Reconnaissance de la parole continue à grand vocabulaire en vietnamien, une langue syllabique tonale

NGUYEN Hong Quang [1,2] *, Pascal NOCERA* [1]*, Eric CASTELLI* [2]*, TRINH Van Loan* [2]

[1]Laboratoire d'Informatique d'Avignon LIA - CERI - 439 chemin des Meinajariès, 84190 Avignon CEDEX 9
(Quang.NGUYEN, Pascal.NOCERA)@univ-avignon.fr
[2]International Research Center MICA, HUT - UMI2954/CNRS - INP Grenoble, Hanoi, Vietnam
(Hong-Quang.Nguyen, Eric.Castelli, Van-Loan.Trinh)@mica.edu.vn

ABSTRACT

This paper proposes a method to build a Vietnamese Large Vocabulary Continuous Speech Recognition system (Vietnamese LVCSR system). The difference between Vietnamese and European languages is analyzed and used to adapt a LVCSR system for European languages to Vietnamese. Experiments are implemented on the VNSPEECHCORPUS. The results show that the accuracy of Vietnamese recognition system is increased by using Vietnamese language characteristics.

Keywords: Automatic speech recognition, Vietnamese language, minority language, tone recognition, compound noun.

1. Introduction

Le vietnamien fait partie du groupe des langues Viêt-Muong de la branche môn-khmer de la famille des langues Austro Asiatiques. Du point de vue du linguiste, le vietnamien est une langue tonale syllabique avec six tons lexicaux. La reconnaissance de la parole vietnamienne est juste au début de son développement. Nous constatons que les différences entre le vietnamien et les langues européennes sont telles que les techniques de reconnaissance de la parole communément employées pour les langues européennes (l'anglais, le français etc.) ne suffisent pas pour obtenir de bonnes performances. En effet, la prise en compte des caractéristiques de la langue vietnamienne au niveau de la représentation des données (lexique, modèle de langage) et des modèles (modèle de tons) permettent d'améliorer significativement les résultats.

La première différence est la quantité de ressources électroniques disponibles. En opposition avec l'anglais et le français, le vietnamien est une langue dite "peu dotée". En effet, bien que parlée par près de 80 millions de personnes, les ressources électroniques utilisables dans les technologies linguistiques humaines (HLT) sont peu nombreuses. Cependant, de récentes études portant sur la construction de LVCSR pour des langues peu dotées permettent de construire néanmoins des système avec peu de données.

Deuxièmement, la segmentation des entités sémantiques d'une phrase est différente entre le vietnamien et les langues européennes. L'anglais et le français sont des langues multisyllabiques où chaque mot/concept est facilement identifiable en utilisant des espaces. En vietnamien, les mots/concepts peuvent être composés d'une ou plusieurs syllabes mais sont systématiquement découpés en syllabes séparées par un espace (langue syllabique). La segmentation en mots/concepts de la phrase est une tâche importante pour les langues isolantes telles que le mandarin, le thaï mais aussi pour le vietnamien. Pour améliorer les résultats des traitements automatiques de ces langues, il est en fait souhaitable de construire un module de segmentation en mots multisyllabiques de la phrase syllabique.

La dernière différence est le ton. Le vietnamien est une langue tonale. Dans chaque syllabe, il y a un ton unique. Ce ton est très important pour trouver la signification d'un mot. Si deux syllabes phonétiquement équivalentes ont des tons différents, ils ont des significations différentes. La reconnaissance des tons est donc très importante pour le traitement d'une langue tonale comme le vietnamien, le mandarin, le thaï, etc. Des travaux ont déjà été effectués sur l'utilisation des tons pour la reconnaissance du vietnamien, cependant, ils portent uniquement sur les mots isolés [2] et non pas sur la parole continue. Dans le cadre de la reconnaissance de la parole continue et en l'absence de traitements spécifiques des tons, on peut supposer que 2 mots/syllabes phonétiquement identiques avec des tons distincts seront traités par le système comme deux homophones (ex : et/est en français) et différenciés par le modèle de langage. Cependant, nos expériences démontrent que l'intégration de l'information du ton dans le processus de reconnaissance améliore significativement les résultats.

