

Using Cross-Decoder Co-Occurrences of Phone N-Grams in SVM-based Phonotactic Language Recognition

Mikel Penagarikano, Amparo Varona, Luis Javier Rodriguez-Fuentes, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain

mikel.penagarikano@ehu.es

Abstract

Most common approaches to phonotactic language recognition deal with several independent phone decoders. Decodings are processed and scored in a fully uncoupled way, their time alignment (and the information that may be extracted from it) being completely lost. Recently, we have presented a new approach to phonotactic language recognition which takes into account time alignment information, by considering cross-decoder co-occurrences of phones or phone n -grams at the frame level. Experiments on the NIST LRE2007 database demonstrated that using co-occurrence statistics could improve the performance of baseline phonotactic recognizers. In this work, the approach based on cross-decoder co-occurrences of phone n -grams is further developed and evaluated. Systems were built by means of open software (Brno University of Technology phone decoders, LIBLINEAR and *FoCal*) and experiments were carried out on the NIST LRE2007 database. A system based on co-occurrences of phone n -grams (up to 4-grams) outperformed the baseline phonotactic system, yielding around 8% relative improvement in terms of EER. The best fused system attained 1,90% EER (a 16% improvement with regard to the baseline system), which supports the use of cross-decoder dependencies for improved language modeling.

Index Terms: Phonotactic Language Recognition, Support Vector Machines, Cross-Decoder Co-occurrences

1. Introduction

Phonotactic language recognizers exploit the ability of phone decoders to convert a speech utterance into a sequence of symbols containing acoustic, phonetic and phonological information. Models for target languages are built by decoding hundreds or even thousands of training utterances and using the phone-sequence (or phone-lattice) statistics (typically, counts of n -grams) in different ways. Since training data include a wide range of speakers and diverse linguistic contents, being *language* the common factor, it is expected that phone statistics reflect language-specific characteristics.

Nowadays, the most common phonotactic approach uses counts of phone n -grams to build a feature vector which feeds a classifier based on Support Vector Machines (SVM) [1]. Typically, N phone decoders are applied in parallel to the input utterance, yielding N phone decodings (or lattices). The output

of the phone decoder i ($i \in [1, N]$) is scored for each target language j ($j \in [1, L]$), by applying the model $\lambda(i, j)$ (estimated using the outputs of the phone decoder i for the training database, taking j as the target language). Scores for the subsystem i are calibrated, typically by means of a Gaussian backend. Sometimes, a t -norm [2] is applied before calibration. Finally, $N \times L$ calibrated scores are fused applying discriminative linear logistic regression, to get L final scores for which a minimum expected cost Bayes decision is taken, according to application-dependent language priors and costs (see [3, 4] for details).

However, the above described structure defines N independent data processing channels, and no cross-decoder dependencies are exploited for language modeling, information being fused only at the score level. The idea of using phonetic information in the cross-stream (cross-decoder) dimension was first applied for speaker recognition in the Johns Hopkins University (JHU) 2002 Workshop [5], where two decoupled time and cross-stream dimensions were modelled separately and integrated at the score level. Some years later, cross-stream dependencies were also used via multi-string alignments in a language recognition application [6].

In a recent work, we have presented a simple approach to phonotactic language recognition which uses statistics of cross-decoder phone co-occurrences at the frame level [7]. Time stamps are extracted as side information from the 1-best phone decoding, so that each frame can be assigned k phone labels, for a combination of $k \leq N$ decoders (N : number of decoders). Finally, sequences of k -phone co-occurrences are used for modeling purposes. As for n -grams, the number of possible k -phone co-occurrences increases exponentially with k , so in practice only 2-phone and 3-phone co-occurrences are considered. In experiments on the NIST LRE2007 database, fusing baseline phonotactic systems with systems based on cross-decoder phone co-occurrences led to improved performance in all the cases (see [7] for details).

The approach described above was extended in [8], by considering counts of up to 3-grams (instead of just unigrams) of 2-phone and 3-phone co-occurrences in a SVM classifier. Additionally, a second approach was also introduced in [8], which did not consider n -grams of phone co-occurrences, but co-occurrences of phone n -grams (up to 3-grams). In this paper, we present the latest developments attained under this second approach, which uses statistics of co-occurrences of phone n -grams (up to 4-grams) in a SVM-based phonotactic language recognizer. Systems have been developed by means of open software (BUT phone decoders, LIBLINEAR and *FoCal*) and evaluation has been carried out on a relevant database (NIST LRE2007).

