitable formats. Each syllable's parameters have a size of 2–6 KB. To make the system more generic, we use external definitions of interval marks and a character table code. The synthetic sound is produced using Log Magnitude Approximation (LMA) filter as the system filter, for which cepstral coefficients are used to characterize the speech sound.

The LMA filter presents the vocal tract characteristics that are estimated in 30 lower-order quefrency elements. The LMA filter is a pole-zero filter that is able to efficiently represent the vocal tract features for all speech sounds. The LMA filter is controlled by cepstrum parameters as vocal tract parameters, and it is driven by fundamental period impulse series for voiced sounds and by white noise for unvoiced sounds. The fundamental frequency (F0) of the speech is controlled by the impulse series of the fundamental period. The gain of the filter or the power of synthesized speech is set by the 0th order cepstral coefficient, c [0].

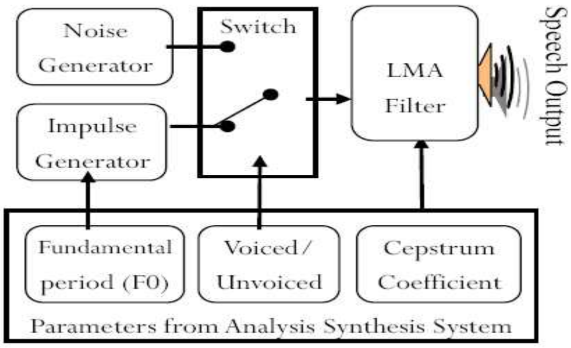


Figure. 5: Diagram of Speech Synthesis Model

4.3. Prosody Modeling

Prosody is a vital component in text-to-speech systems and preparation of appropriate prosodic control method is very important for the naturalness of synthesized speech of any language. In Amharic language segments duration is the most important and useful component in prosody control. It is shown that, unlike English language in which the rhythm of the speech is mainly characterized by stress (loudness), rhythm in Amharic is mainly marked by longer and shorter syllables depending on gemination of consonants, and by certain features of phrasing [4]. Therefore it is very important to model the geminates duration in AmhTTS system. In this paper we propose a new prosodic control method for synthesizing a high quality speech.

The task of the prosodic generation component of our system is to reliably predict and model the duration of geminates, sixth order syllables, and intonation contour. Our system uses a compact rule-based prosodic generation method in four phases:

- Syllables connection rules,
- Duration modeling,
- Sixth order syllables rules and,
- Intonation rule.

5. DURATION MODELING

The following section discusses the proposed duration modeling we employed in our system.

The duration modeling is programmed in the system and generates geminates from singletons by lengthening the duration of consonant part of the syllables following the gemination mark (’).

In general, we modeled the durations of:

- Syllable durations,
- Pause durations
  - between words,
  - between sentences and,
  - between intermediate phrases
- Geminates duration

5.1. Geminates duration

Two types of durational model were prepared for two groups of consonants, continuant (voiced and unvoiced) and non-continuant (stops and ejective/glottalized) consonants. If a gemination mark (’) is followed by syllable with voiced or unvoiced consonant, the last three frames of the cepstral parameters (c[0]) of vowel is adjusted linearly and then 120 ms of frame 1, 2 and 3 of second syllable is added. Then the second syllable is connected after frame 4. Totally 90 ms of cepstral parameters is added. Otherwise, if, a gemination mark (’) is followed by syllable with glottal or non-glottal consonant then, the last three frames of the cepstral parameters (c[0]) of vowel is adjusted linearly and then 100 ms of silence is added. Finally, and the second syllable is directly connected.

The following figures fig.7 and fig.9 show sample words synthesized by applying the durational model and, Fig.6 and Fig.8 show the waveform of original words just for comparison purpose only. The synthesized words are comparative words which differ only by presence or absence of gemination. In fig 6, the synthesized word /sxixfta/ ሽፍታ meaning ‘rebel’, the sixth order singleton consonant /f/ is unvoweled and short. However, in fig. 8, the word /sxix’fixta/ ሽፍታ meaning ‘rash’, the consonant /f/ is voweled and longer /’fix/ because it is geminated. Note that the sixth order syllables always need vowels to be pronounced as geminates.

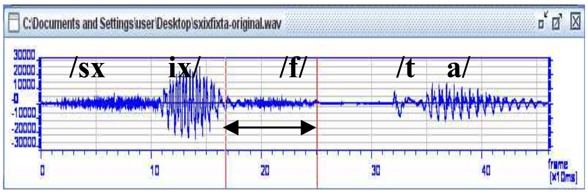


Figure 6: Waveform & duration of original word ሽፍታ/sxixfta/, meaning ‘rebel’

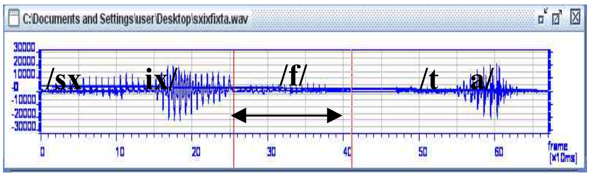


Figure 7: Waveform & duration of synthesized word ሽፍታ/sxixfta/, meaning ‘rebel’

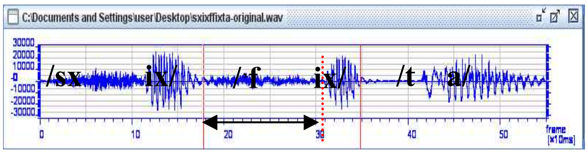


Figure 8: Waveform & duration of original word ሽፍታ /sxix’fixta/, meaning ‘rash’

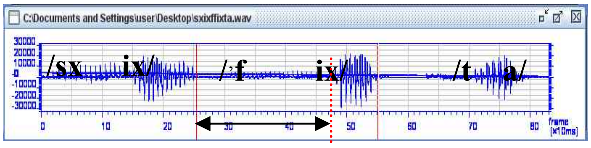


Figure 9: Waveform & duration of synthesized word ሽፍታ /sxix’fixta/, meaning ‘rash’

5.2. Silence Generation

The duration of the closure portion of a stop sounds plays an important role in the quality of synthesis. For stops, because units are picked from different contexts and partly because of faulty labeling of silence, concatenating them results in a duration of silence that may not be appropriate for the target context. More importantly, this duration is the one that distinguishes gemination and non-gemination (i.e., long and short consonants). Hence, inappropriate duration of silence will lead to stops sounding unnatural, or a geminate being perceived as a non-geminate (and vice versa).

For example, on synthesizing Amharic word /a’keke/, silence duration for the geminated syllable /ke/ (middle) was 59 msec and sounded like /akeke/, and not natural. However, on changing the silence duration to 100 msec it sounded natural. Fig. 10 show the stop duration in /a’keke/ where the first syllable /’ke/ is with geminate consonant/k/ and the last syllable /ke/ is with singleton consonant /k/.

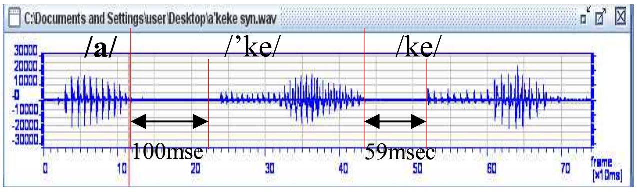


Figure 10: Synthesized word አከከ /a’keke/, meaning ‘scratch’ showing significance of silence duration for stop consonants

In our system, the duration rule generates the appropriate silence. As discussed above, when a gemination mark (’) is followed by syllable with stop consonants the last three frames of the cepstral parameters (c[0]) of vowel is adjusted linearly and then 100 ms of silence is added and the final syllable will be connected directly. Fig.7 shows a sample work synthesized by applying silence generation rule.

6. EVALUATION AND DISCUSSION

The effectiveness of the duration modeling employed in our system was evaluated using word and sentence listening tests. The word listening test was performed to evaluate the intelligibility and the sentence listening test was used to evaluate the naturalness of the system.

6.1. Recordings

The recording was done in a soundproof room, with a digital audio tape (DAT) recorder with sampling rate of of 48 kHz. Then from DAT the recorded data were transferred to a PC via a digital audio interface (A/D, D/A) converter. All recording was done by male native speaker who is not included in the listening tests.

6.2. Speech Materials

The stimuli for the first listening test consisted of 200 words which were selected from Amharic-English dictionary. The selected words are commonly and frequently used words in the day-to-day activities. Among the 200 words we selected, 80 words (40% of words) contain one or more geminated syllables and 75% of the words contain sixth order syllables. Using these words, two types of synthesized speech data were prepared: Analysis/synthesis sounds and rule-based synthesized sounds using AmhTTS system. The original speech sounds were also added in the test for comparison purpose.

For the second listening test we selected five sentences which contain words with either geminated syllables or sixth order syllables or both from Amharic grammar book [12]. Then, we prepared three kinds of speech data: original sentences, analysis/synthesis sentences, and synthesized sentences by our system by applying prosodic rules. In total we prepared 15 sounds.

6.3. Methods

Both listening tests were conducted by four Ethiopian adults who are native speakers of the language (2 female and 2 male). For both listening tests we prepared listening test programs and a brief introduction was given before the listening test.

In the first listening test, each sound was played once in 4 second interval and the listeners wrote the corresponding Amharic scripts to the word they heard on the given answer sheet.

In the second listening test, for each listener, we played all 15 sentences together and randomly. Each subject listens to 15 sentences and gives his/her judgment score using the listening test program by giving a measure of quality as follows: (5 – Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 – Bad). They evaluated the system by considering the naturalness aspect. Each listener did the listening test fifteen times and we took the last ten results considering the first five tests as training.

6.4. Results and Discussion

After collecting all listeners’ response, we calculated the average values and we found the following results.

In the first listening test, the average correct-rate for original and analysis-synthesis sounds were 100% and that of rule-based synthesized sounds was 98%. We found the synthesized words to be very intelligible.

In the second listening test the average mean opinion score (MOS) for synthesized sentences were 3.2 and that of original and analysis/synthesis sentences were 5.0 and 4.7 respectively. The result showed that the durational control method employed in our system is effective and produced fairly good prosody. However, the durational modeling only may not be enough to properly generate natural sound. Appropriate syllable connections rules and proper intonation modeling are also important. Therefore studying typical intonation contour by modeling word level prosody and improving syllables connection rules by using quality speech units is necessary for synthesizing high quality speech.

We also synthesized a paragraph shown in fig.11 which is taken from study [7] and asked the listeners to compare it with “Eruxelf Amatets”[1] speech synthesizer [7] which is a commercial synthesis system. All the listeners preferred the speech synthesized by our system and clearly able to understand what it says. However, the speech synthesized by Eruxelf Amatets2 is not intelligible as our system and lacks naturalness. Especially the prosody of Eruxelf Amatets is worse and our system is by far better.

