Improved Modeling of Cross-Decoder Phone Co-occurrences in SVM-based Phonotactic Language Recognition

Mikel Penagarikano, Amparo Varona, Luis J. Rodríguez-Fuentes, Germán Bordel

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
email: mikel.penagarikano@ehu.es

Odyssey 2010, Brno, Czech Republic
July 1, 2010

Outline

Motivation

- Most common approaches to phonotactic language recognition deal with several independent phone decodings.
- These decodings are processed and scored in a fully uncoupled way and no cross-decoder dependencies are exploited for language modeling, information being fused only at the score level.
- Certain sounds from languages not covered by (not matching) the decoders may be better represented by cross-decoder outputs.

Background

- Cross-stream (cross-decoder) information previously applied for speaker recognition in the Johns Hopkins University (JHU) 2002 Workshop, where two decoupled time and cross-stream systems were integrated at the score level.

  Q. Jin, J. Navratil, D.A. Reynolds, J.P. Campbell, W.D. Andrews, and J.S. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition", in Proceedings of ICASSP, 2003, pp. 800-803.

Background

- Cross-stream (cross-decoder) information previously applied for speaker recognition in the Johns Hopkins University (JHU) 2002 Workshop, where two decoupled time and cross-stream systems were integrated at the score level.

  Q. Jin, J. Navratil, D.A. Reynolds, J.P. Campbell, W.D. Andrews, and J.S. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition", in Proceedings of ICASSP, 2003, pp. 800-803.

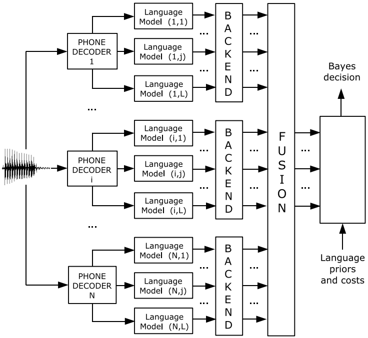
- Some years later, cross-stream dependencies were also used via multi-string alignments in a language recognition application

  Christopher White, Izhak Shafran, and Jean-Luc Gauvain, "Discriminative classifiers for language recognition", in Proceedings of ICASSP, 2006, pp. 213-216.

Architecture

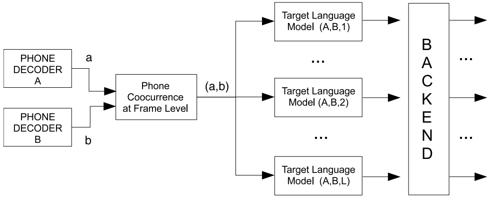
Common approach to phonotactic language recognition:

N Phone Decoders + L SVM-based Language Models + Gaussian Backend + Linear Fusion



Introduction

- Exploit cross-decoder dependencies using time-synchronous (frame level) phone co-occurrences.
- In a two decoder scenario:

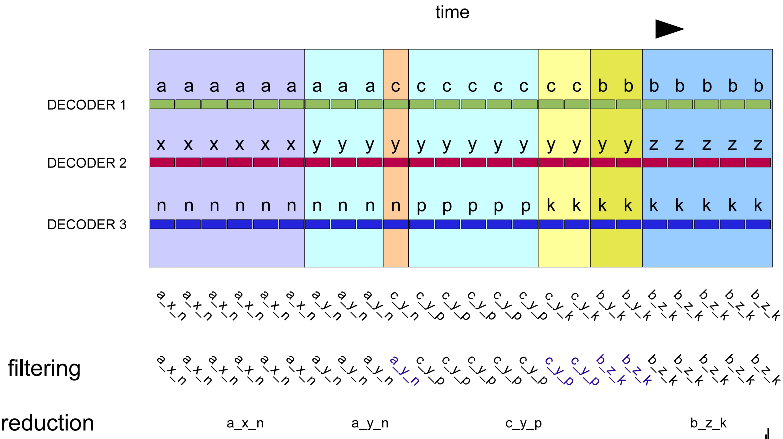


- In a D-decoder scenario:
  - Build a single D-phone co-occurrence system
  - Build $D!/k!(D-k)!$ k-phone co-occurrence systems

Approach 1: n-grams of phone co-occurrences

- First introduced in *Penagarikano et al., ICASSP 2010.*
- Get a frame-synchronous sequence of multi-phone (k-phone co-occurrence)
- Two type of sequence segments can be identified
  - *Stationary segments*: relatively long portions of speech for which decoders keep the same labels
  - *Transitional segments*: mainly appearing at phone borders (cross-decoder desynchronization)
- Transitional segments are removed and stationary segments are collapsed.