Dans cet article :
- La Section 2 présente les ressources vietnamiennes disponibles qui ont été utilisées dans nos expériences et nos évaluations.
- La Section 3 présente l'apprentissage du modèle acoustique, le premier résultat et l'adaptation au locuteur.
- La Section 4 développe notre approche sur la constitution et l'utilisation d'un lexique de mots composés.
- La Section 5 décrit notre module de reconnaissance de ton vietnamien et l'intégration de ce module dans le système de reconnaissance de la parole SPEERAL.
- Enfin, la Section 6, donne des conclusions et des perspectives sur les futurs travaux.

2. Ressources vietnamiennes disponibles

2.1. Corpus de la parole

Le corpus de la parole utilisé dans nos expériences est le VNSPEECHCORPUS. C'est un corpus de parole lue, enregistré dans un studio [4]. Il y a deux types de texte dans les enregistrements : paragraphe (80 %) et conversation (20 %). Le signal issu du microphone est échantillonné à une fréquence de 16kHz (16 bits). Nous avons utilisé seulement les enregistrements des 18 locuteurs (10 hommes et 8 femmes) de dialecte standard (au Nord du Viêt-Nam) ce qui représente environ 14,4 heures de parole. Nous avons divisé ce corpus en deux parties : 8 hommes et 6 femmes pour l'apprentissage (environ 11,2 heures de parole), et 2 hommes et 2 femmes pour les tests (environ 3,2 heures de parole).

2.2. Dictionnaire de prononciations vietnamien

Le vietnamien est une langue syllabique tonale compremant 6 tons. Elle utilise environ un vocabulaire de 6698 syllabes. Un lexique de prononciation pour le vietnamien a été construit en appliquant VNPhoneAnalyzer sur ce vocabulaire [5].

2.3. Corpus de Texte et Modèle de Langage

Pour des langues peu dotées telles que le vietnamien, une méthode de construction de corpus de texte et de modèle de langage a été présentée par [6]. Selon cette méthode, le corpus de texte est construit à partir des données issues de sites Web de journaux électroniques vietnamiens. Les données superflues (les menus, les références, les publicités, les annonces, etc.) ont été filtrées et les phrases contenant uniquement des mots du vocabulaire (le lexique des mots monosyllabes) ont été extraites. Dans le corpus final, il y a environ 2.7 millions de phrases avec 45 millions syllabes.

L'apprentissage de notre modèle de langage trigramme, a été effectué à l'aide des outils du CMU[1] avec l'algorithme de Good-Turing et Katz pour le back-off. La valeur de la perplexité du modèle de langage ainsi obtenu sur le corpus de test est de 72.74.

3. Première expérience

3.1. Modèle acoustique

Des modèles acoustiques non contextuels ont été utilisés dans nos expériences. Chaque phonème vietnamien est représenté par un modèle de Markov caché (HMM) à 5 états : 3 états émetteurs, le prémier état et le dernier état sont des états non émetteurs. Chaque état émetteur est modélisé par 64 gaussiennes. Le vecteur de paramètre a été extrait toutes les 10 ms et contient 39 coefficient (13 MFCC, leurs dérivées premières et secondes).

La méthode utilisée pour l'apprentissage des modèles acoustique est basée sur des méthodes récemment développées pour le traitement des langues minoritaires. Elle utilise un modèle acoustique français (qui a été appris sur un corpus français conséquent) comme modèle de bootstrapping [1][5]. Un tableau de correspondances phonémiques entre le français et le vietnamien a été préalablement créé en utilisant les connaissances phonétiques de chacune des langues. Ce tableau nous a permis de construire le premier modèle acoustique vietnamien à partir du modèle acoustique français. Le modèle, quant à lui, nous a permis d'effectuer un alignement du corpus d'apprentissage et d'apprendre des modèles spécifiquement vietnamiens. Dans notre expérience, le modèle acoustique français a été appris sur BREF80[2].

Pour nos expériences, nous avons utilisé le système Speeral [7]. Speeral est un système de reconnaissance automatique de la parole continue (système LVASR) développé par le Laboratoire d'Informatique d'Avignon (LIA). Le premier résultat (WER : le taux d'erreur de mot) sur le corpus de test était 34,7%.

