USING CROSS-DECODER PHONE COOCURRENCES IN PHONOTACTIC LANGUAGE RECOGNITION

Mikel Penagarikano, Amparo Varona, Luis Javier Rodríguez-Fuentes, Germán Bordel

GTTS, Department of Electricity and Electronics, University of the Basque Country, Spain
e-mail: mikel.penagarikano@ehu.es

ABSTRACT

Phonotactic language recognizers are based on the ability of phone decoders to produce phone sequences containing acoustic, phonetic and phonological information, which is partially dependent on the language. Input utterances are decoded and then scored by means of models for the target languages. Commonly, various decoders are applied in parallel and fused at the score level. A kind of complementarity effect is expected when fusing scores, since each decoder is assumed to extract different (and complementary) information from the input utterance. This assumption is supported by the performance improvements attained when fusing systems. However, decodings are processed in a fully uncoupled way, their time alignment (and the information that may be extracted from it) being completely lost. In this paper, a simple approach is proposed, which takes into account time alignment information, by considering cross-decoder phone coocurrences at the frame level. To evaluate the approach, a choice of open software (BUT front-end and phone decoders, SRI-LM toolkit, libSVM, FoCal) is used, and experiments are carried out on the NIST LRE2007 database. Adding phone coocurrences to the baseline phonotactic systems provides slight performance improvements, revealing the potential benefit of using cross-decoder dependencies for language modeling.

Index Terms— Language Recognition, Phone Decoding, Phone Coocurrence

1. INTRODUCTION

Phonotactic language recognizers exploit the ability of phone decoders to convert a speech utterance into a sequence of phones 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 (unigrams, bigrams, etc.) in different ways. Since training data feature a wide range of speakers and diverse linguistic contents, being *language* the common factor, it is expected that phone statistics reflect language-specific characteristics.

The most common approaches are the so called PPRLM (Parallel Phone Recognizers followed by Language Models) [1] (referred to as Phone-LM in this paper) and the Phone-SVM (Support Vector Machines applied on phone n-gram counts) [2]. In both cases, N phone decoders are applied 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 subset of the training database corresponding to language j). Scores for the subsystem i are calibrated, typically by means of a Gaussian backend. Sometimes, a t-norm [3] is applied before calibration. Finally, $N \times L$ calibrated scores are fused applying 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 [4, 5] for details). Figure 1 shows the structure of a phonotactic language recognizer.

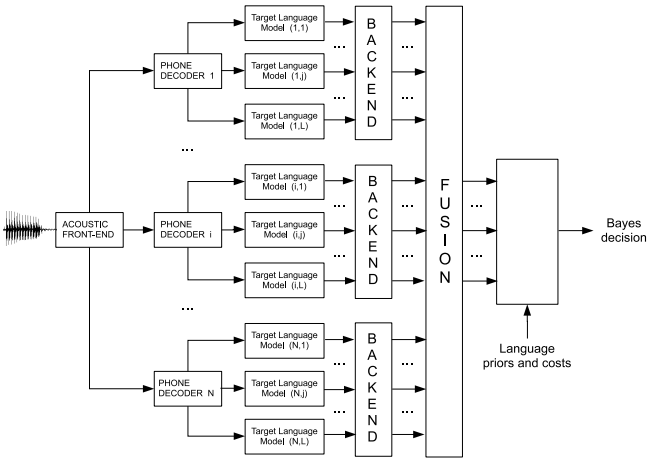

Fig. 1. A phonotactic language recognition (LR) system.

However, the above described structure defines N independent data processing channels, and no cross-decoder dependencies are exploited for language modeling, information

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).

being fused only at the score level. In this paper, a simple approach is proposed which takes into account cross-decoder phone coocurrences at the frame level. Time stamps are extracted as side information from the 1-best phone decoding. This way, each frame may be assigned N phone labels (one per decoder), and a sequence of multi-phones could be defined and used for modelling purposes. The simplest case would consist of sequences of two-phone labels corresponding to two decoders A and B, which could be processed and modelled exactly the same way as phone sequences (see Figure 2). In fact, there could be $N(N-1)/2$ of such 2-decoder subsystems. This configuration can be easily generalized to k-decoder subsystems, being the most general approach considering a single N-decoder system producing N-phone coocurrence sequences. As for n-grams, the number of k-phone coocurrences increases exponentially with k, so in practice it will only make sense to model 2-phone and 3-phone coocurrences.