Approach 1: n-grams of phone co-occurrences



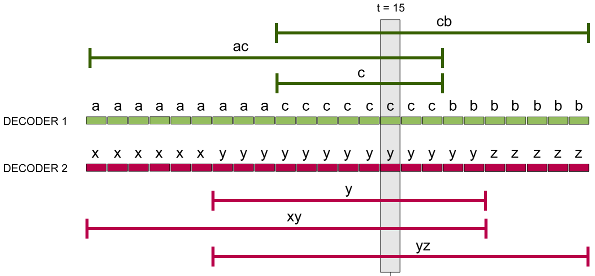
Approach 1: n-grams of phone co-occurrences

- Standard phonotactic approach is performed on the resulting k-phone sequence.
- ... not so standard
  - Number of different k-phones (1-grams): 2500 ($k = 2$), 124000 ($k = 3$)
  - The number of n-grams increases exponentially.
  - A full bag of n-grams strategy is infeasible.
- Only the most frequent n-gram counts are included in the supervector.

Approach 2: co-occurrences of phone n-grams

- In the previous approach, cross-decoder desynchronization affects the time modeling (n-grams)
- Exploit cross-decoder dependencies using time-synchronous (frame level) phone n-gram co-occurrences.
- Directly compute the n-gram co-occurrence counts from the decodings.
  - Each phone n-gram is counted once for each decoder, so its count is distributed among all the frames it spans.
  - The contribution corresponding to a given phone n-gram at a given frame is distributed among all the co-occurrences.
  - The sum of the counts of phone n-grams co-occurrences is equal to the average number of n-grams.
- Only the most frequent co-occurrence counts are included in the supervector.

Approach 2: co-occurrences of phone n-grams



$$count(c\_y) = \frac{1}{2}\left(\frac{1}{8\cdot 1} + \frac{1}{13\cdot 1}\right)$$

$$count(ac\_xy) = \frac{1}{2}\left(\frac{1}{17\cdot 2} + \frac{1}{19\cdot 2}\right)$$

$$count(ac\_yz) = \frac{1}{2}\left(\frac{1}{17\cdot 2} + \frac{1}{18\cdot 2}\right)$$

$$count(cb\_xy) = \frac{1}{2}\left(\frac{1}{15\cdot 2} + \frac{1}{19\cdot 2}\right)$$

$$count(cb\_yz) = \frac{1}{2}\left(\frac{1}{15\cdot 2} + \frac{1}{18\cdot 2}\right)$$

Training, development and test corpora

- Limited to those distributed by NIST to all LRE2007 participants
  - Call-Friend Corpus
  - OHSU Corpus provided by NIST for LRE05
  - development corpus provided by NIST for LRE07
- 10 conversations per language randomly selected for development purposes.
- Each development conversation was further split in segments containing 30 seconds of speech.
- Evaluation was carried out on the LRE07 evaluation corpus, specifically on the 30-second, closed-set condition.

Evaluation measures

- Most usual performance measures used in language recognition systems.
  - *DET* plots & *EER*: not providing calibration information.
  - C_{avg} & C_{min}: application dependent costs.

Evaluation measures

- Most usual performance measures used in language recognition systems.
  - *DET* plots & *EER*: not providing calibration information.
  - C_{avg} & C_{min}: application dependent costs.
- We prefer C_{llr} (more precisely, C_{mxe})
  - It is used as an alternative performance measure in NIST evaluations.
  - It evaluates the application independent system performance by means of a single numerical value (and appealing units: **bits**).
  - $\Delta = \log_2 N - C_{mxe}$ gives the effective amount of information that the recognizer delivers to the user, given no prior information.
  - The lower C_{mxe} is, the more informative our system is.

Evaluation measures

- Most usual performance measures used in language recognition systems.
  - *DET* plots & *EER*: not providing calibration information.
  - C_{avg} & C_{min}: application dependent costs.
- We prefer C_{llr} (more precisely, C_{mxe})
  - It is used as an alternative performance measure in NIST evaluations.
  - It evaluates the application independent system performance by means of a single numerical value (and appealing units: **bits**).
  - $\Delta = \log_2 N - C_{mxe}$ gives the effective amount of information that the recognizer delivers to the user, given no prior information.
  - The lower C_{mxe} is, the more informative our system is.
- However, we keept *DET* plots, *EER* and detection cost.

Software Components

Freely available software was used in all the stages

- **Phone Decoders:** The TRAPS/NN phone decoders developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU).
- **SVM modeling:** *LIBLINEAR* (a fast linear-only version of libSVM). Modified by adding some lines of code to get the regression values (instead of class labels).
- **Gaussian Backend & Fusion:** *FoCal Multi-class* toolkit by Niko Brummer.

Configuration

BUT TRAPS/NN CZ, HU & RU phone decoders

- Before doing phone tokenization, an energy-based voice activity detector is applied to split and remove non-speech segments.
- Non phonetic units (*int*, *pau* and *spk*) are mapped to silence (*sil*).
- Number of resulting phonemes: 43 (CZ), 59 (HU) and 49 (RU).
- 1-best decoding.