ምአራፍ 6 ነጻነት

**በዛብህ፡ወላጆቹን፡ከድቶ፡ማንኩሳን፡አንደለቀቀ፡ዋሸራ፡ወደም
ትባል፡አገር፡ሄደ፡ ቅኔ፡ቤት፡ገባ።በዋሸራም፡አንድ፡ዓመት
እንደቆየ፡ቅኔ፡ተቀኝቶ፡ዝማሬ፡መዋሲት፡ለማካሄድ፡ወደ፡ዙር፡
አምባ፡ሄደ።ዙር፡አምባ፡ሁለት፡ዓመት፡ያክል፡ቆይቶ፡ዝማሬ፡
መዋሲትና፡ጽህፈት፡አወቀ።በመጨረሻ፡ደብረወርቅ፡አምትባ
ል፡አገር፡አዲስ፡የመጽሃፍ፡መምህር፡ከጎንደር፡መምጣታችው
ን፡ስለስማ፡ወደዚያ፡ሄዶ፡መጽሓፍ፡ለመቀጸል፡ወሰነ።**

Figure 11: Paragraph taken from [7]

7. ISSUES TO BE ADDRESSED

Segmented units occurring just before geminates will be left with some co-articulation due to the effect of geminate. Therefore, there is a need to handle such co-articulation related issues in syllable based synthesis.

8. CONCLUSION

In this paper, we presented a preliminary result on the durational modeling of geminates for AmhTTS system. We demonstrate how intelligibility of words can be improved by correct duration modeling of geminates. Our durational modeling of geminates greatly improved the quality of synthesized speech. However, the system still lacks naturalness and for better durational modeling, it needs automatic gemination assignment and modeling mechanisms.

Therefore, as a future work, we are planning to improve the duration model using the data obtained from the annotated speech corpus, properly model the co-articulation effect of geminates and to study the typical intonation contour. We are also planning to integrate a morphological analyzer for automatic gemination assignment and machine learning techniques for generating appropriate prosodic parameters.

7. REFERENCES

[1] Dennis H. Klatt, “Synthesis by rule of segmental durations in English sentences”, In B. Lindblom and S. Ohman, Editors, Frontiers of Speech Communication Research, pages 287–300, Academic Press, New York, 1979.

[2] Mixdorff, H., Nguyen, D.T. and Wu, N.T. (2005): Duration Modeling in a Vietnamese Text-to-Speech System. Proceedings of Specom2005, Patras, Greece, 2005.

[3] Alexandros Lazaridis, Panagiotis Zervas, Nikos Fakotakis, George Kokkinakis, "A CART approach for Duration Modeling of Greek Phonemes", In Proceedings of the 12th International Conference "Speech and Computer", Moscow, Russia, October, pp. 287-292, 2007.

[4] M.L Bender, J.D.Bowen, R.L. Cooper and C.A. Ferguson. Language in Ethiopia, London, Oxford University Press 1976.

[5] A. Alemu, L. Asker, and M. Getachew, “Natural language processing for Amharic: Overview and suggestions for a way forward,” in 10th Conf. Traitement Automatique des Langues Naturelles, Batz-sur-Mer, France, 2003.

[6] Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, Unit Selection Voice for Amharic Using Festvox, 5th ISCA Speech Synthesis Workshop, Pittsburgh, 2004.

[7] http://www.eruxelf.com.et/praoverv.htm

[8] Tadesse Anberbir and Tomio Takara., Amharic Speech Synthesis Using Cepstral Method with Stress Generation Rule, INTERSPEECH 2006 ICSLP, Pittsburgh, pp. 1340-1343, 2006.

[9] M. Lionel Bender, Hailu Fulass. 1978. Amharic Verb Morphology: A Generative Approach, Carbondale.

[10] Maddieson I., “Glides and gemination”, Lingua, (2007), doi:10.1016/j.lingua.2007.10.005

[11] Takara, T.; Kochi, T, General Speech Synthesis System for Japanese Ryukyu Dialect. Proc. of the 7th WestPRAC, 173-176, 2000.

[12] Baye Yimam, **የአማርኛ ሰዋሰው** (“Amharic Grammer”), Addis Ababa. (in Amaharic), 2008.

[1] Eruxelf Amatets is an Amharic speech synthesis program desgined based on the Festival Speech Synthesis System. It uses the Ked voice available in Festival to read out Amharic documents.

Appendix 1: Amharic consonants with their features shown using IPA, transcription we used and script in Amharic.

Manner of Articulation		**Place of Articulation**											
		Labials		**Alveolar**		**Palatals**		**Velars**		**Labio-Velar**		**Glottal**	
Stops	**Voiceless**	P [p]	ፕ	t [t]	ት			k [k]	ክ	kx	ኳ	ax[?]	እ
	Voiced	b [b]	ብ	d [d]	ድ			g [g]	ግ	gx	ጓ		
	Glottalized	p [p’]	ጵ	tx [t’]	ጥ			q [q]	ቅ	qx	ቋ		
Fricatives	**Voiceless**	f [f]	ፍ	S [s]	ስ	sx [ʃ]	ሽ					h [h]	ህ
	Voiced	V [v]	ቭ	z [z]	ዝ	zx[z’]	ዥ						
	Glottalized			xx [s’]	ጽ							hx	ኋ
Africatives	**Voiceless**					c [tʃ]	ች						
	Voiced					j [g’]	ጅ						
	Glottalized					cx c’]	ጭ						
Nasals	**Voiced**	m [m]	ም	n [n]	ን	nx [n’]	ኝ						
Liquids	**Voiced**			l [l]	ል								
				r [r]	ር								
Glides		w [w]	ው			y [j]	ይ						

UNSUPERVISED SMT FOR A LOW-RESOURCED LANGUAGE PAIR

Thi-Ngoc-Diep Do [1,2], *Laurent Besacier*[1], *Eric Castelli*[2]

[1]LIG Laboratory, CNRS/UMR-5217, Grenoble, France
[2]MICA Center – HUT, Hanoi, Vietnam – CNRS/UMI2954, Grenoble INP
ngoc-diep.do@mica.edu.vn

ABSTRACT

This paper presents an unsupervised method in application of extracting parallel sentence pairs from a comparable corpus. A translation system is used to mine the comparable corpus and to withdraw the parallel sentence pairs. An iteration process is implemented not only to increase the number of extracted parallel sentence pairs but also to improve the quality of translation system. A comparison between this unsupervised method and a semi-supervised method is also presented. The unsupervised extracting method was tested in a hard condition: the parallel corpus did not exist and the comparable corpus contained up to 50% of non parallel sentence pairs. However, the result shows that the unsupervised method can be really applied in the case of lacking parallel data.

Index Terms— *unsupervised method, extract parallel sentence pairs, comparable corpus.*

1. INTRODUCTION

Over the past fifty years of development [1], machine translation (MT) has obtained good results when applied to several pairs of languages such as English-French, English-Italia, etc. Many approaches for MT have been proposed, such as: rule-based (direct translation, interlingua-based, transfer-based), corpus-based (statistical, example-based) as well as hybrid approaches. However, research on SMT for low-resourced languages always faces the challenge of getting enough data to support any particular approach.

Statistical machine translation tries to generate translations using statistical methods based on large parallel bilingual corpora for source and target languages. These corpora are used to build a statistical translation model for source/target languages and a statistical language model for target language. The two models and a search module are then used to decode the best translation [2], [3]. Thus, a large parallel bilingual text corpus is a prerequisite. Such a corpus is not always available, especially for low-resourced languages.

The most common methods to build parallel corpora consist in automatic methods which collect parallel sentence pairs from the Web [4], [5], or alignment methods which extract parallel documents/sentences from two monolingual corpora [6], [7], [8]. Beside these "traditional" methods, there is also the method of extracting parallel sentence pairs from a comparable corpus. For instance, Sadaf and Schwenk present a semi-supervised extracting method [9]. This kind of method requires an initial parallel corpus (see more in section 2.1). We assume that in the case of a low-resourced language pair, even a small parallel corpus might not be available to start developing a SMT system. So, does a fully unsupervised method, starting with a noisy comparable corpus, is able to solve the problem of lacking parallel data?

This paper presents a fully unsupervised extracting method, in comparison with a semi-supervised extracting method. The first results show that the unsupervised method can be really applied in the case of lacking parallel data. The rest of the paper is organized as follows. Section 2 describes the two methods of extracting parallel sentence pairs from a comparable corpus: semi-supervised method versus fully unsupervised method. Section 3 gives our experiments and our results on testing the unsupervised method. The next section presents an application of this method for a real low-resourced language pair: Vietnamese-French. The last section concludes and gives some perspectives.

2. SEMI-SUPERVISED V/S UNSUPERVISED LEARNING

2.1 Semi-supervised learning method

Using a comparable corpus to extract parallel data has been presented in some previous works. D.S. Munteanu and D. Marcu present a method for extracting parallel sub-sentential fragments from comparable bilingual corpora [10].

Each source language document is translated into target language, using a bilingual lexicon/dictionary. The target language document which matches this translation is extracted from a collection of target language documents. Parallel sentence pairs are then filtered and parallel sub-sentential fragments are extracted from this document pair (see more in [10]).

S. Abdul-Rauf and H. Schwenk also present a method for extracting parallel data from a comparable corpus. To mine a comparable French-English corpus, for example, a statistical machine translation system is used to translate the French side to English. These translated texts are then compared with the English side, using the evaluation metric TER, and the parallel sentence pairs are filtered out. A post-processing is then applied to smooth the results. This technique is similar to that of [10], but a proper statistical machine translation system is used instead of the bilingual dictionary, and an evaluation metric is used to decide the degree of parallelism between two sentences.

All these methods are presented as effective methods to extracting parallel fragments/sentences from a comparable corpus.

2.2. Unsupervised learning method

The two above mentioned methods can be considered as semi-supervised methods, which need an initial parallel corpus to build the extracting system. We assume that in the case of low-resourced languages, this parallel corpus, even small, may be not available. So, we try to propose a fully unsupervised method, here, where the starting point is a simple noisy comparable corpus containing a significant amount of non parallel sentences. One of the challenges of this work is to see if such a different starting point (noisy comparable corpus, versus truly parallel corpus) can lead to the design of an acceptable SMT system.

Firstly, a baseline statistical machine translation system S_0 is built based on a comparable corpus (C_2) (in semi-supervised method, the system S_0 is built from a parallel corpus (C_1)). Of course the quality of S_0 is not high. We propose to use this system to mine another comparable corpus (D), and also to improve the quality of the translation system.

Secondly, the source side of the corpus D (in our case the source language is French) is translated by the system S_0. The translated output is then compared with the target side (English in our case) of the corpus D. The evaluation metric is calculated for each sentence pairs. The pairs are considered as parallel sentence pairs if the evaluation metric is larger than a threshold.

In our research, several evaluation metrics are used to determine which one is the most suitable. The scores are estimated at sentence level. Four common evaluation metrics are used: BLEU [11], NIST [12], TER [13] and a modified PER* (see in section 3.3).