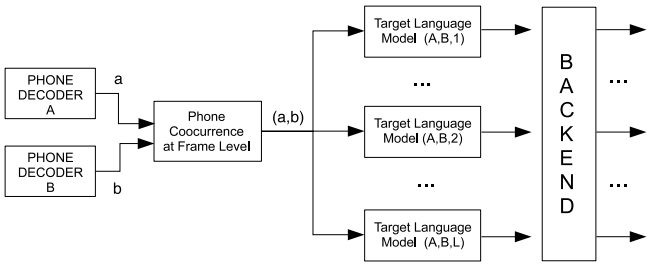

Fig. 2. A 2-decoder phone coocurrence LR subsystem.

The idea of using phonetic information in the cross-stream (cross-decoder) dimension was first applied for speaker recognition in the JHU 2002 Workshop [6]. However, the approach proposed in [6] considered two decoupled time and cross-stream dimensions, which were modelled separately and integrated at the score level.

The rest of the paper is organized as follows. The baseline phonotactic systems used in this work and the cross-decoder phone coocurrence approach are described in Sections 2 and 3, respectively. The experimental setup is described in Section 4. Results of language recognition experiments on the NIST LRE2007 database (pooled for all the target languages) are presented and discussed in Section 5. Finally, conclusions and some ideas for future work are outlined in Section 6.

2. BASELINE PHONOTACTIC SYSTEMS

In this work, we aimed to test the proposed approach with regard to a well-established methodology (phonotactic systems) and a relevant corpus of data (NIST LRE2007 database). We also aimed to allow other researchers to easily verify our results. So we made a choice of open software resources to build two phonotactic systems. Three phone decoders developed by the Brno University of Technology (BUT) for Czech, Hungarian and Russian [7] are the core elements of the baseline systems. BUT decoders have demonstrated a high accuracy and have been previously used by other groups (besides BUT [8], the MIT Lincoln Laboratory [9]) as the backend for phonotactic language recognition. Since each BUT decoder runs an acoustic front-end, it can be seen as a black box which takes a speech signal as input and gives the 1-best phone decoding as output. Regarding channel compensation, noise reduction, etc. all the systems presented in this paper rely on the acoustic front-end provided by BUT decoders.

Two baseline systems were built matching the structure shown in Figure 1. The Phone-LM system scored phone decodings with language models estimated by means of the SRI Language Model toolkit [10]. The Phone-SVM system was based on bag-of-N-grams vectors that were weighted as proposed in [11]. SVMs were trained by means of libSVM [12], applying a linear kernel, and using the log-likelihoods provided by the package (instead of the scores). In fact, both the Phone-LM and Phone-SVM systems were built by fusing three sub-systems, for Czech, Hungarian and Russian decoders.

3. USING CROSS-DECODER PHONE COOCURRENCES

Given an input sequence of feature vectors $X = (X_1, \ldots, X_T)$, T being the length of X, assuming that N phone decoders are available, consider the 1-best phone decodings: $D^{(i)}(X) = \{d_1^{(i)}, \ldots, d_{K(i)}^{(i)}\}$, $i \in [1, N]$, $K(i)$ being the length of $D^{(i)}(X)$, and segmentations (defined by considering phone labels at frame level): $S^{(i)}(X) = \{s_1^{(i)}, \ldots, s_T^{(i)}\}$, $i \in [1, N]$. Language modeling should consider not only intra-decoder phone (time) dependencies (phone n-grams: $p(d_t^{(i)}|d_1^{(i)}, \ldots, d_{t-1}^{(i)})$), but also cross-decoder time-synchronous (frame level) phone dependencies, which we call *coocurrences*: $p(s_t^{(i)}|s_t^{(1)}, \ldots, s_t^{(i-1)}, s_t^{(i+1)}, \ldots, s_t^{(N)})$), because language-specific information may be also extracted from these latter. In fact, the most general approach to language modeling should combine both dependencies into one single model, so that $s_t^{(i)}$ would depend on all the $s_k^{(j)}$, with $k \in [1, t-1]$ and $j \in [1, N]$, $j \neq i$. However, taking into account that each phone decoder handles between 30 and 60 phones, the resulting model would be too complex and the number of parameters too large, making it unfeasible computing robust estimations.

