

ETAPE

Speech Transcription

Speech Group - LORIA
18-19 June 2012

ETAPE – Speech Transcription

- Data overview
 - Speech data
 - Text data
- Focus on some modules
 - Vocabulary selection: NN-based approach
 - Language modeling
 - Lexicon: CRF-based grapheme-to-phoneme conversion
 - Acoustic modeling: Class-based speech recognition
- Speech transcription systems
 - Overview of individual systems
 - Systems combination

Speech Data for Acoustic Modeling

	Train set	Dev set
Ester2	Train (<i>audio</i>) ~ 217 h Dev (<i>audio</i>) ~ 7 h Test (<i>audio</i>) ~ 7 h	---
Epac	(<i>audio</i>) ~ 112 h	---
Etape	Train (<i>audio</i>) ~ 26 h	Dev (<i>audio</i>) ~ 8 h
Total	<i>Useful training frames</i> ~ 261 h	

Text Data for Language Modeling

Data type		87 ... 97	98 ... 05	06 ... 11
News papers (87..07)	Le Monde, L'Humanité	265 Mw	231 Mw	28 Mw
Radio show transcripts (98..05)	Ester, ...		113 Mw	
Web data	News paper web sites			162 Mw
	Radio & TV web sites			40 Mw
	Other sites			131 Mw
GigaWord (94..08)	AFP + APW	152 Mw	427 Mw	180 Mw
<i>Total</i>		417 Mw	771 Mw	541 Mw
		<i>1729 M word tokens</i>		

Etape Train set : 276 000 word tokens – used for optimizing language models

Vocabulary Selection

Automatic NN-based Selection

- New selection method based on Neural Networks
- Principle
 - Use a Neural Network to compute a score for each possible word (objective: 1 for words to be selected, and 0 for useless words)
 - Training the NN parameters
 - Target = 1 for words present in the ETAPE training data
 - Skip words belonging to known lexicon BDLex
 - Target = 0 for remaining words
 - Vocabulary selection
 - Apply the trained NN to each word
 - Select the ones with largest score
- Vocabulary selection
 - Slightly better results than with unigram-based approaches
 - Almost as good as manual method when using simple set of occurrence counts as features
 - Final features: occurrence count in 20 text data subsets
+ number of different n-grams in which word occurred (n=1, 2, 3 & 4) in 4 data types

Vocabulary Selection

- Combining manual-based and NN-based lexicons
 - Manual-based lexicon (85000 items) - words occurring at least once in various data subsets, plus large French cities and countries
 - NN-based lexicon (85000 items)
 - Merging lexicons → final lexicon (97 000 items)
- And some manual adjustments
 - From the analysis of most frequent errors on ETAPE training data
correction of some pronunciations and removal of unwanted unaccented words (*ex. “etats” instead of “états”*)

Absolute WER variation - <u>compared to manual selection</u>	ETAPE
Merging manual-based and NN-based selection (→ 97 000 items)	-0.1%
+ Manual adjustments	-0.4%

Language Modeling

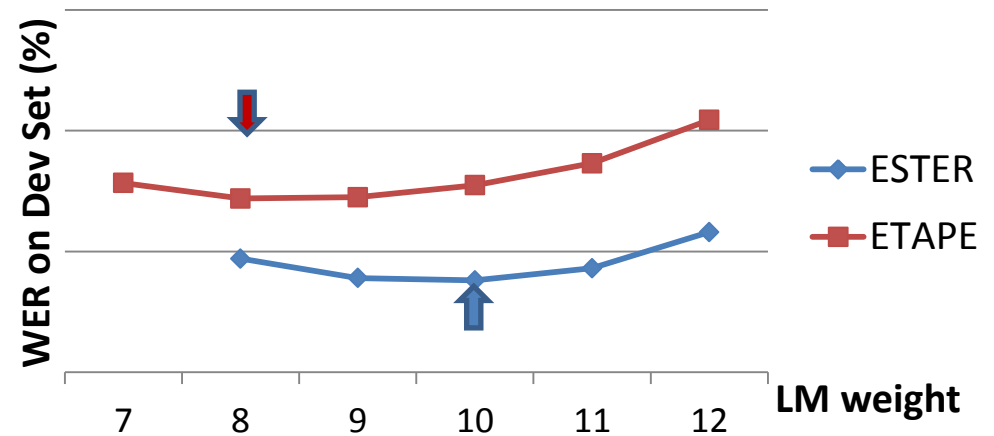
- Ngram-based – SRILM Toolkit
 - Estimating counts on several text data subsets (4 or 5)
 - Optimizing linear combination weights using transcripts of Etape training data