The extracted sentence pairs are then combined with the baseline system S_0 in several ways to create a new translation system. An iteration process is performed which re-translates the French side by this new translation system, re-calculates the evaluation metric and then re-filters the parallel sentence pairs. We hope that each iteration not only increases the number of extracted parallel sentence pairs but also improves the quality of the translation system.

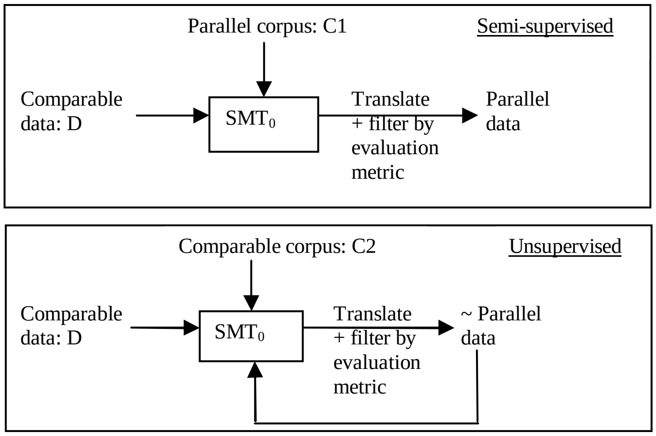


Figure 1: Semi-supervised v/s unsupervised methods

Again, to reuse the extracted parallel data in translation system, different combinations can be proposed:

- W1: The translation system at step i is retrained on a training corpus consisting of C2 and E_{i-1} (the extracted data from the last iteration); E_0 being the data extracted when translation system is trained on C2 only (S_0).
- W2: The translation system at step i is retrained on training corpus consisting of C2 and $E_0+E_1+\ldots+E_{i-1}$ (the extracted data from the previous iterations).
- W3: At iteration i, a new separate phrase-table is built based on the extracted data E_{i-1}. The translation system decodes using both phrase-table of S_0 and this new one (log-linear model) without weighting them.
- W4: The same combination as W3, but the phrase-table S0 and the new one are weighted, e.g. 1:2.

The section 3 presents our experiments on this unsupervised method.

3. PRELIMINARY EXPERIMENTS FOR FRENCH-ENGLISH SMT

In this section, we present experiments on unsupervised method, in comparison with those on semi-supervised method. Two systems were built, one based on semi-supervised method (Sys1), another based on unsupervised method (Sys2).

3.1. Data preparation

We chose French-English languages for these preliminary experiments. Data was chosen from the Europarl corpus

[6], version 3. The *correct parallel sentence pairs* were extracted directly from the Europarl corpus and a comparable corpus was simulated by introducing a significant amount of wrong sentence pairs in the data (about 50%).

To make it comparable with the real case treated in section 4 (low-resourced language pair), the size of the experimental data was chosen small. The corpus C1 contains only 50K correct parallel sentence pairs. The corpus C2 contains 25K correct parallel sentence pairs (withdrawn from C1) and 25K wrong sentence pairs. The corpus D, the input data for extracting process, was built from 10K correct parallel sentence pairs and 10K wrong sentence pairs, which were different from sentence pairs of C1 and C2. The correct and the wrong sentence pairs were marked to calculate the precision and the recall later.

3.2. System construction

Both systems Sys1 and Sys2 were constructed using the Moses toolkit [14]. This toolkit contains all of components needed to train the translation model. It also contains tools for tuning these models using minimum error rate training and for evaluating the translation result using the BLEU score.

The English language model was built from English part of the entire Europarl corpus. The baseline translation models were built from corpus C1 and C2.

3.3. Starting with parallel or comparable corpus?

One question that we want to answer first is whether the translation system based on a comparable corpus can be used to filter the input data like the translation system based on parallel corpus does. To examine this problem, the French side of corpus D was translated by Sys1 and Sys2. Then, the translated outputs were compared with the English side of the corpus D. Four evaluation scores were used in this comparison: BLEU, NIST, TER and PER*. Our modified position-independent word error rate (PER*) is calculated based on the similarity, while the PER [15] measures the difference, of words occurring in hypotheses and reference.

$$\text{PER*} = \frac{2 * \text{number of identical words}}{(\text{length of hypothesis} + \text{length of reference})}$$

Then the distributions of evaluation scores for correct parallel sentence pairs and wrong sentence pairs were calculated and presented in figure 2.

From these distributions, we can make the following comments: first, the distributions of scores have the same shape between Sys1 and Sys2. Especially, the distributions of scores for the wrong pairs were nearly identical in both systems. So, a comparable corpus can replace a parallel corpus for constructing an initial translation system. Remember that the initial comparable corpus here contains up to 50% non-parallel sentence pairs. Therefore, this kind of unsupervised method can be really applied in the case of lacking parallel data. Another important result is that the PER*, a simple and easily calculated score, can be considered as the best score to filter the correct parallel sentence pairs and filter out the wrong ones. Table 1 presents the precision and recall of filtering parallel sentence pairs from two systems.

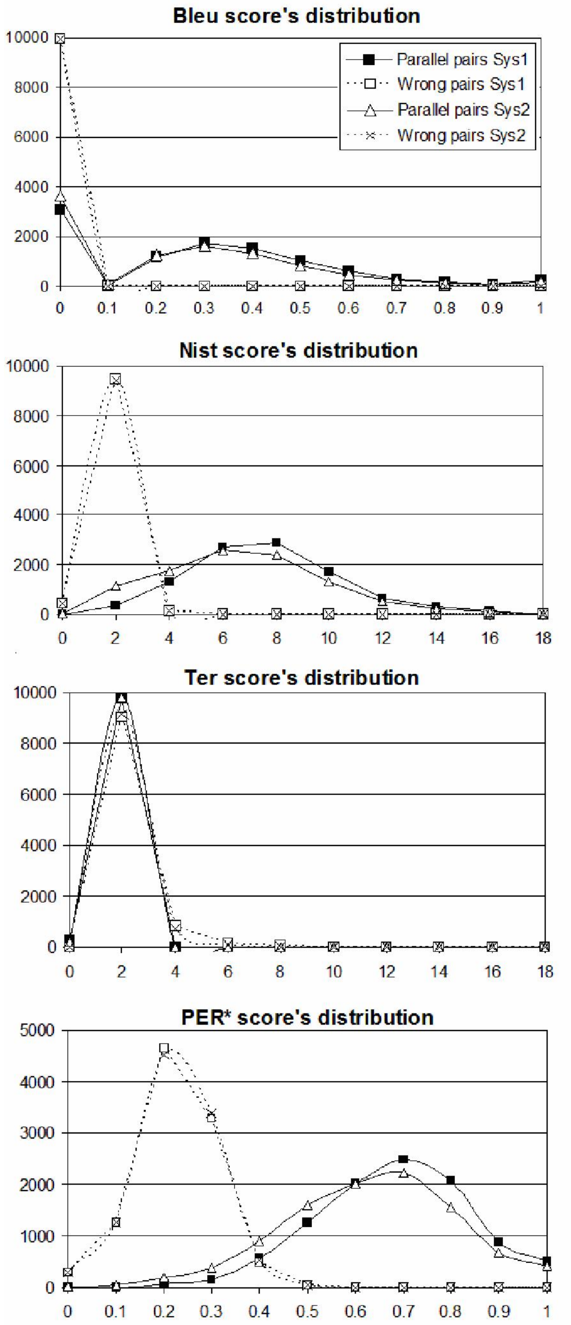


Figure 2: Score distributions for Sys1, Sys2

Sys1 – semi-supervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	6908	6892	99.76	68.92	81.52
Nist=0.4	8350	8347	99.96	83.47	90.97
Per=0.3*	10342	9785	94.61	97.85	96.20
Per=0.4*	9390	9333	99.39	93.33	96.27
Sys2 – unsupervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	6233	6218	99.75	62.18	76.61
Nist=0.4	7110	7108	99.97	71.08	83.08
Per=0.3*	10110	9468	93.65	94.68	94.16
Per=0.4*	8682	8629	99.38	86.29	92.37

Table 1: Precision and recall of filtering parallel sentence pairs (given 10K correct pairs)

3.4. The iterations of the unsupervised method

Section 3.3 has shown that an unsupervised method can be also used to filter the parallel sentence pairs from a comparable corpus. However the result of filtering in Sys2 is lower than that in Sys1 (for example, the number of correct extracted sentence pairs is reduced (table1)). So, we propose, in this section, an iterative process in order to improve the quality of the translation system, and then to increase the number of correctly extracted sentence pairs.

3.4.1. The number of correct extracted sentence pairs

The extracted sentence pairs were combined with the baseline system in four ways (as mentioned in section 2.2). The iteration experiment was carried out with Sys2. In order to receive the maximum number of correct extracted sentence pairs, for all iterations we chose the evaluation score PER* and the threshold=0.3, which gave the maximum recall=94.68% in the baseline system.

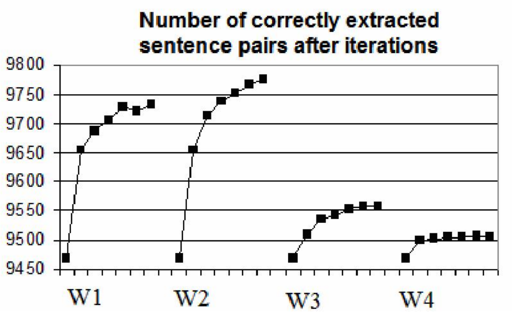


Figure 3: Number of correctly extracted sentence pairs after 6 iterations for four different combinations

Figure 3 presents the number of correctly extracted sentence pairs after 6 iterations for four different combinations: W1, W2, W3 and W4. The number of correct extracted pairs was increased in all cases; however the combination W2 brought the largest number of correct extracted sentence pairs.

3.4.2. The precision and the recall of filtering process

The precision and the recall of these four combinations are presented in figure 4. Because the filtering process focused on extracting the largest number of correct extracted sentence pairs, the precision was decreased. However, using the combination W2, the recall after 6 iterations (97.77) nearly reached the recall of Sys1 (97.85) (PER*=0.3).

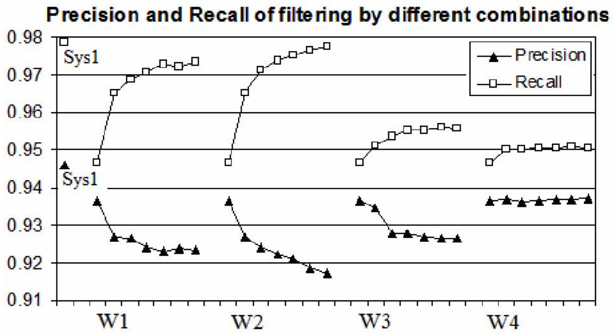


Figure 4: Precision and recall of filtering using different combinations

3.4.3. Translation system evaluation

The quality of the translation systems was also evaluated. A test set containing 400 French-English parallel sentence pairs was extracted from Europarl corpus. Each French sentence had only one English reference. The quality was reported in BLEU and TER. Figure 5 gives the evaluation scores for the systems after each iteration.