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

The rest of the paper is organized as follows. Section 2 presents the main features of the baseline phonotactic language recognition system used in this work. Section 3 describes the approach based on cross-decoder co-occurrences of phone n -grams. The experimental setup is briefly described in Section 4. Results obtained in language recognition experiments on the NIST LRE2007 database (pooled for all the target languages) are presented in Section 5. Finally, conclusions and future work are outlined in Section 6.

2. Baseline SVM-based Phonotactic Language Recognizer

As in [8], in this work a SVM-based phonotactic language recognizer is used as baseline system, and the NIST LRE2007 database is used for development and evaluation. Systems have been built by means of open software. In particular, the TRAPS/NN phone decoders for Czech (CZ), Hungarian (HU) and Russian (RU) developed by the Brno University of Technology (BUT) [9] are the core elements of all the systems in this work. Each BUT decoder takes a speech signal as input, runs an acoustic front-end, applies a set of acoustic models and gives the 1-best phone decoding as output. Non phonetic units (*int*, *pau* and *spk*) are mapped to silence (*sil*), so that output dimensions for BUT decoders are 43 (CZ), 59 (HU) and 49 (RU), respectively. Before phone tokenization, an energy-based voice activity detector is applied to split and remove non-speech segments from the signals. Regarding channel compensation, noise reduction, etc. all the systems presented in this paper rely on the acoustic front-end embedded in BUT decoders.

In the baseline system, phone sequences are modelled by means of SVM. SVM vectors consist of counts of phone n -grams (up to trigrams), weighted as proposed in [10]. A Cramer and Singer solver for multiclass SVMs with linear kernels has been applied, by means of LIBLINEAR [11] (adding some lines of code to retrieve regression values). Final scores are computed by fusing the scores of three calibrated SVM-based phonotactic sub-systems, for Czech, Hungarian and Russian decoders. The *FoCal* toolkit is used for calibration and fusion [3, 4].

3. Co-occurrences of phone n -grams

The approach presented in this paper is based on computing and using statistics of cross-decoder co-occurrences of phone n -grams. For any given decoder, up to n n -grams can overlap at each frame t , which means that up to n^k phone n -grams can co-occur at the same frame for a choice of k decoders. So, a procedure must be designed for distributing co-occurrence counts at frame level. This procedure will allow us to circumvent the issue of lack of synchronization among decoders at phone borders. In this work, we consider only cross-decoder co-occurrences of n -grams with the same n . Though possible, mixed co-occurrences (unigrams with bigrams, bigrams with trigrams, etc.) are not considered.

Let consider an input sequence of feature vectors $X = (X_1, \dots, X_T)$ and a choice of k decoders $\pi = (d_1, \dots, d_k)$. Let $\Gamma_n^{(d)}(t)$ be the set of n -grams overlapping at frame t in decoding d . Let $w_n^{(d)}(t, i)$ be one of such n -grams and $f_n^{(d)}(t, i)$ the number of frames it spans, with $i \in [1, |\Gamma_n^{(d)}(t)|]$. Note that $|\Gamma_n^{(d)}(t)| = n$ for all t except for a number of frames at the borders of X , where $1 \leq |\Gamma_n^{(d)}(t)| < n$. Let $c_n^\pi(t, \nu) = (w_n^{d_1}(t, i_1), \dots, w_n^{d_k}(t, i_k))$ be a co-occurrence of k phone n -grams, for a choice of n -grams $\nu = (i_1, \dots, i_k)$, with $1 \leq i_j \leq |\Gamma_n^{(d_j)}(t)|$, for $j \in [1, k]$.