It is even unfeasible to model time-synchronous phone dependencies for a high number of decoders. Here we propose a simple approach where a restricted model is defined by taking into account time-synchronous phone dependencies for a choice of k decoders (out of N). There can be defined $N!/k!(N-k)!$ of such models, which could be estimated and applied on an independent way, and their scores fused with those of other models.

In this work, a sequence of frame-level phone coocurrences is built for each combination of $k = 2$ and $k = 3$ decoders (with $N = 3$). A sequence of coocurrences can be used the same way as a sequence of phones, either to estimate n-gram language models (like in Phone-LM systems) or to compose a bag-of-N-grams vector (like in Phone-SVM systems). This way, the resulting models include information from both time and cross-stream dimensions.

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. Phone decoders

The phonotactic language recognition systems used in this work were all based on the Brno University of Technology (BUT) TRAPS/NN decoders for Czech, Hungarian and Russian [7]. Non phonetic units appearing in the decodings (*int*, *pau* and *spk*) were mapped to silence (*sil*). Prior to phone tokenization, an energy-based voice activity detector was applied to split and remove non-speech segments from the signals. Phone decoder outputs were used in three different ways:

PHONE: 3 $\times$ phone outputs (CZ, HU and RU). Output dimensions were respectively 43, 59 and 49.

COOC2: 3 $\times$ frame-level 2-phone coocurrence outputs (CZ_HU, CZ_RU and HU_RU). Output dimensions were respectively 2537, 2107 and 2891.

COOC3: 1 $\times$ frame-level 3-phone coocurrence output (CZ_HU_RU). Output dimension was 124313.

4.3. Language models

Two different language modeling techniques were compared: (1) 4-gram language models (LM) with Witten-Bell smoothing; and (2) Support Vector Machine-based language models (SVM), using weighted bag-of-N-grams vectors. In the SVM case, for *PHONE* outputs, up to 3-grams where estimated, whereas for *COOC2* and *COOC3* outputs, only 1-gram statistics where estimated, due to the high number of coocurrences, which made unfeasible even the use of 2-grams.

5. RESULTS

Tables 1 and 2 show the Equal Error Rate (EER) performance of single systems and various system combinations (fusions), respectively. Although the performance of the best 2-phone coocurrence system (*COOC2-LM*) was far from those of the baseline systems, improvements were obtained when fused with any of them, yielding a relative 11% improvement with regard to *PHONE-LM* performance, and a relative 23% improvement with regard to *PHONE-SVM* performance (see Figure 3). The *COOC2-LM* system contributed useful information even when fused with the high-accuracy *PHONE-LM + PHONE-SVM* system, yielding a 7% relative improvement (see Figure 4). On the other hand, *COOC3* systems yielded worse performance than *COOC2* systems, specially in the case of the LM (n-gram) approach (see Table 1). Finally, though contributing useful information when fused with the baseline phonotactic systems, COOC3 did not outperform COOC2 (see Table 2). These results could be explained by the high number of 3-phone coocurrences and the limited amount of training data, which may be leading to unreliable estimates. To overcome this, a kind of selection strategy could be defined which would allow to apply only statistically reliable and/or discriminant 3-phone coocurrences.

Table 1. Performance (EER) of baseline phonotactic and coocurrence-based LR systems.