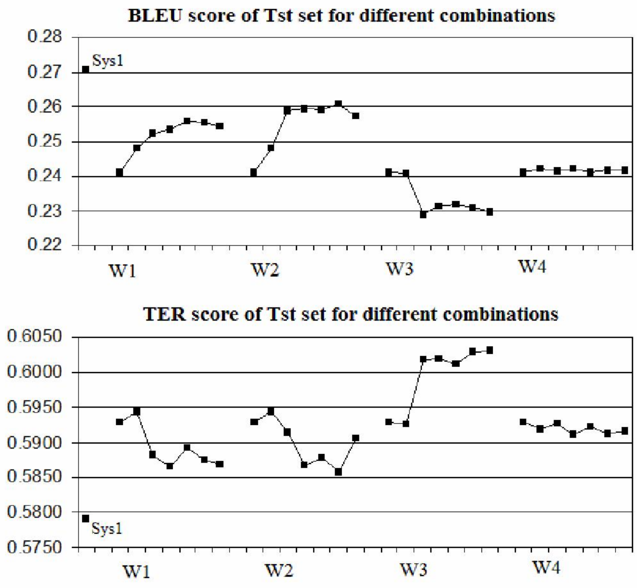


Figure 5: Translation system evaluations

The translation system evaluation revealed an important result. The quality of the translation system can increase quickly during some first iterations, then increase slowly and then it can be decreased after several iterations. It can be explained that for the first iterations, the new parallel sentence pairs are included into the translation model, so it increase the translation quality. However, for the next iterations, the precision of the extracting process was decreased, more wrong sentence pairs were added to the system, so the translation model got worse and the quality of translation system was reduced.

After about 3 iterations, the Bleu score can increase about 2 points. Note that there is no tuning for the statistical models (no development data set was used).

4. APPLICATION FOR FRENCH-VIETNAMESE LANGUAGE PAIR

Vietnamese is the 14th widely-used language in the world; however research on MT for Vietnamese is rare. The earliest MT system for Vietnamese is the system from the *Logos Corporation*, developed as an English-Vietnamese system for translating aircraft manuals during the 1970s [1]. Until now, in Vietnam, there are only four research groups working on MT [16]. However the results are still modest.

We focus on building a French-Vietnamese statistical machine translation (SMT) system. The training corpus was created by mining a bilingual news corpus from the Web. The mining process was presented in [17]. In [17], the parameters of mining process were adjusted to obtain parallel sentence pairs. But, in this research, to test the unsupervised method, we adjust the parameters to obtain comparable sentence pairs corresponding to a comparable corpus similar to that of previous section (including wrong parallel sentence pairs).

The initial translation system was built from a comparable training corpus C2 of 30.000 French-Vietnamese sentence pairs. The corpus D contains 21.000 French-Vietnamese sentence pairs. In these corpora, we do not know how many correct parallel sentences are included. The unsupervised method was applied. There is no tuning process for the statistical models. The number of extracted sentence pairs after several iterations was reported in figure 6.

The quality of the translation systems was also evaluated on a test set of 400 manually extracted French-Vietnamese parallel sentence pairs [17]. Each French sentence has only one Vietnamese reference. The evaluation scores were reported in figure 7.

The unsupervised method was applied in a real low-resourced language pair: French-Vietnamese. The result shows that this method can be really applied in the case of lacking parallel data. The quality of the translation system increased during several iterations. We intend to apply this method on a large scale of mining the real comparable data stream extracted from the web.

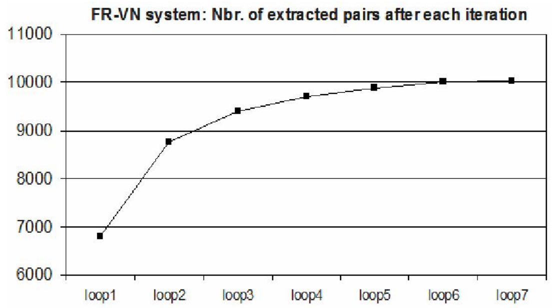


Figure 6: Number of extracted sentence pairs after each iteration in FR-VN translation system

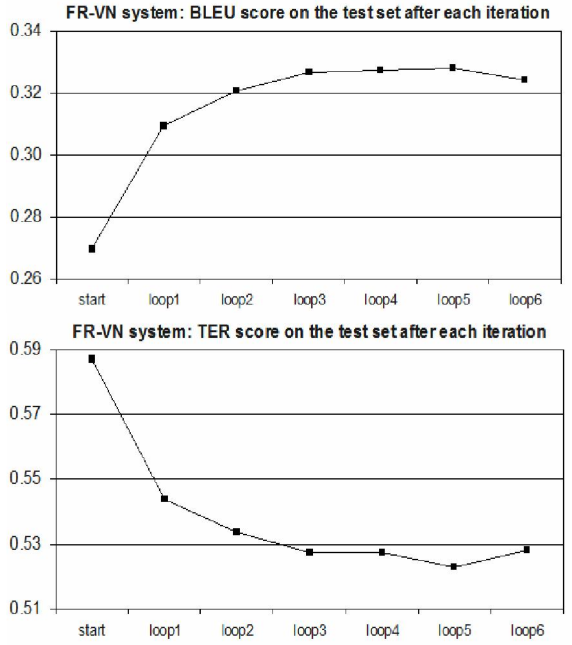


Figure 7: FR-VN translation system evaluations

5. RELATED WORKS

Beside several researches on the semi-supervised method mentioned in previous sections ([9], [10]), there are also researches involving our work. In [18], Zhao and Vogel propose a maximum likelihood criterion which combines sentence length models and a statistical translation lexicon model extracted from an already existing aligned parallel corpus. An iterative process is applied to retrain the translation lexicon model by using the extracted data. Sarikaya et al. present a semi-supervised method with iterations and the initial translation system is based on parallel corpus [19]. The method is also presented as an efficient method in filtering the parallel sentence pairs from a comparable corpus. In this research, authors use a different evaluation metric (Bleu), and use the type of combination like our W2 type. However, their research does not provide a full explanation about how they choose evaluation metric, or combination method, and further more, the problem of decreasing the quality of translation system after several iterations is not mentioned.

6. CONCLUSION AND PERSPECTIVES

This paper presents an unsupervised method for extracting parallel sentence pairs from a comparable corpus. An initial translation system was built based on a comparable corpus, instead of a parallel corpus. The initial translation system was then used to translate another comparable corpus, to withdraw the parallel sentence pairs. An iteration process was implemented to increase the number of extracted parallel sentence pairs and to improve the quality of translation system. The method was tested in a hard condition: the parallel corpus does not exist and the comparable corpus contains up to 50% of non parallel sentence pairs. However, the result shows that this method can be really applied, especially in the case of lacking parallel data. Several ways of using this method was also presented, with different evaluation metrics and different ways of combining the extracted data with the initial translation system. An interesting result is that the quality of the translation system can be improved during the first iterations, but it becomes worse later because of adding the noisy data into the statistical models.

The next work of this research focuses on how to decrease this undesired problem. After some first iterations, the filtering may be altered to respect the precision, instead of the recall. Additionally the new way of reuse extracted parallel sentence pairs will be researched.

11. REFERENCES

[1] Hutchins, W.J. *Machine translation over fifty years*. Histoire, epistemologie, langage: HEL, ISSN 0750-8069, 2001.

[2] Brown, P.F., S.A.D. Pietra, V.J.D. Pietra and R.L. Mercer. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics. Vol. 19, no. 2, 1993.

[3] Koehn, P., F.J. Och and D. Marcu. *Statistical phrase-based translation.* Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Vol. 1, 2003.

[4] Resnik, P. and N.A. Smith. *The Web as a parallel corpus*. Computational Linguistics, 2003.

[5] Kilgarriff, A. and G. Grefenstette. *Introduction to the special issue on the Web as corpus*. Computational Linguistics, volume 29, 2003.

[6] Koehn, P., *Europarl: a parallel corpus for statistical machine translation*. Machine Translation Summit, 2005.

[7] Gale, W.A. and K.W. Church. *A program for aligning sentences in bilingual corpora*. Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1993.

[8] Patry, A. and P. Langlais. *Paradocs: un système d'identification automatique de documents parallèles.* 12e Conference sur le Traitement Automatique des Langues Naturelles, 2005.

[9] Abdul-Rauf, S. and H. Schwenk, *On the use of comparable corpora to improve smt performance*, Proceedings of the 12th Conference of the European Chapter of the ACL, 2009.

[10] Munteanu, D.S. and D. Marcu. *Extracting parallel sub-sentential fragments from non-parallel corpora*. 44th annual meeting of the Association for Computational Linguistics, 2006.

[11] Papineni K., S. Roukos, T. Ward, and W. Zhu. *BLEU:a method for automatic evaluation of machine translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002.

[12] Doddington G. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. In Human Language Technology Proceedings, 2002

[13] Snover M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, *A study of translation edit rate with targeted human annotation*, Proceedings of Association for Machine Translation in the Americas, 2006.

[14] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, R. Zens, M. Federico, N. Bertoldi, B. Cowan, W. Shen and C. Moran. *Moses: open source tool-kit for statistical machine translation*. Proceedings of the ACL, 2007.

[15] Tillmann C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. *Accelerated DP based search for statistical translation*. In 5th European Conf. on Speech Communication and Technology,1997.

[16] Ho, T.B. *Current status of machine translation research in vietnam, towards asian wide multi language machine translation project*. Vietnamese Language and Speech Processing Workshop, 2005.

[17] Do, T.N.D., V.B. Le, B. Bigi, L. Besacier and E. Castelli, *Mining a comparable text corpus for a Vietnamese-French statistical machine translation system*, 4th Workshop on Statistical Machine Translation, 2009.

[18] Zhao B., S. Vogel, *Adaptive parallel sentences mining from Web bilingual news collection*, ICDM, 2002.

[19] Sarikaya R., S. Maskey, R. Zhang, E. Jan, D. Wang, B. Ramabhadran, S. Roukos, *Iterative sentence–pair extraction from quasi–parallel corpora for machine translation*, Interspeech, 2009.