Combination weights	Speech transcripts		Text data		
	Ester	Other radios	News papers	Web data	Giga word
3-gram LM (forward)	0.68		0.06	0.25	0.01
4-gram LM (backward)	0.30	0.35	0.10	0.22	0.03

Model size	1g	2g	3g	4g
3-gram LM (forward)	97 k	43 M	79 M	
4-gram LM (backward)	96 k	40 M	255 M	139 M

Language Modeling

- Optimal LM weight different between ESTER & ETAPE



Absolute WER variation	ETAPE
Optimizing language model weight (from 10 down to 8)	-0.5%
And adjusting word insertion penalty (from 0.7 down to 0.5)	-0.1%

Acoustic Modeling

Standard Modeling

- Standard modeling relies on 4 models
 - Corresponding to speech quality: studio vs. telephone
 - And speaker gender: male vs. female

Absolute WER variation	ETAPE
Increasing training set (<u>from 190 hours to 260 hours</u>)	-1.0%
Increasing number of senones (<u>from 4500 to 6500</u>) with ETAPE training	-0.5%
Increasing again (<u>from 6500 to 8500</u>) with larger artificial corpus (warped mfcc)	-0.3%

Acoustic Modeling

Class-based Modeling

- Use an arbitrary number of data classes instead of standard male/female classes
 - Automatic classification of training data
 - Standard adaptation of acoustic models on class data → ok for 2 or 4 classes, but then, WER degrades when number of classes gets larger
 - So, introduction of a **tolerance margin in classification process** (data close to class boundary may belong to several classes)
→ more data for each class, better training & better results [ICASSP'2012, JEP'2012]
 - Further improvement obtained using phonetic models (instead of a single GMM) and a KL divergence criterion for data classification (in test, requires a previous decoding pass for determining the sequence of phones of the speech segment)

Absolute WER variation – <u>compared to standard 4 models (quality & gender)</u>	ESTER2
GMM-based classification, tolerance margin, 16 classes, MLLR+MAP adaptation	-0.5%
KL-based classification, tolerance margin, 8 classes, MLLR+MAP adaptation	-0.7%

Pronunciation Modeling

Grapheme-to-Phoneme Conversion

- CRF-based approach developed and investigated (Conditional Random Field) [IS'2011]
 - Input: window of 9 letters
 - Output: phoneme (or null-phoneme)
 - Training needs letter-to-phoneme alignments – obtained with a discrete HMM
- CRF provides slightly better results than JMM (Joint Multigram Models – state of the art)
- Both CRF and JMM based approaches can generate multiple pronunciation variants
- CRF takes benefit of language origin for generating pronunciation variants of proper names [JEP'2012]
- Generation of pronunciation lexicons for Etape
 - First, from available lexicons (BdLex, some proper names, acronyms)
 - Automatic for remaining items, using CRF and JMM models

Systems Summary & Results on ETAPE Development Set

	Lexicon & LM	Features	Acoustic	Second pass
ANTS Diarization + SPHINX decoding	Lex 97K words (NN + manual)	Aurora mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female <5s → CMN, 6500s, 64g ≥5s → CMN+VNorm, 4500s, 128g	MLLR (>2.5s)
	Pron. new words (CRF + JMM)	Sphinx mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female noCMN] 2.5s [CMN] 5s [CMN+Vnorm, 8500s, 128g	----
	LM 3g (cutoff 1.1.2)		Studio/Tel – 16 classes GMM-based, CMN, 8500s, 64g	VTLN + MLLR (>2.5s)
			Studio/Tel – 8 classes KL-based, CMN, 8500s, 64g	VTLN + MLLR (>2.5s)