3.2. Adaptation au locuteur

Une adaptation non supervisée au locuteur a été effectuée. Pour cela, nous avons fait plusieurs itérations "reconnaissance/adaptation" en utilisant la technique d'adaptation MLLR. Etant donné la faible taille du corpus d'apprentissage des modèles acoustiques de départ, cette tâche nous a permis d'augmenter significativement les performances du système puisque le WER est passé de 34,7% à 25,2%.

4. Mots multisyllabiques

Pour utiliser des mots multisyllabiques dans un système LVASR vietnamien, il est nécessaire d'effectuer un certain nombre de traitements. Il faut bien sur avoir un lexique multisyllabiques mais également un corpus de texte composés de mots multisyllabiques pour l'apprentissage du modèle de langage.

4.1. Représentation des mots multisyllabiques

En français, les espaces servent à séparer les mots. Dans le vietnamien, les espaces séparent les syllabes. Cependant, il est difficile d'isoler les mots multisyllabiques dans un corpus, puisque les syllabes elles même ont aussi une signification sémantique et que le regroupement n'est pas unique et conditionné par la sémantique générale de la phrase. Par exemple, le mot bisyllabique 'xe1_bo2'[3] (une charrette) a deux syllabes : 'xe1' (véhicule) et 'bo2'(bœuf). La phonétisation d'un mot multi syllabique a été obtenue par la concaténation des phonétisations des syllabes qui le composent.

Plusieurs lexiques ont été testés : un premier lexique sémantique (Section 4.2) et des lexiques obtenus automatiquement (Section 4.3)

[1] http ://www.speech.cs.cmu.edu/SLM_info.html

[2] http ://www.elda.fr/catalogue/en/speech/S0006.html

[3] Les chiffres '1' et '2' représentent les tons de chaque syllabe.

4.2. Utilisation du lexique sémantique

Afin d'apprendre le modèle de langage sur le nouveau lexique, il a fallu transformer le texte d'apprentissage en fusionnant les syllabes en mots multisyllabiques. En cas d'ambiguïté (il y a plusieurs mots commençant par la même syllabe), différentes stratégies ont été testées pour choisir un regroupement :
– le mot au hasard (RM : Random Matching),
– le mot le plus long (MM : Maximum Matching),
– l'hypothèse de segmentation de la phrase avec le meilleur total de l'information mutuelle des mots (MMIS : Maximum Mutual Information of the sentence).

4.3. Utilisation du lexique construit automatiquement

Nous avons testé plusieurs méthodes pour construire des lexiques de reconnaissance en regroupant les syllabes automatiquement. L'objectif était d'extraire et d'isoler des séquences de syllabes souvent rencontrées, sans pour autant qu'elles aient une signification sémantique. Dans ce cas, les séquences peuvent être des mots, des parties de mots, ou bien un groupe de mots. Deux méthodes ont été étudiées pour ce traitement : utilisation des informations mutuelles et du critère de perplexité maximale, utilisation du modèle de langage des syllabes et de l'information mutuelle des mots.

La première méthode regroupe une suite de mots si son information mutuelle est supérieure à S_{MI} et son nombre d'apparition supérieur à S_{min}. Elle permet de construire un nouveau corpus sur lequel on peut relancer l'algorithme pour trouver de nouvelles séquences. Ce processus s'arrête si la perplexité du corpus de test ne décroît plus [3]. Cette méthode est appelée MI_PERP.

Méthode	Nb mots	SER	SER + Adapt
Lexicon + RM	40342	30.5	22.6
Lexicon + MM	40342	30.1	22.3
Lexicon + MMIS	40342	30.0	22.2
MI_PERP	36558	27.5	20.6
DP_MI	39823	26.2	19.4

Tab. 1: Résultat des expériences sur LVASR vietnamien avec les modèle de langages différents

Dans la deuxième méthode, le découpage des phrase en une suite de mots (monosyllabiques ou multisyllabiques) est réalisé par un algorithme de programmation dynamique en utilisant le modèle de langage des syllabes (la durée d'un mot multi syllabique n'était pas dépassée de 4 syllabes). Chaque liste de groupe de syllabes est ensuite analysée et la liste la plus significative en terme d'informations mutuelles est retenue. L'information mutuelle des groupes de syllabes est calculée sur le corpus d'apprentissage de texte et les groupes ayant un nombre d'apparition inférieur à un seuil K sont éliminés. Cette méthode est appelée DP_MI.