PRODUCTION AND PERCEPTION OF VIETNAMESE FINAL STOP CONSONANTS /P, T, K/

Viet Son Nguyen[1], *Eric Castelli*[1], *René Carré*[2]

[1] MICA Center - HUT CNRS/UMI2954 INP Grenoble - C10 Hanoi university of Technology
01 Dai Co Viet street, Hai Ba Trung, Hanoi, Vietnam
Viet-Son.Nguyen@mica.edu.vn - Eric.Castelli@mica.edu.vn
[2] Laboratoire Dynamique du Langage, UMR 5596, CNRS - Université Lyon 2
14 Avenue Marcelin Berthelot, 69363 Lyon cedex 07, France
Recarre@orange.fr

ABSTRACT

The bursts and voiced formant transitions are well known as separate cues to the place of articulation of initial stop consonant. The Vietnamese presents three final voiceless stop consonants /p, t, k/ without bursts. It is an opportunity to study these final stop consonants and to compare their characteristics with those of the corresponding initial stop consonants. This paper analyses these final consonants in terms of the vowel-consonant (VC) transition duration, the starting formant transition values and the slopes of the VC transition. Measurements have shown that in the same vocalic contexts (the same preceding vowel contexts), the three final stop consonants /p, t, k/ are always clearly different by at least one of the three slopes of F1, F2 and F3. In perception tests, synthesized consonant C in the context /a/-C are recognized as /p/, or /t/, or /k/ when the slopes of the /a/-C transition of F2 and F3 are varied. It means that slopes of the VC transition is an important parameter that allows Vietnamese distinguishing three final voiceless stop consonant /p, t, k/ in Vietnamese language.

***Index Terms*—** Final stop consonant, Vietnamese

1. INTRODUCTION

The problem of perceptual constancy of initial stop consonants with the following vowels in a consonant - vowel or/and consonant - vowel - consonant (CV/CVC) context were studied for a long time ago. In 1954, Liberman showed the role of the transitions of second and third formant in the perception tests of consonants in which the second formant transition F2 is more important than the third formant one [1]. In 1957, Lisker in the researches of perception signs of the initial consonants /w, j, r, l/ concluded that the third formant transition F3 is a good parameter to distinguish two initial consonants /r/ and /l/ [2]. In 1958, Harris also presented that the third formant transition is a good parameters for discriminating two consonants /d/ and /g/ [3]. However, Cole in [4] suggested that the stop consonants pronounced before different vowels may be recognized in terms of a context independent acoustic cue, namely, the bursts produced at the release of initial stop occlusion. Recently, Dorman in [5] noted that bursts and transitions complement each other in the sense that when one cue is weak, the other is usually strong.

In Vietnamese, linguists [6, 7] have demonstrated the existence of six final stop consonants /p, t, k, m, n, ŋ/. However, as opposed to the initial consonants, the three final voiceless stop consonants /p, t, k/ are produced without bursts at the end. We do the hypothesis that the directions and the rates of the formant transitions at the end of the vowel allow distinctiveness. In this paper, these characteristics are studied in the vocalic contexts of all Vietnamese vowels. On the other side, the results of our analysis were validated by the perception tests.

2. STRUCTURE OF VIETNAMESE SYLLABLE

According to studies of linguists, a Vietnamese syllable in its complete form has three parts: initial part, final part and tone. The final part can be divided into three smaller components, i.e. medial part, nucleus part and ending part. So the full form of a syllable has five components: initial part, medial part, nucleus part, ending part and tone (Fig.1). The nucleus part and tone always exist obligatory in a syllable, but the others are optional.

	Tone		
Initial part	Final part		
	Medial	Nucleus	Ending

Figure 1. Structure of Vietnamese syllable [8]

The centre of the Vietnamese syllable, the nucleus part, is always a vowel or diphthong. Vietnamese presents twelve vowels /a, ɛ, e, i, u, o, ɔ, ɤ, ɯ, ă, ɔ̆, ɤ̆/ [6, 7, 9, 10]. The ending part can be one of the six final consonants /p, t, k, m, n, ŋ/ or the two final semi-vowels /w, j/. In this paper, we present the results obtained with the three final voiceless stop consonants /p, t, k/.

3. FINAL STOP CONSONANT ANALYSIS

In order to study the final voiceless stop consonants, a Vietnamese corpus was built from the speech of four male native Vietnamese speakers with mean age of 29. All speakers were born and live in the North of Vietnam, and they speak the standard (Hanoi) dialect. Each subject was asked to pronounce a series of VC2/C1VC2 syllables (five repetitions each) in a Vietnamese carrier phrase meaning "say VC2/C1VC2 softly" where C1 was the initial consonant /b/, C2 was one of the three final stop consonants /p/, /t/, /k/ and V was one of twelve Vietnamese vowels /a/, /ɛ/, /e/, /i/, /u/, /o/, /ɔ/, /ɤ/, /ɯ/, /ă/, /ɔ̆/, and /ɤ̆/. Note that the short vowel /ɔ̆/ is never combined with the two voiceless final consonants /p/ and /t/ (as a consequence, they do not exist in Vietnamese). Vietnamese language is a tonal language with six tones: plat tone (tone A1), falling tone (tone A2), rising tone (ton B1 on sonorant-final syllables and tone D1 on obstruent final syllables) drop tone (tone B2 on sonorant-final syllables and tone D2 on obstruent final syllables), curve tone (ton C1), and broken tone (ton C2) [11, 12]. In order to reduce the influence of tone, it would have been preferable to study Vietnamese syllables in a flat monotonous tone context (tone A1). However, closed syllables ending with /p, t, k/ in Vietnamese may only bear the rising or drop tone (tone D1 or tone D2). As a result, we chose the rising tone configuration which is easily pronounced in Vietnamese. So, the three final voiceless stop consonants /p, t, k/ were combined with twelve Vietnamese vowels, yielding 1360 tokens (1 final consonant /k/ x 12 vowels x 2 contexts VC2/C1VC2 x 5 repetitions x 4 speakers, and 2 final consonants /p, t/ x 11 vowels x 2 contexts VC2/C1VC2 x 5 repetitions x 4 speakers). There are 860 tokens which are actual lexical items, and 500 tokens are not.

Nguyen in his study on the production and perception of nine Vietnamese vowels /a, ɛ, i, u, ɔ, ɤ, ă, ɔ̆, ɤ̆/ showed that the effect of final voiceless stop consonants /p, t, k/ on the vowel duration in both contexts (VC2 and C1VC2) has not been considered. So, we measured the following parameters: formant transition durations, starting formant transition values and formant transition slopes. All the measurements were obtained using the WinSnoori[1] software program.

Fig. 2 represents the formant transition duration of the three final stop consonants /p, t, k/ in twelve preceding vowel contexts (the mean value is calculated for all productions of the four subjects). In the (C1)VC2 context, the vowel /ɔ̆/ is never combined with the two final consonants /p, t/ (these combinations do not exist in Vietnamese). It is easy to realize that in the same vocalic context, the formant transition durations of the three final consonants /p, t, k/ remain constant. Thus, the formant transition duration VC2 cannot bring any distinctive characteristic to the final voiceless stop consonants.

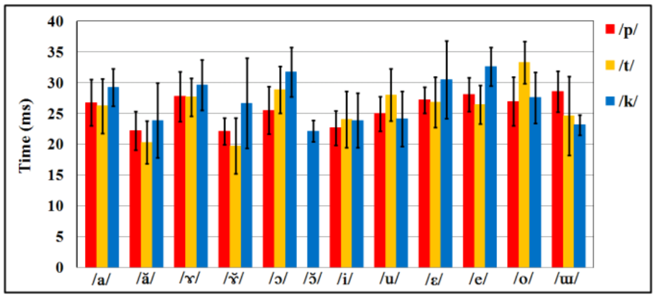


Figure 2. Formant transition duration of VC2 in the (C1)VC2 productions (the vowel /ɔ̆/ is never combined with the two final consonants /p, t/)

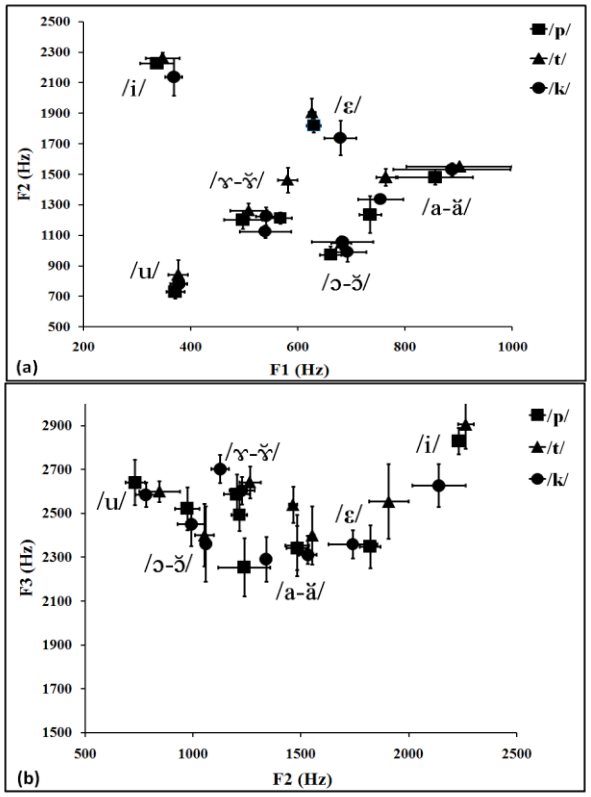


Figure 3. Starting formant transition values of the final consonants /p, t, k/ in different preceding vowel contexts: F1-F2 in (a), F2-F3 in (b)

[1] http://www.loria.fr/~laprie/WinSnoori/

Table I. F1, F2, F3 formant transition slopes (Hz/ms) (mean value and standard deviation (s.d)) in (C1)VC2 productions (the vowel /ɔ̆/ is never combined with the two final consonants /p, t/)

Vowel	/p/			/t/			/k/		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
/a/ s.d	-38 24	-96 31	-39 26	-69 32	7 28	38 45	-82 50	-23 7	-15 41
/ă/ s.d	-48 40	-72 17	-10 34	-47 34	-4 27	69 36	-39 37	-34 29	5 43
/ɤ/ s.d	-30 16	-132 22	-16 19	-17 10	148 16	-7 8	-31 39	1 24	26 43
/ɤ̆/ s.d	-47 19	-121 31	12 13	-33 15	52 22	33 54	-67 33	-51 42	46 15
/ɔ/ s.d	-41 10	-27 12	-42 26	-33 7	53 35	-55 84	-53 29	-27 27	-36 49
/ɔ̆/ s.d							-21 26	-75 46	-9 33
/i/ s.d	-5 7	-123 20	-223 74	-6 7	-81 26	-38 62	-48 17	79 31	41 41
/u/ s.d	-11 13	-4 40	-20 43	-7 22	134 44	-44 35	-31 8	-65 5	34 36
/o/ s.d	-30 20	-6 18	14 29	-1 7	142 6	-109 80	-76 8	-102 24	27 9
/e/	-26	-201	-43	-22	-46	19	-84	217	11
s.d	14	40	37	3	14	26	8	83	28
/ɯ/	-14	-70	-21	-17	98	36	-28	-5	13
s.d	11	42	47	15	54	58	11	5	4
/ɛ/	-24	-175	-88	-24	-51	-26	-33	-126	-62
s.d	6	32	46	19	6	34	46	68	116

In the (C1)VC2 context, the starting transition values of F1, F2, and F3 were defined as the first point where each formant begins its transition from the vowel V to the final consonant C2. Fig. 3 illustrates the starting formant transition values of the final stop consonant in different preceding vowel contexts (the mean value is calculated for all productions of four subjects). Once again, one can observe that in the same context of a preceding vowel, the starting transition values of F1, F2 and F3 of the final stop consonants /p, t, k/ are not clearly distinctive.