In this approach, each decoder $d_j \in \pi$ makes its own contribution to the count of a given co-occurrence of phone n -grams at a given frame. The key concepts are: (1) each phone n -gram is counted once for each decoder, so its count is distributed among all the frames it spans; and (2) the contribution corresponding to a given phone n -gram at a given frame for a given decoder is distributed among all the combinations of phone n -grams at that frame for the remaining decoders. Taking into account these principles, we get the following expression:

$$count(c_n^\pi(t, \nu), d_j) = \frac{1}{f_n^{(d_j)}(t, i_j) \cdot \prod_{\substack{l=1 \\ l \neq j}}^k |\Gamma_n^{(d_l)}(t)|} \quad (1)$$

The count for $c_n^\pi(t, \nu)$ is computed as the average contribution over all the decoders:

$$count(c_n^\pi(t, \nu)) = \frac{1}{k} \sum_{j=1}^k count(c_n^\pi(t, \nu), d_j) \quad (2)$$

Finally, the count corresponding to a given co-occurrence of phone n -grams $b_n^\pi = (v_n^{(d_1)}, \dots, v_n^{(d_k)})$ is computed by adding the counts for all the frames in the sequence where it appears:

$$count(b_n^\pi) = \sum_{t=1}^T \sum_{\forall \nu} \delta(b_n^\pi, c_n^\pi(t, \nu)) \cdot count(c_n^\pi(t, \nu)) \quad (3)$$

In practice, counts are computed in two passes. The first pass computes and stores $|\Gamma_n^{(d)}(t)|$ and $f_n^{(d)}(t, i)$ for each decoder d and each frame t . Starting from the previously stored values, the second pass accumulates the counts of phone n -grams on a frame-by-frame basis, applying equation 2 for each combination ν of phone n -grams appearing at frame t .

In this work, we consider cross-decoder co-occurrences of unigrams, bigrams, 3-grams and 4-grams. An example for $k = 2$ decoders ($\pi = (1, 2)$) including up to bigrams, is shown in Figure 1. Let consider the shaded frame ($t = 15$) in Figure 1. The sets of n -grams appearing at that frame are:

$$\begin{aligned} \Gamma_1^{(1)}(15) &= \{c\} & \Gamma_2^{(1)}(15) &= \{ac, cb\} \\ \Gamma_1^{(2)}(15) &= \{y\} & \Gamma_2^{(2)}(15) &= \{xy, yz\} \end{aligned}$$

and the number of frames they span:

$$\begin{aligned} f_1^{(1)}(15, 1) &= 8 & f_2^{(1)}(15, 1) &= 17 \\ f_1^{(2)}(15, 1) &= 13 & f_2^{(1)}(15, 2) &= 15 \\ & & f_2^{(2)}(15, 1) &= 19 \\ & & f_2^{(2)}(15, 2) &= 18 \end{aligned}$$

Starting from these values and according to equation 2, the counts of co-occurrences of phone n -grams are computed as

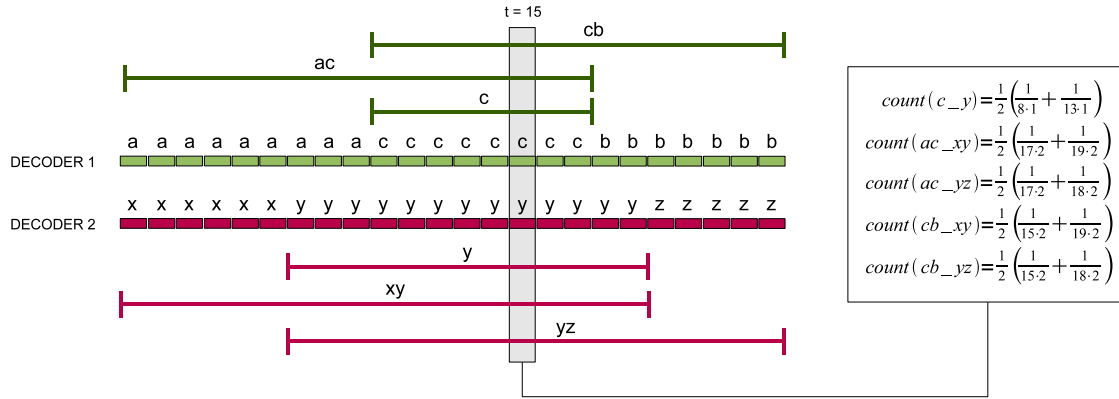


Figure 1: Co-occurrences of phone n -grams (2-decoder configuration, up to bigrams): (1) each n -gram is counted once for each decoder, so its count is distributed among all the frames it spans; (2) the contribution corresponding to a given n -gram at a given frame for a given decoder is distributed among all the combinations of n -grams appearing at that frame for the remaining decoders; and (3) the count corresponding to a given n -gram at a given frame is computed as the average contribution over all decoders.