GTTS Tecnologías Software

Configuration

BUT TRAPS/NN CZ, HU & RU phone decoders

- Before doing phone tokenization, an energy-based voice activity detector is applied to split and remove non-speech segments.
- Non phonetic units (*int*, *pau* and *spk*) are mapped to silence (*sil*).
- Number of resulting phonemes: 43 (CZ), 59 (HU) and 49 (RU).
- 1-best decoding.

LIBLINEAR

- Phone sequences are modelled by means of Support Vector Machines
- SVM vectors consist of counts of phone n-grams (up to trigrams), converted to frequencies and weighted with regard to their background probabilities as $w_i = \min\left(C, \frac{1}{\sqrt{p(d_i|background)}}\right)$, with $C = 300$

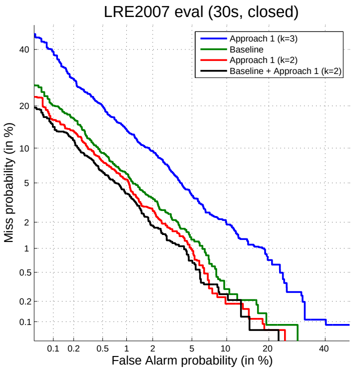
Single Systems Performance

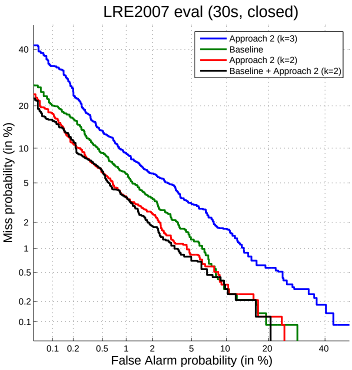
		EER	C_{LLR}
Baseline	CZ	5,67%	0,8259
	HU	5,10%	0,7434
	RU	5,64%	0,8016
	Fusion	2,69%	0,3981
Approach 1 (k=2)	CZ-HU	4,07%	0,5661
	CZ-RU	4,53%	0,6526
	HU-RU	3,79%	0,5109
	Fusion	2,27%	0,3393
Approach 1 (k=3)	CZ-HU-RU	4,34%	0,6500
Approach 2 (k=2)	CZ-HU	3,32%	0,4506
	CZ-RU	3,58%	0,5276
	HU-RU	2,75%	0,4140
	Fusion	2,24%	0,3223
Approach 2 (k=3)	CZ-HU-RU	3,90%	0,5724

Fused Systems Performance

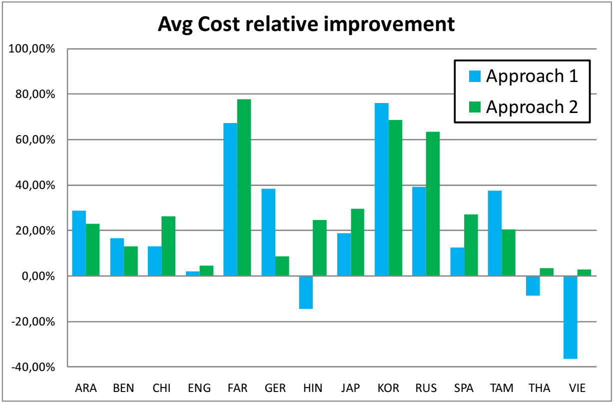
Fused Systems	EER	C_{LLR}
Baseline	2,69%	0,3981
A1 (k=2)	2,27%	0,3393
A2 (k=2)	2,24%	0,3223
A1 (k=3)	4,34%	0,6500
A2 (k=3)	3,90%	0,5724
A1 (k=2) + A1 (k=3)	2,21%	0,3388
A2 (k=2) + A2 (k=3)	2,28%	0,3280
Baseline + A1 (k=2)	**1,92%**	**0,3054**
Baseline + A2 (k=2)	**1,88%**	**0,3064**
Baseline + A1 (k=3)	2,38%	0,3472
Baseline + A2 (k=3)	2,15%	0,3582
Baseline + A1 (k=2) + A1 (k=3)	2,02%	0,3056
Baseline + A2 (k=2) + A2 (k=3)	1,90%	0,3158

DET plots


LRE2007 eval (30s, closed)
Approach 1 (k=3)
Baseline
Approach 1 (k=2)
Baseline + Approach 1 (k=2)
Miss probability (in %)
False Alarm probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40


LRE2007 eval (30s, closed)
Approach 2 (k=3)
Baseline
Approach 2 (k=2)
Baseline + Approach 2 (k=2)
Miss probability (in %)
False Alarm probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40

C_{avg} relative improvement per target language



Summary

- Two approaches to the modeling of cross-decoder phone co-occurrences in SVM-based Phonotactic Language Recognition have been proposed and evaluated.
- Both approaches outperformed the baseline system when using combinations of $k = 2$ decoders.
- Co-occurrence information is more effectively extracted in 2-decoder configurations and recovered by means of fusion.
- Under 3-decoder configuration, both approaches showed a poor performance compared to the baseline system. This may reveal robustness issues related to: the higher amount of transitional segments and the huge number of phone co-occurrence combinations.

Thank you!