In fact, the places of articulation of these consonants are naturally different and the starting transition values of F1, F2, and F3 are more or less close, it is therefore necessary that the slopes are different. To test this hypothesis, we calculate the formant transition slopes VC2 in the same context of the preceding vowel V.

Table I represents the formant transition slopes of each final stop consonant /p/, /t/, and /k/ in the twelve preceding vowel contexts. Fig. 4 illustrates the comparison of the three final consonants /p, t, k/ in the same context of the three preceding vowels /a, i, u/. It is interesting to note that: (1) depending on the vowel context (/a/, or /i/, or /u/), three final consonants /p, t, k/ can be distinguished by at least one of formant transition slope F1 and/or F2 and/or F3; (2) in the context of three preceding vowels /a/, /i/, and /u/, the formant transition slope of F2 is always a good parameter to differentiate the three final consonants /p, t, k/. Nevertheless, to verify and estimate if the formant transition slopes VC allows distinguishing the three final consonants /p, t, k/, we need to perform statistical tests.

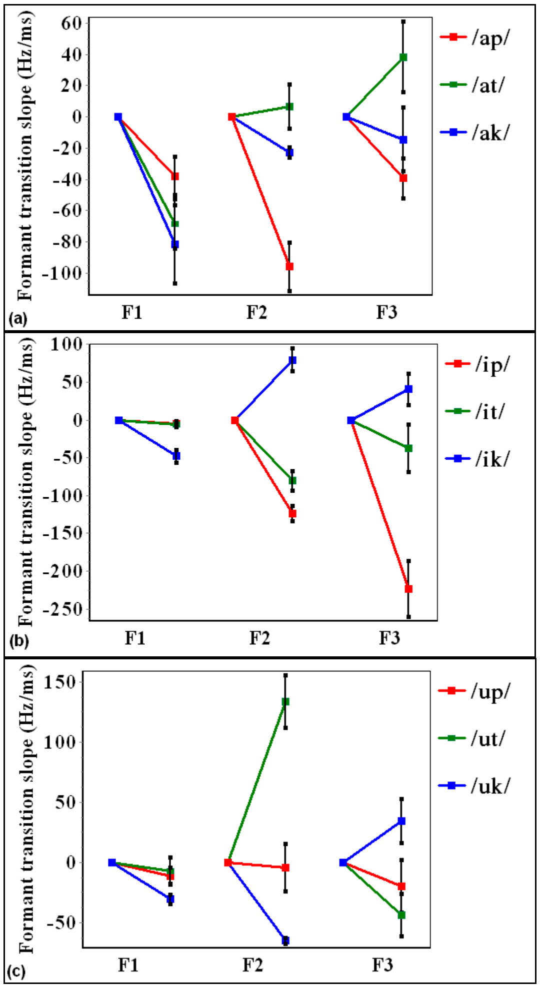


Figure 4. Comparison of the formant transition slopes F1, F2, F3 of the three final consonants /p, t, k/ in the same context of a preceding vowel: /a/ in (a), /i/ in (b), and /u/ in (c). The slope and standard deviation were calculated for all productions (C1)VC2 of four speakers

A statistical test (one-way ANOVA test) of the formant transition slopes comparing the three final stop consonants /p, t, k/ in the same context of a preceding vowel is presented in Table II. In each statistical test (for each preceding vowel and for each formant F1, F2, F3), the formant transition slopes of the three final stop consonants /p, t, k/ were compared. The significant thresholds of 0.05, 0.01, 0.005, and 0.001 were also used to compare with the p-value (significance value) of statistical test. If p-value of one test is smaller than the significant threshold, the hypothesis of the differentiation of these consonants by the corresponding formant in that test is true. From Table II, it is interesting to note that: (1) in all preceding vowel contexts, the three final stop consonants /p, t, k/ are always

Formants		Preceding vowel context										
		/a/	/ɤ/	/ɔ/	/ă/	/ɤ̆/	/i/	/u/	/o/	/e/	/ɯ/	/ɛ/
F1	**F-stat**	1.4	1.8	1.13	0.55	13.63	51.67	7.18	37.2	43.37	2.22	2.49
	p-value	ns	ns	ns	ns	****	****	***	****	****	ns	ns
F2	**F-stat**	21.49	156.8	29.99	16.92	47.35	78.07	84	313.82	89.14	57.22	20.62
	p-value	****	****	****	****	****	****	****	****	****	****	****
F3	**F-stat**	6.75	1.73	0.25	5.94	4.16	45.04	8.49	18.53	5.18	2.68	6.17
	p-value	***	ns	ns	***	*	****	****	****	***	ns	***

Table II. ANOVA tests (p-value and F-statistic) of the formant transition slopes in comparing the three final consonants /p, t, k/ in the same context of a preceding vowel. * = test is significant at 0.05, ** = test is significant at 0.01, *** = test is significant at 0.005, **** = test is significant at 0.001, and ns = not significant

distinguished by at least one of the three slopes of F1, F2, F3 (p-values is always smaller than one of significant thresholds), and (2) the F2 slope is a strong significant parameter that always makes possible the discrimination of these three final stop consonants (p-value of F2 slope is always smaller than the significant threshold of 0.001). It seems that statistically, we can conclude that the formant transition slopes play an important role to distinguish the three final occlusive consonants /p, t, k/. In other words, these three final consonants with bursts articulated with a preceding vowel, change the end of the vowel. The formant transition slopes VC (F1, F3 and especially F2) are characteristics that could allow Vietnamese to recognize these consonants. Although there are three cases (/ɤ/, /ɔ/, /ɯ/ context) where the F3 formant transition slope does not play a significant role, in general, most of our results agree with the results obtained in studies of Liberman on the F2 formant transition [1], and in studies of Harris on the F3 formant transition [3]. However, to confirm this suggestion, we continue with the perception tests to estimate the role of the formant transition slopes in the discrimination of the final consonants /p, t, k/.

4. FINAL STOP CONSONANT PERCEPTION

For the perception tests, a VC syllable is synthesized in which V is the vowel /a/ (with duration of 120 ms). The VC transition duration is 20 ms. The final consonant C is synthesized without burst at the end and with a variation of formant transition slopes as following : (1) the offset value of the first formant (F1offset) is constant (250 Hz) (see Fig. 5); (2) the offset values of the evolution of the two formants F2 and F3 vary as shown on the formant plan F2 / F3 (see the yellow points in Fig. 6) : F2offset value varies from 500 Hz to 2300 Hz, and that of F3offset varies from 1500 Hz to 3300 Hz. In the perception tests, there are forty-three VC syllables synthesized with different formant transition slopes of F2 and F3. Ten listeners (five men and five women) listen five times the syllables presented in random order and they choose what the final consonant is recognized : /p/, or /t/, or /k/, or NAK (NAK (not acknowledgment) is selected for the case where the sound synthesized is not recognized, or it is not /p/, or /t/, or /k/). Table III shows the main results of the perception tests. The average correct recognition rates are calculated for ten listeners. It is interesting to note that by varying the F2 and F3 formant transition slopes: (1) listeners can distinguish the three final consonants /p, t, k/; the best score of the final consonant /p/, /t/ and /k/ are 88%, 92% and 80%, respectively; (2) in the plan of F2 / F3, we can find out three distinct regions corresponding to the three final consonants /p, t, k/ where each one is well recognized; (3) the two final consonants /p/ and /t/ are perceived more easily than the final consonant /k/ (the average score of the consonant /p/ and /t/ is high and the region of these two consonants in the plan F2 / F3 is greater than the one of /k/; (4) generally, in the VC context (V is the vowel /a/), the final consonant C is recognized with the best score as /p/ if F2offset = 1100 Hz, and F3offset = 1500 Hz; as /t/ if F2offset = 1700 Hz, and F3offset = 3000 Hz; and as /k/ if

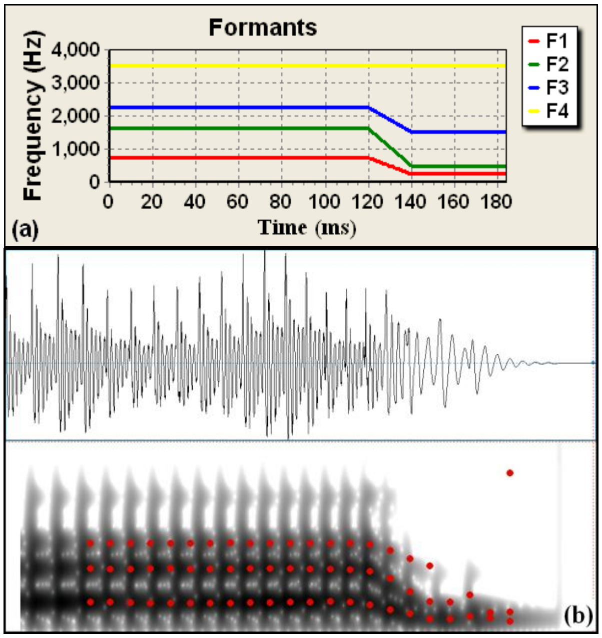


Figure 5. Perception tests of the final stop voiceless consonants /p, t, k/. A Vietnamese syllable VC is synthesized where V is the vowel /a/, the final consonant C is synthesized without burst and with a variation of F2, F3 formant transition slope: the controlled formants in (a), the synthesized signal and the first three formants measured in (b)

both values of F2offset and F3offset are close, 200 Hz and 2100 Hz, respectively.