follows:

$$\begin{aligned} \text{count}(c_y) &= \frac{1}{2} \cdot \left(\frac{1}{8 \cdot 1} + \frac{1}{13 \cdot 1} \right) \\ \text{count}(ac_xy) &= \frac{1}{2} \cdot \left(\frac{1}{17 \cdot 2} + \frac{1}{19 \cdot 2} \right) \\ \text{count}(ac_yz) &= \frac{1}{2} \cdot \left(\frac{1}{17 \cdot 2} + \frac{1}{18 \cdot 2} \right) \\ \text{count}(cb_xy) &= \frac{1}{2} \cdot \left(\frac{1}{15 \cdot 2} + \frac{1}{19 \cdot 2} \right) \\ \text{count}(cb_yz) &= \frac{1}{2} \cdot \left(\frac{1}{15 \cdot 2} + \frac{1}{18 \cdot 2} \right) \end{aligned}$$

For estimating the SVMs corresponding to target languages, counts computed this way are accumulated for a training database, SVM vectors being built with the M highest counts ($M = 200000$ in this work). Note that counts of co-occurrences of unigrams, bigrams, 3-grams and 4-grams are put together in a single representation, which includes information from both time (phone n -grams) and cross-stream (co-occurrence) dimensions.

For scoring purposes, given an input sample X , we first obtain 1-best decodings and segmentations, then count phone n -gram co-occurrences and use them to build an M -dimensional vector. Finally, this vector is scored with regard to SVMs. Note that, since a sparse representation is used, co-occurrences not appearing among those with the M highest counts in the training database are not used for scoring.

4. Experimental Setup

4.1. Training, development and test corpora

Training and development data were limited to those distributed by NIST to all LRE2007 participants: (1) the Call-Friend Corpus; (2) the OHSU Corpus provided by NIST for LRE05; and (3) the development corpus provided by NIST for LRE07. For development purposes, 10 conversations per language were randomly selected, the remaining conversations being used for training. 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 (primary evaluation task for the LRE07).

4.2. Evaluation measures

Most authors compare the performance of language recognition systems either globally (but not numerically) by means of Detection Error Tradeoff (DET) plots, or numerically (but not globally, and not at the optimal operation point) by means of Equal Error Rates (EER). In this work, systems will be also compared in terms of the so called C_{LLR} [12], which is used as an alternative performance measure in NIST evaluations. We internally consider C_{LLR} as the most relevant performance indicator, for two reasons: (1) C_{LLR} allows us to evaluate system performance globally by means of a single numerical value, which is somehow related to the area below the DET curve, provided that scores can be interpreted as log-likelihood ratios; and (2) C_{LLR} does not depend on application costs; instead, it depends on the calibration of scores, an important feature of detection systems.

5. Results

Table 1 shows EER and C_{LLR} performance in language recognition experiments on the LRE2007 database applying the baseline phonotactic system and a system using statistics of 2-decoder co-occurrences of phone n -grams (up to 4-grams). For the sake of completeness, the performance of subsystems (1-decoder configurations for the baseline system and 2-decoder configurations for the proposed approach) and partial fusions is also shown in Table 1, rows corresponding to final (fused) systems being shown in boldface.

The system using statistics of 2-decoder co-occurrences of phone n -grams yielded 2,08% EER (64 misses and 831 false alarms) and $C_{LLR} = 0,3083$, which means improvements of around 8% and 12%, respectively, with regard to the baseline system, which yielded 2,26% EER (69 misses and 903 false alarms) and $C_{LLR} = 0,3496$.

Regarding subsystems, note that 2-decoder co-occurrence subsystems performed consistently better than 1-decoder baseline subsystems, the co-occurrence subsystem HU-RU yielding best results. On the other hand, fusing two 1-decoder subsystems yielded better results than the corresponding 2-decoder co-occurrence subsystems, but the fusion of the latter performed better than the fusion of the former. Finally, as shown in Figure 2, the fusion of the baseline system and the system using co-occurrences of phone n -grams attained 1,90% EER (58 misses

and 759 false alarms), which means around 16% EER improvement and reveals that co-occurrence subsystems provide information that is not present in the baseline system.