Les résultats obtenus avec ces nouveaux lexiques et corpus avant et après adaptation au locuteur sont résumés dans la Table 1.

5. Utilisation d'informations du ton

Dans cette section, nous décrivons notre module de reconnaissance de tons vietnamiens et notre méthode pour utiliser ce module dans le LVASR vietnamien.

5.1. Reconnaissance de ton vietnamien

Corpus de ton Le corpus d'apprentissage des tons est un corpus de parole continue pour lequel nous connaissons la nature et la frontière des tons. Au début, la frontière des syllabes a été détectée automatiquement par l'algorithme de Viterbi en utilisant le modèle acoustique précédent. Ces frontières ont ensuite été corrigées manuellement et le segment voisé de la syllabe a été choisi pour caractériser le segment de chaque ton.

Nos expériences ont montré que les résultats ne sont pas homogènes entre les locuteurs masculins et féminins et qu'un modèle dépendant du genre était plus robuste. C'est pourquoi, nous avons travaillé uniquement sur les locuteurs masculins dans les expériences suivantes.

Deux approches ont été expérimentées pour la caractérisation des tons. Dans la première approche, chaque trame du signal est représentée par un vecteur et un modèle de Markov caché (HMM) calcule le score du ton comme pour les phonèmes. Dans la deuxième approche, plus globale, chaque ton est représenté par un seul vecteur et des réseaux de neurones artificiels (ANN) a été appliquée pour effectuer la reconnaissance. Les résultats pour ces 2 méthodes sont présentés dans la Table 2.

Reconnaissance de ton en utilisant des réseaux de neurones Nous avons utilisé pour cela un perceptron multicouches. Il se compose d'une couche d'entrées qui reçoit les paramètres du ton, d'une couche cachée et d'une couche de sortie qui représente le score de chaque ton. La règle de la rétropropagation du gradient de l'erreur a été appliquée pour l'apprentissage du réseau.

Méthode	TTC (%)
Utilisant des réseaux de neurones	
Non contextuel	71,31%
Non contextuel + détaillé (meilleur N = 4)	72,37%
Non contextuel + coupant 25% du ton	72,76%
Contextuel	73.67%
Utilisant des modèles de Markov cachés	
Non contextuel	75.80%
Non contextuel + coupant 25% du ton	77.02%

Tab. 2: TTC (Taux de tone correct) des tests de reconnaissance de ton

Deux modélisation ont été testées, des modèles non contextuels et des modèles contextuels de tons. Les modèles non contextuels sont estimés sur l'intégralité du segment voisé du ton alors que pour les modèles contextuels, on a pris 3 parties : le segment voisé du ton principal, la partie finale du segment voisé du ton précédent et le segment non voisé entre le ton principal et le ton précédent.

Pour modéliser le ton, le segment voisé du ton principal a été divisé en **N** parties. Les courbes de F_0 et de l'énergie de chaque partie ont été modélisées par 4 paramètres de Légendre ou par la moyenne et la pente. Cette technique a également été appliquée sur la partie finale du segment voisé du ton avant. Pour le segment non voisé entre le ton principal et le ton avant, nous avons choisi de le représenter par la durée et la moyenne de l'énergie du segment. De plus, la durée du segment voisé et du segment non voisé au début du ton principal a été ajoutée dans le jeu de paramètre de ton.

Pour nous affranchir des phénomènes de coarticulation, nous avons utilisé soit le model contextuel soit le model non contextuel auquel nous avons enlevé K% des trames du début du ton (K a été choisi empiriquement).

Reconnaissance de ton en utilisant le modèle de Markov caché Dans nos expériences, nous avons utilisé un HMM pour chaque ton. Ils sont composés de 3 états émetteurs chacun et sont modélisés par 16 gaussiennes. Chaque trame du signal est représenté par un vecteur de 6 paramètres : $[f_0, \Delta f_0, \Delta\Delta f_0, e, \Delta e, \Delta\Delta e]$, f_0 et e sont les valeurs normalisées de la fréquence fondamentale et de l'énergie du trame. Comme dans le cas utilisant des réseaux de neurones, nous avons fait deux tests : avec estimation et sans estimation de l'événement de coarticulation.