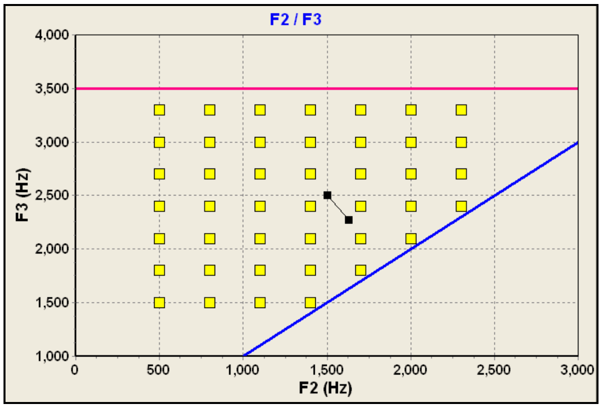


Figure 6. Distribution of the F2, F3 offset values in the plan of F2 / F3 in the perception tests of the three final consonants /p, t, k/: the yellow points are the offset values of the evolution of these two formants (F2offset and F3offset), the blue line and red line are the border values of the evolution of F2offset and F3offset, respectively

5. CONCLUSIONS AND PERSPECTIVES

Our results have shown that, for the three final voiceless stop consonants /p, t, k/, the static characteristics (formant transition durations and starting formant transition values of F1, F2, and F3) are more or less close. However, in all (C1)VC2 contexts, the dynamic characteristics (formant transition slopes) can be considered as good parameters to differentiate the three final voiceless stop consonants /p, t, k/. At the statistic level, the results of analyze have confirmed that in the same vowel context, three final stops /p, t, k/ are always discriminated by at least one of three slopes of F1, or F2, or F3, and the F2 formant transition slope is a particularly significant parameter which always makes the discrimination of these three final consonants. In the perception tests, the results showed that by varying the formant transition slope of F2 and F3, most listeners can recognize the final consonant C in the sequence synthesized VC as one of three stops /p, t, k/, and in the F2 / F3 plan, three regions corresponding to the three final stop consonants /p, t, k/ are distinct. Then, we confirm the assertion of Dorman that if the weight perceptual of bursts is weak, the one of transition is very important [5]. For three Vietnamese final stop consonants /p, t, k/ realized without burst at the end of occlusion, the formant transition slope of F2 and F3 is the only discriminating sign.

On the other hand, Carré [13] pointed out that in V1V2 production, the transition rates of F1 and F2 are necessary and sufficient to represent V2 at the very beginning of the transition and throughout the transition, there is sufficient information to detect V2. In (C1)VC2 production, our results also showed that the formant transition slopes from different preceding vowels to the same final stop consonant are distinguishable. So the formant transition slopes in the identification of the vowel in CV contexts could also be important. Since the formant transition duration remains constant (as in V1V2 production, Carré [13], and in VC, CV combinations, Kent [14]), the corresponding formant trajectories in the acoustic space can be described in term of formant transition rates. Therefore, the time domain could play an important role in the identification of vowels. Such a representation leads to new interpretations of co-articulation, normalization, invariance and vowel reduction.

Table III. Main results of the perception tests for the F2, F3 formant transition slope in the VC context. The average correct recognition rates are calculated for ten listeners

Offset value (Hz)		Correct recognition rate		
F2offset	F3offset	/p/	/t/	/k/
800	1500	80%	8%	12%
800	1800	80%	10%	10%
800	2100	78%	12%	6%
800	2400	76%	18%	4%
800	2700	74%	20%	6%
1100	1500	88%	2%	10%
1100	1800	68%	12%	16%
1100	2100	78%	8%	10%
1100	2400	70%	10%	14%
1100	2700	74%	14%	10%
1100	3000	62%	28%	8%
1400	3000	16%	70%	14%
1400	3300	16%	72%	8%
1700	2400	10%	68%	22%
1700	2700	8%	78%	12%
1700	3000	2%	92%	6%
1700	3300	4%	78%	18%
2000	2100	0%	20%	80%
2000	2400	0%	36%	62%
2000	2700	0%	90%	10%
2000	3000	0%	84%	16%
2000	3300	2%	80%	14%
2300	2400	0%	26%	70%

ACKNOWLEDGMENT

This work is supported by the Vietnamese government project KC.03.15/06-10 (Researching, Designing and Building intelligent automatic devices and systems using human-machine interface by voice in control problems).

6. REFERENCES

[1] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant vowel transitions in the perception of the stop and nasal consonants," *Psychological Monographs*, vol. 68, pp. 1-13, 1954.

[2] L. Lisker, "Minimal cues for separating /w, r, l, y/ in intervocalic position," *Word*, vol. 13, pp. 257-267, August, 1957 1957.

[3] K. S. Harris, H. F. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, "Effect of third-formant transitions on the perception of the voiced stop consonants," *jasa*, vol. 30, pp. 122-126, 1958.

[4] R. A. Cole and B. Scott, "The phantom in the phoneme: Invariant cues for stop consonants," *Perception & Psychophysics*, vol. 15, pp. 101-107, 1974.

[5] M. F. Dorman, M. Studdert-Kennedy, and L. J. Raphael, "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," *Perception & Psychophysics*, vol. 22, pp. 109-122, 1977.

[6] T. T. Doan, *Ngữ âm tiếng Việt (Vietnamese phonetics)*: Hanoi National University Publishing House, 1999.

[7] H. Q. Nguyen, *Ngữ pháp tiếng Việt (Vietnamese Grammar)*: Encyclopedia Publishing House, 2007.

[8] D. D. Tran, E. Castelli, J. F. Serignat, V. L. Trinh, and X. H. Le, "Influence of F0 on Vietnamese syllable perception," in *InterSpeech - EuroSpeech*, Lisbon, Portugal, 2005, pp. 1697-1700.

[9] T. Hoang and M. Hoang, *Remarques sur la structure phonologique du vietnamien* vol. 40. Hanoi: Études vietnamiennes, 1975.

[10] V. S. Nguyen, R. Carré, and E. Castelli, "Production and perception of Vietnamese short vowels," in *Acoustical Society of America Meeting*, Paris, 2008, pp. 3509-3514.

[11] Q. C. Nguyen, "Reconnaissance de la parole en langue Vietnamienne," in *INP-Grenoble dissertation* France: Institut national polytechnique de Grenoble, 2002.

[12] A. Michaud, "Final consonants and glottalization: New perspectives from Hanoi Vietnamese," *Phonetica*, vol. 61, pp. 119-46, Apr-Sep 2004.

[13] R. Carré, "Production and perception of V1V2 described in terms of formant transition rates," in *Proceedings of the Acoustical Society of America Meeting*, Paris, 2008, pp. 2339-2344.

[14] R. D. Kent and K. L. Moll, "Vocal-tract characteristics of the stop cognates," *jasa*, vol. 46, pp. 1549-1555, 1969.

LANGUAGE IDENTIFICATION OF CODE SWITCHING MALAY-ENGLISH WORDS USING SYLLABLE STRUCTURE INFORMATION

Yin-Lai Yeong , Tien-Ping Tan

School of Computer Science, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
yyl09_em0751@student.usm.my, tienping@cs.usm.my

ABSTRACT

This paper introduces a language identification approach using syllable structure information. We also review and compare other approaches. Most of these approaches use linguistic information for language identification. The information used for language identification is Malay affixation information, English vocabulary list, alphabet n-gram, grapheme n-gram. The approach using syllable structure information has the highest accuracy at 93.73% compared to other approaches. Based on the accuracy result of comparison, by using syllable structure 1.91% accuracy had increased for language identification compare with the second higher result in this paper. Syllable structure information is able to gain a better result for language identification.

Index Terms— Language identification, code switching, syllable structure information, Malay, English

1. INTRODUCTION

Language identification is an approach to identify the languages used in a text or speech. In this study, we deal with the problem of language identification primary in text. Language identification is important in many areas of natural language processing. In automatic speech recognition and speech synthesis, language identification is used for identifying the languages of the words before they are converted to their respective pronunciations. In the field of machine translation, the languages need to be determined before the text can be automatic translated. Language identification is also a part of the document categorization system, where it classifies text documents based on the language categories.

Early language identification approaches are only capable of identifying one language given a sentence or document, for instance using *n*-grams or the statistics of short words frequency to determine the language of the words [1], [2], [3]. The approach may fail when there are few languages in the text.

Malaysia is a multilingual society where most people are capable of speaking more than one language. Code switching is a common phenomenon in Malaysian conversation where more than one language is used at the same time. In Malay conversation for example, speakers often switch between Malay and English. There are many reasons why languages are switched from one to another. The most common is to overcome the inability to express one's opinion in the target language. It can also be social where code switching is used to show the social position of the group. Besides speech, code switching also occurs in writing.

Abu Bakar explained that one of the reasons is the education system different between fifteenth century and twentieth century that causes people to mix up English and Malay [4]. Code-switching can occur in Malay speech where single word or string of words are imported from English, and assimilated through a range of phonological and morphological processes [4]

This paper examines a few language identification approaches. We also propose an improvement to the Malay and English language identification which uses syllable information. Section 2 gives an introduction and overview of Malay language. Section 3 discusses five language identification approaches, while the experiment and results are described in section 4. Section 5 present the conclusion and future works.

2. MALAY

Malay is the national language of Malaysia, Indonesia and Brunei. It is also widely spoken in southern Thailand and Singapore. There are many varieties or dialects of Malay. In this paper, we focus only on the standard Malay used in Malaysia. Malay like other languages is also very much influenced by English. A lot of English words have been absorbed into Malay especially in the field of science and technology [5]. However, most people are still more comfortable to combine Malay and English in writing or speech.

Malay is an agglutinative language. One of the features of agglutinative language is the ability of the base word to combine with the prefix, suffix or circumfix to form a new word with different meaning [6]. Prefix is added in front of the word, while suffix is appended at the end of the word.

Ranaivo-Malacon states that only five native prefixes (ber-, per-, ter-, me-, pe-) may create deletion, insertion and assimilation contact with the base word [6].

Table 1 shows some example of Malay prefix, suffix and circumfix.

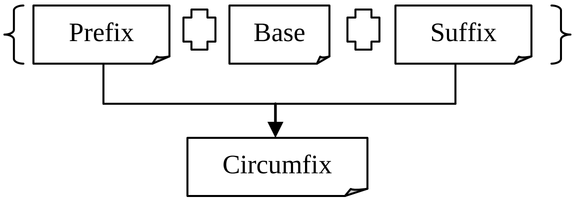


Figure 1: A Malay word structure.

Prefix	‘ber’, ‘di’ , ‘juru’, ‘ke’, ‘pem’, ‘meng’, ‘peng’, ‘per’, ‘ter’, ‘mem’, ‘men’, ‘pen’, ‘me’, ‘pe’, ‘be’, ‘se’, ‘te’
Suffix	‘nya’, ‘kan’, ‘an’, ‘i’, ‘kah’, ‘lah’, ‘tah’
Circumfix	‘ber…an’, ‘per…an’, ‘ter…kan’, ‘mem…kan’, ‘pem…an’, ‘pen…an’, 'pe…an’, ‘ke…an’, ‘se…an’, ‘te…kan’, ‘di…kan’, ‘ber…kan’, ‘me…i’, ‘men…i’, ‘meng…i’, ‘menge…kan’, ‘penge…an’, ‘peng…an’

Table 1: Sample for prefix, suffix and circumfix.

A grapheme is “a minimal unit of a writing system” or “a unit of a writing system consisting of all the written symbols or sequences of written symbols that are used to represent a single phoneme” [7]. Table 2 shows the Malay grapheme and their respective phoneme class [5]. It consists of a maximum of two characters.