Systems Summary & Results on ETAPE Development Set

	Lexicon & LM	Features	Acoustic		Second pass
ANTS Diarization + SPHINX decoding	Lex 97K words (NN + manual)	Aurora mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female <5s → CMN, 6500s, 64g ≥5s → CMN+VNorm, 4500s, 128g		MLLR (>2.5s)
	Pron. new words (CRF + JMM)	Sphinx mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female noCMN] 2.5s [CMN] 5s [CMN+Vnorm, 8500s, 128g		----
	LM 3g (cutoff 1.1.2)		Studio/Tel – 16 classes GMM-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)
			Studio/Tel – 8 classes KL-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)
LIUM Diarization + HTK+Julius decoding	Lex 96K words (manual)	HTK mfcc + HLDA 9 fr. → 40 coef.	CMN+Vnorm per speaker, 6000 shared states, 30g	All phoneme apertures Studio/Tel – Gender independent	SAT CMLLR per speaker
	Pron. new words (CRF proper names)			Merged vowel aperture for o/e/ø Studio/Tel – Gender independent	
	LM 4g (cutoff 1.1.1.2)			All phoneme apertures Studio/Tel – Male/Female	

Systems Summary & Results on ETAPE Development Set

	Lexicon & LM	Features	Acoustic		Second pass	WER
ANTS Diarization + SPHINX decoding	Lex 97K words (NN + manual)	Aurora mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female <5s → CMN, 6500s, 64g ≥5s → CMN+VNorm, 4500s, 128g		MLLR (>2.5s)	31.56%
	Pron. new words (CRF + JMM)	Sphinx mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female noCMN] 2.5s [CMN] 5s [CMN+Vnorm, 8500s, 128g		----	30.27%
	LM 3g (cutoff 1.1.2)		Studio/Tel – 16 classes GMM-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)	29.01%
			Studio/Tel – 8 classes KL-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)	29.35%
LIUM Diarization + HTK+Julius decoding	Lex 96K words (manual)	HTK mfcc + HLDA 9 fr. → 40 coef.	CMN+Vnorm per speaker, 6000 shared states, 30g	All phoneme apertures Studio/Tel – Gender independent	SAT CMLLR per speaker	31.25%
	Pron. new words (CRF proper names)			Merged vowel aperture for o/e/ø Studio/Tel – Gender independent		30.59%
	LM 4g (cutoff 1.1.1.2)			All phoneme apertures Studio/Tel – Male/Female		30.94%

Systems Summary & Results on ETAPE Development Set

	Lexicon & LM	Features	Acoustic		Second pass	WER
ANTS Diarization + SPHINX decoding	Lex 97K words (NN + manual)	Aurora mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female <5s → CMN, 6500s, 64g ≥5s → CMN+VNorm, 4500s, 128g		MLLR (>2.5s)	31.56%
	Pron. new words (CRF + JMM)	Sphinx mfcc + Δ + $\Delta\Delta$ → 39 coef.	Studio/Tel – Male/Female noCMN] 2.5s [CMN] 5s [CMN+Vnorm, 8500s, 128g		----	30.27%
	LM 3g (cutoff 1.1.2)		Studio/Tel – 16 classes GMM-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)	29.01%
			Studio/Tel – 8 classes KL-based, CMN, 8500s, 64g		VTLN + MLLR (>2.5s)	29.35%
LIUM Diarization + HTK+Julius decoding	Lex 96K words (manual)	HTK mfcc + HLDA 9 fr. → 40 coef.	CMN+Vnorm per speaker, 6000 shared states, 30g	All phoneme apertures Studio/Tel – Gender independent	SAT CMLLR per speaker	31.25%
	Pron. new words (CRF proper names)			Merged vowel aperture for o/e/ø Studio/Tel – Gender independent		30.59%
	LM 4g (cutoff 1.1.1.2)			All phoneme apertures Studio/Tel – Male/Female		30.94%
Combination of all system outputs using ROVER					24.52%	

Summary

- No refined optimization of a multipass system
- No work carried out on diarization
- Focus on a few modules
 - Several small improvements achieved in various places (vocabulary, pronunciations, language models, acoustic models)
- Leading to a set of speech transcription systems having different characteristics and performance in the range 29.0% to 31.6% WER (on Dev data, excluding overlapped speech)
- Efficient combination of the output results using ROVER
 - ➔ 24.5% WER on Dev data (excluding overlapped speech)

Thank you