5.2. Utilisation de l'information des tons dans un système LVASR vietnamien

Dans la littérature, deux approches ont été proposées pour traiter les tons dans un système de reconnaissance du vietnamien. Dans la première approche, le modèle acoustique est un modèle de phonème "tonal" et la valeur de F0 est insérée dans le vecteur de paramètres [8]. Dans la deuxième technique, le modèle de phonème est indépendant du ton et la décision entre les mots ayant la même suite de phonèmes dans le lexique mais des tons différents est prise par le modèle de langage [5].

Dans les expériences précédentes, c'est la deuxième approche qui a été utilisée par le système Speeral puisque qu'aucune information prosodique sur des tons n'était présente dans le processus de décision. Dans les expériences qui suivent, la décision entre les mots différents sera faite non seulement par le modèle de langue, mais aussi par le modèle de reconnaissance de ton. Cette méthode a été mise en oeuvre comme suit : dans Speeral, chaque hypothèse de mot est calculée en utilisant le modèle de langage comme dans (1).

$$Score_{word} = \alpha \cdot Score_{acous} + \beta \cdot Score_{ML} \qquad (1)$$

$Score_{acous}$ et $Score_{ML}$ sont le score acoustique et le score du modèle de langage du mot. Cette expérience a été présenté dans la section 3. Le SER sur les locuteurs masculins était 34.6% et 21.3% après adaptation au locuteur.

Pour chaque hypothèse de mot, nous connaissons précisément la frontière du mot ainsi que celle des phonèmes qui le composent. Nous pouvons donc utiliser notre module de reconnaissance de ton présenté précédemment pour calculer le score du ton de ce mot et l'intégrer au score de l'hypothèse (2).

$$Score_{word} = \alpha \cdot Sc_{acous} + \beta \cdot Sc_{ML} + \gamma \cdot Sc_{tone} \qquad (2)$$

α , β, γ ont été choisis empiriquement. Des premières expériences ont été effectuées en utilisant le lexique syllabique. Le SER sur le corpus de test (uniquement les locuteurs masculins) est alors de 26.0% et de 24.9% en utilisant les ANN et les HMM, et chute à 15,2% et 13,5% si on utilise les modèles acoustiques adaptés au locuteur.

6. Conclusions

Ce papier présente notre étude pour la construction du vietnamien LVASR en utilisant différentes techniques prévues pour les langues minoritaires. L'intégration de caractéristiques du vietnamien dans le système est aussi présentée. Les résultats montrent que les performances du système LVASR vietnamien augmentent significativement en intégrant ces caractéristiques.

Pour les travaux futurs, le module de reconnaissance des tons sera intégré dans le système de reconnaissance de la parole qui utlise le modèle de langage multi syllabique.

Références

[1] Nimaan Abdillahi, Nocera Pascal, and Bonastre Jean-François. Towards automatic transcription of somali language. In *LREC 2006*, Genoa, Italy, 24-26 May, 2006.

[2] N.Q. Cuong, E. Castelli, and Ngoc-Yen. Pham. Tone recognition for vietnamese. In *EUROSPEECH'2003*, Geneva, 2003.

[3] I.Zitouni and K.Smaïli. Vers une meilleure modélisation du langage : le prise en compte des séquences dans les modèles statistiques. In *JEP'00*, AUSSOIS – France, 2000.

[4] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, and J-F. Serignat. Spoken and written language resources for vietnamese. In *LREC 2004*, volume II, pages 599–602, Lisbon, Portugal, May 26-28, 2004.

[5] Viet Bac LE and Laurent BESACIER. First steps in fast acoustic modeling for a new target language : application to vietnamese. In *ICASSP 2005*, pages 821–824, Philadelphia, USA, May, 2005.

[6] Viet Bac LE, Brigitte BIGI, Laurent BESACIER, and Eric CASTELLI. Using the web for fast language model construction in minority languages. In *Eurospeech 2003*, pages 3117–3120, Geneva, Switzerland, September 2003.

[7] P. Nocera, G. Linares, D. Massonié, and L. Lefort. Phoneme lattice based asearch algorithm for speech recognition. In *TSD'02*, Brno, Czech Republic, September, 2002.

[8] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, and John-Paul Hosom. Vietnamese large vocabulary continuous speech recognition. In *Interspeech 2005*, Lisbon, Portugal, September, 2005.