Class	Graphemes
Vowel	‘a’, ‘e’, ‘i’, ‘o’, ‘u’
Diphthong	‘ai’, ‘au’, ‘oi’
Plosive	‘p’, ‘b’, ‘t’, ‘d’, ‘k’, ‘q’, ‘g’
Fricative	‘f’, ‘v’, ‘s’, ‘z’, ‘sy’, ‘sh’, ‘kh’, ‘gh’, ‘h’
Affricate	‘c’, ‘j’
Vibrante	‘r’
Lateral	‘l’
Lateral	‘l’
Nasale	‘m’, ‘n’, ‘ny’, ‘ng’
Glide	‘w’, ‘y’

Table 2: Graphemes

Malay syllable structures are shown in Table 3. Most of the words have two or three syllables. Original Malay words have a simple syllable structure, whereas many of the words with two or more consonants that form the coda of a syllable are borrowed from English [5].

Syllable	Word	Description
V	i.kan	**V**.CVC
VC	in.tan	**VC.CVC**
CV	sa.tu	**CV**.CV
CVC	ban.tu	**CVC**.CV
CCV	dwibahasa	**CCV**.CV.CV.CV
CCVC	prak.tik	**CCVC**.CVC
CCCV	stra.tegi	**CCCV**.CVCV
CCCVC	struk.tur	**CCCVC**.CVC

Table 3: Malay syllable structures [4].

3. APPROACHES TO LANGUAGE IDENTIFICATION

This paper discusses five language identification approaches. Most of the approaches discussed make use of linguistic information for language identification. The information used for language identification is affixation, English vocabulary list, alphabet n-gram, grapheme n-gram and syllable structure.

3.1. Affixation information

Malay words can be formed by adding affixation to the base word. By verifying whether the base word belongs to a known base list, we can know whether a word is Malay. This will involve the stripping of the affixation, and then checking the base word in the base word list. However, this approach is not capable of determining whether a word is English. Thus, we assume a given word is either Malay or English.

3.2. English vocabulary list

English vocabulary list can simply be used to determine whether a word is English. If we find the word in the list, then we assume it is an English word, if it is not found in the list then we assume it is a Malay word.

3.3. Alphabet n-gram

We can also model the orthography of the words using alphabet n-gram, and then calculate the probability of an unknown word in different languages [8]. The language with the highest probability is assumed to belong to that language. .

Before the orthography model of different languages can be trained, we must prepare a set of words with their language identified. The words are then segmented to sequence of alphabet based on a particular n-gram order. The segmentation of the word is from left to right. For example bigram alphabet segmentation for the Malay word “BAIK” would be {“_B”, “BA”, “AI”, “IK”, “K_”}. The probability of the alphabet sequence is then calculated for each language.

During the test, given a word with the sequence of alphabet $A = A_1, A_2, A_3,\ldots.A_n$. We want to find the most probable language L, given the alphabet sequence:

$$\begin{aligned}\hat{L} &= \arg\max\ P(L \mid A)\\ &= \arg\max\ \frac{P(L)\,P(A \mid L)}{P(A)}\\ &= \arg\max\ P(L)\,P(A \mid L)\end{aligned}$$

We assume P(L) is the same for all languages. Thus, the probability of the sequence of alphabet, P(A|L) (bigram) is calculated as below.

For indentifying Malay or English word, we will need to calculate the probability of the alphabet sequence for English and Malay, and subsequently selecting the one with the highest probability as the most probable language.

$$P_{English}(A_1, A_2, ..., A_n \mid English) = \prod_{i=1}^{n} P(A_i \mid A_{i-1})$$

$$P_{Malay}(A_1, A_2, \ldots, A_n \mid Malay) = \prod_{i=1}^{n} P(A_i \mid A_{i-1})$$

If the bigram is not found, a small value (1^{-10}) is assigned. The most probable language will be selected.

3.4. Grapheme n-gram

The calculation for this approach is the same as the previous approach. The only difference is the segmentation, where words are segmented to grapheme units instead of alphabets. During the test, given a word with the sequence of grapheme G = G_1, G_2, G_3,….G_n. We want to find the most probable language L, given the grapheme sequence:

$$\begin{aligned}\hat{L} &= \arg\max\ P(L \mid G)\\ &= \arg\max\ \frac{P(L)\,P(G \mid L)}{P(G)}\\ &= \arg\max\ P(L)\,P(G \mid L)\end{aligned}$$

We assume P(L) is the same for all languages. The most probable language will be selected.

3.5. Syllable structure

Before a word is converted to syllable structure, the word is first converted to grapheme sequence (refer to Table 2). The grapheme sequence is then segmented to syllables by determining the largest syllable that can be formed from right to left [5]. Figure 2 shows the syllables forming after converted the word to grapheme..

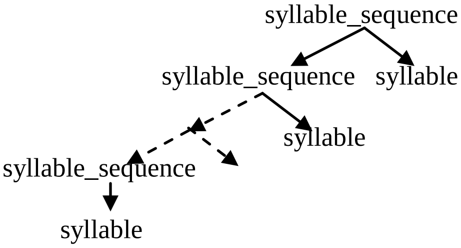


Figure 2: Left-branching structure of syllable sequence.

For example with the previous example, the word “KEBAIKAN” will be segmented to {“_KE”, “KEBAI”, “BAIKAN”, “KAN_”} [8].

The calculation for this approach is same as before. Given the sequence of syllable S = S_1, S_2, S_3,….S_n. We want to find the most probable language L, given the syllable sequence:

$$\begin{aligned}\hat{L} &= \arg\max\ P(L \mid S)\\ &= \arg\max\ \frac{P(L)\,P(S \mid L)}{P(S)}\\ &= \arg\max\ P(L)\,P(S \mid L)\end{aligned}$$

For indentifying Malay or English word, calculate the probability of the sequence of the grapheme in the word, given it is English or Malay.

4. EXPERIMENT

For examining language identification using affixation information, base word vocabulary list contains around three thousand number of base words. As for the language identification using English vocabulary list, the word list was extracted from Carnegie Mellon University (CMU) pronunciation dictionary [10]. The English vocabulary list contains twenty thousand English words.

The approach using alphabet n-gram, grapheme n-gram and syllable n-gram used a separate training and testing word list. Bigram was used for the testing. If the bigram sequence is zero, a small value (1^{-10}) was applied. In the training, twenty thousand vocabularies were selected randomly. As for the testing, ten thousand vocabularies were selected. From the total of ten thousand vocabularies selected for training, 3584 is English and 6416 is Malay. The vocabularies selected for training and testing are different and unique.

	Affi-xation Info.	English Vocab. List	Alphabet Bigram	Grapheme Bigram	Syllable Struc.
Malay Words	4000	8774	6443	6482	6542
English Words	5940	1226	3557	3518	3499
Malay and English	-	-	0	0	41

Table 4: Result for each approach.

	Affi-xation Info.	English Vocab. List	Alphabet Bigram	Grapheme Bigram	Syllable Struc.
Malay as Malay	4051	6356	5994	6040	6167
English as Malay	9	2418	449	442	375
English as English	3575	1166	3135	3142	3244

Malay as English	2365	60	422	376	255
Accuracy %	**76.26**	**75.22**	**91.29**	**91.82**	**93.73**

Table 5: Analyze and accuracy for each approach.

Based on the result, language identification approach using Malay affixation information and English vocabulary list have the lowest accuracy compared with other three approaches. The reasons for the high error rates are not unexpected because these approaches depend on the number of Malay base words and English words.

There is some gain in accuracy of 0.5% from alphabet bigram to grapheme bigram approach. This is because the difference between alphabet and grapheme is not much. Using syllable structure for language identification gives the highest accuracy at 93.73%.

It is interesting to note that Table 4 shows that with the syllable information, it identified 41 words with the same probability. This occurred for those words that can be Malay or English word. The other possibility is because the number of syllables for the word is short and these words have the same value probability.

5. CONCLUSION & FUTURE WORKS

This paper proposed language identification using syllable structure to identify Malay and English word. Experiment results shows the proposed method is able achieve a 93.73% accuracy on 10,000 testing vocabularies. The result is better than other similar approaches using alphabet information and grapheme information. For future work, we will incorporate the syllable structure information with word sequence information. By combining syllable n-gram and word n-gram, the accuracy of the language identification system will be increased further.

6. ACKNOWLEDGEMENTS

This project is supported by the short term grant 304/PKOMP/6310020 “Malay Pronunciation Modeling: Language Identification, Pronunciation Rules and Foreign Word Modeling.

7. REFERENCES

[1] G. Greffenstette, “Comparing two language identification schemes,” in 3rd International Conference on Statistical Analysis of Textual Data, 1995.

[2] J. Prager, “Linguini: language ldentification for multilingual documents,” in 32nd Hawaii International Conference on System Sciences, 1999, pp. 1–11.

[3] J. C. Schmitt, “Trigram-based method of language identification,” U.S. Patent number: 5062143, 1991.

[4] H. Abu Bakar, “Code-switching in Kuala Lumpur Malay, The Rojak Phenomenon”, journal of Southeast Studies, University of California, Volume 9, Spring 2009.

[5] Tien-Ping Tan and Bali Ranaivo-Malançon, “Malay Grapheme to Phoneme Tool for Automatic Speech Recognition”, Third International Workshop on Malay and Indonesian Language Engineering, Singapore, 2009.

[6] B. Ranaivo-Malacon, "Computational Analysis of Affixed Words in Malay Language," Internal Publication, USM, 2004.

[7] http://dictionary.reference.com/browse/grapheme

[8] B. Ranaivo-Malacon and P. K. Ng, Language Identifier for Bahasa Malaysia and Bahasa Indonesian, Proceeding of the 1st Malaysian Software Engineering Conference (MySEC ’05), December2005, Penang, Malaysia, pp.257-259.

[9] Kamus Dewan Edisi Ketiga, 1994

[10] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Acknowledgement

The organising of SLTU Workshop 2010 is a joint effort of many individuals and organisations. They come from within and outside Universiti Sains Malaysia. Their combined technical, administrative and financial support as well as advisory and knowledge-based skills has enabled the Organising Committee to proceed with the various tasks smoothly and systematically. Hence, the Organiser of the 2nd International Workshop on Spoken Languages Technologies for Under-Resourced Languages would like to extend its heartfelt thanks and appreciation to the following and whoever concerned, for their collaboration and support in making this workshop a success:

Vice Chancellor, USM
Deputy Vice Chancellor, Academic and International Affairs, USM
Deputy Vice Chancellor, Students Affairs and Development, USM,
Deputy Vice Chancellor, Research and Innovation, USM
School of Computer Sciences, USM
Centre for Instructional Technology and Multimedia (PTPM), USM
Centre for Knowledge, Communication and Technology (PPKT), USM
Development Department, USM
Security Department, USM
Registry Department, USM
Busary Department, USM
Public Relation Unit, USM
Multimedia, Information, Communication and Applications
(MICA) Research Center, Vietnam
Laboratoire d’Informatique de Grenoble (LIG), France

~Thank You~

Sponsors

Association Francophone de la Communication Parlee

International Speech Communication Association

Speech Processing Asian Network

Centre National de la Recherche Scientifigue

Grenoble INP

National Library of Malaysia

Index of Authors