

Development of Polish Large Vocabulary Continuous Speech Recognition acoustic models

Grażyna Demenko^{1,2} Stefan Grochowski^{1,3} Katarzyna Klessa²
Marek Lange¹ Bartosz Rapp¹ Marcin Szymański¹

¹Poznań Supercomputing and Networking Center

²Adam Mickiewicz University, Institute of Linguistics

³Poznań University of Technology, Institute of Computing Science

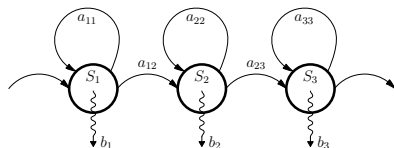
i3 Conference, Nov. 5, 2009

LVCSR – Large Vocabulary Continuous Speech Recognition:

- typically able to use vocabularies of 20k–100k words
- mostly developed for non-inflectional languages
- acoustic models trained on large corpora (over 100h of speech)
 - speakers selected to represent typical distribution of age, sex and dialect

Hidden Markov Model

Context-dependent phoneme realizations are represented by HMMs



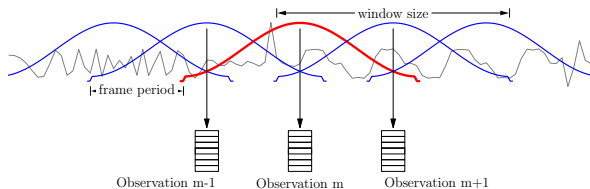
Phone-level hidden Markov model with
three emitting states
and no skip-transitions.

Legend

a_{ij} transition probabilities

b_j observation emission distribution

Speech parametrization



Short-time signal analysis by overlapping Hamming windows.
MFCC parameters are commonly used as feature vectors.

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation

Training procedure

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation
- Conversion to *plain* triphones (context-dependent models)

Training procedure

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation
- Conversion to *plain* triphones (context-dependent models)
 - /o/ /s'/ /e/ /m/

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation
- Conversion to *plain* triphones (context-dependent models)
 - /o/ /s'/ /e/ /m/
 - /o+s'/ /o-s'+e/ /s'-e+m/ /e-m/

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation
- Conversion to *plain* triphones (context-dependent models)
 - /o/ /s'/ /e/ /m/
 - /o+s'/ /o-s'+e/ /s'-e+m/ /e-m/
- Context-dependent *clustering* of triphones
 - Expert-prepared contextual questions are used for clustering

- Initial estimates for monophones (context-independent HMMs)
 - based on reference segmentation
- Conversion to *plain* triphones (context-dependent models)
 - /o/ /s'/ /e/ /m/
 - /o+s'/ /o-s'+e/ /s'-e+m/ /e-m/
- Context-dependent *clustering* of triphones
 - Expert-prepared contextual questions are used for clustering
- Splitting of single Gaussians into multi-modal distributions

Baum-Welch reestimation procedure is routinely run after each step.

Experiment: stressed vowels as separate phonemes

- Polish phonetic alphabet consists of 39 phonemes
 - including 6 vowel sounds: /i/, /y/, /e/, /a/, /o/, /u/

Experiment: stressed vowels as separate phonemes

- Polish phonetic alphabet consists of 39 phonemes
 - including 6 vowel sounds: /i/, /y/, /e/, /a/, /o/, /u/
- Distinction between lexically stressed and unstressed instances is introduced
- Modification made on *dictionary* level only
 - no acoustic analysis of stress on training set or during recognition
 - accentuation rules are used

Experiment: stressed vowels as separate phonemes

- Polish phonetic alphabet consists of 39 phonemes
 - including 6 vowel sounds: /i/, /y/, /e/, /a/, /o/, /u/
- Distinction between lexically stressed and unstressed instances is introduced
- Modification made on *dictionary* level only
 - no acoustic analysis of stress on training set or during recognition
 - accentuation rules are used
- Further extension: sonorants (/m/, /n/, /n'/, /N/, /j/, /l/, /w/) placed within stressed syllables are also distinguished from those inside unstressed syllables

Experimental results

- ca. 50h used for experiments
- ca. 36k word-list
- 5-fold cross-validation

Setup	Num. phonemes	%Acc ¹	±
Standard	39	45.82	1.27
Stressed vowels	45	49.10	1.18
Vowels & sonorants	52	49.62	1.52

¹Accuracy is $A = \frac{H-I}{N}$, where H is number of correctly recognized words, I is the number of inserted words, N is total number of words in reference.

Speaker adaptation experiment

- Variability of speaker characteristics has negative impact on performance of speaker-independent recognition systems
- Speaker adaptation is recommended

Speaker adaptation experiment

- Variability of speaker characteristics has negative impact on performance of speaker-independent recognition systems
- Speaker adaptation is recommended
- Maximum-likelihood linear regression (MLLR) yields ca. 12.5% of accuracy boost

- Significant accuracy boost introduced by vowels' stress “distinction”
- Overall accuracy still below levels acceptable for real applications

- More advanced speech parametrization (LDA)
- Environmental noise and out-of-vocabulary words
- Dictionary of over 100k words (possibly over 1M words)
 - Polish has complex inflection rules
- High-order language modeling (trigrams or higher)
 - Polish has comparably flexible word-order
- Decoder speed optimization

Thank you.