Table 1: Performance (EER and C_{LLR}) of: (1) the baseline phonotactic system; (2) a system using statistics of 2-decoder co-occurrences of phone n -grams (up to 4-grams); and the fusion of (1) and (2).

		EER	C_{LLR}
Baseline	CZ	5,13%	0,7289
	HU	4,67%	0,6560
	RU	4,64%	0,6939
	Fusion (CZ,HU)	3,06%	0,4298
	Fusion (CZ,RU)	3,13%	0,4830
	Fusion (HU,RU)	2,69%	0,3891
	(1) Fusion (all)	2,26%	0,3496
2-decoder co-occurrences of phone n -grams	CZ-HU	3,06%	0,4363
	CZ-RU	3,62%	0,5380
	HU-RU	2,92%	0,4355
	(2) Fusion (all)	2,08%	0,3083
(1) + (2)	Fusion	1,90%	0,2943

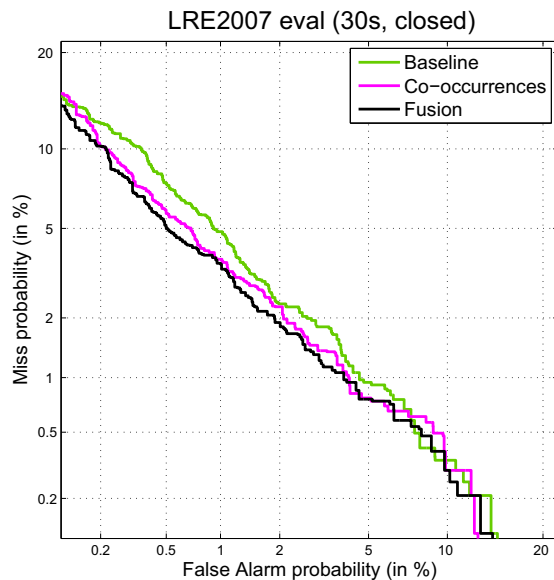


Figure 2: Pooled DET curves for: (1) the baseline phonotactic system; (2) a system using statistics of 2-decoder co-occurrences of phone n -grams (up to 4-grams); and the fusion of (1) and (2).

6. Conclusions

In this paper, latest developments under an approach using cross-decoder co-occurrences of phone n -grams in SVM-based phonotactic language recognition have been presented and evaluated. The proposed approach relies on the assumption that cross-decoder co-occurrence information is somehow specific to each target language. The approach does not involve significant additional computation with regard to a baseline phonotactic system. It represents just a means to extract more information from existing decodings.

A system using statistics of 2-decoder co-occurrences of phone of n -grams (up to 4-grams) outperformed the baseline system in language recognition experiments on the LRE2007

database. Fusing the baseline system and the system using co-occurrences of phone n -grams led to best performance: 1,90% EER and $C_{LLR} = 0,2943$ (around 16% relative improvement in both cases).

We are currently working on various co-occurrence selection schemes, with the aim to reduce the size of SVM vectors while keeping or even improving performance. Future work will focus on increasing the robustness of phonotactic approaches that integrate time and cross-stream dependencies.

7. References

- [1] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2–3, pp. 210–229, 2006.
- [2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, January 2000.
- [3] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *Proc. Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.
- [4] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. on ASLP*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [5] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in *Proceedings of ICASSP*, vol. 4, 2003, pp. 800–803.
- [6] C. White, I. Shafran, and J.-L. Gauvain, "Discriminative classifiers for language recognition," in *Proceedings of ICASSP*, 2006, pp. 213–216.
- [7] M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "Using cross-decoder phone co-occurrences in phonotactic language recognition," in *Proceedings of ICASSP*, Dallas, Texas (USA), 2010, pp. 5034–5037.
- [8] M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "Improved modeling of cross-decoder phone co-occurrences in SVM-based phonotactic language recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.
- [9] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology BUT, <http://www.fit.vutbr.cz>, Brno, CZ, 2008.
- [10] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Proceedings of ICASSP*, 2008, pp. 4145–4148.
- [11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, software available at <http://www.csie.ntu.edu.tw/~cjlin/liblinear>.
- [12] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, 2006.