	EER	
	LM	SVM
PHONE-CZ	8,31%	5,87%
PHONE-HU	5,88%	5,49%
PHONE-RU	6,71%	5,85%
PHONE	2,96%	2,71%
COOC2	3,62%	4,84%
COOC3	11,26%	5,63%

Table 2. Performance (EER) of various fused LR systems.

Fusions	EER
PHONE-LM + PHONE-SVM	2,15%
COOC2-LM + COOC3-SVM	3,06%
PHONE-LM + COOC2-LM	2,39%
PHONE-LM + COOC3-SVM	2,59%
PHONE-SVM + COOC2-LM	2,00%
PHONE-SVM + COOC3-SVM	2,40%
PHONE-LM + PHONE-SVM + COOC2-LM	1,95%
PHONE-LM + PHONE-SVM + COOC2-LM + COOC3-SVM	1,99%

6. CONCLUSIONS AND FUTURE WORK

A simple approach to phonotactic language recognition, which takes into account cross-decoder phone information, has been proposed and evaluated. Adding phone coocur-

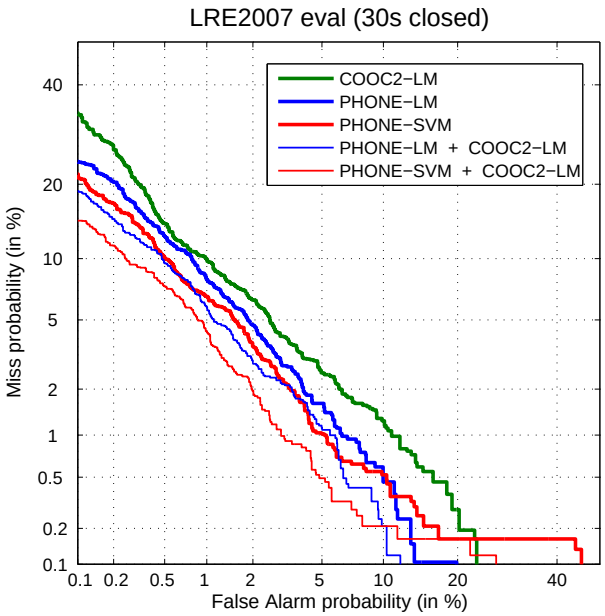

Fig. 3. Pooled DET curves for the baseline phonotactic language recognition systems (PHONE-LM and PHONE-SVM), the coocurrence approach COOC2-LM and the fused systems PHONE-LM+COOC2-LM and PHONE-SVM+COOC2-LM.

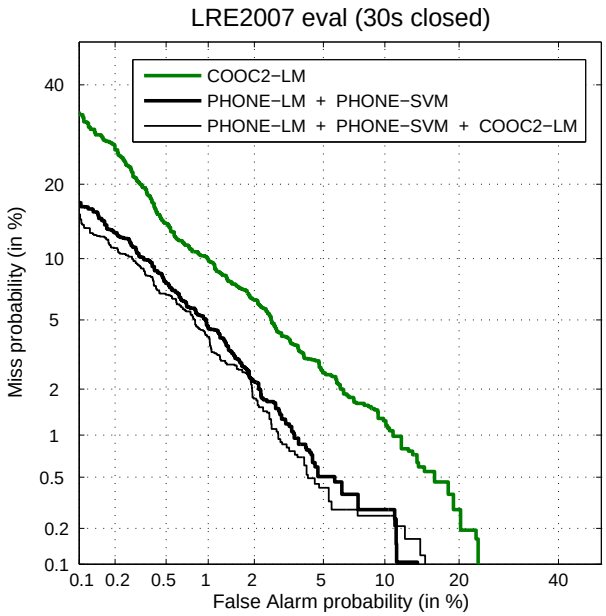

Fig. 4. Pooled DET curves for the fusion of baseline phonotactic systems, the coocurrence approach COOC2-LM and the fusion of all of them.

rences to the baseline phonotactic systems provides slight performance improvements, revealing the potential benefit of using cross-decoder dependencies for language modeling. On the other hand, this approach does not involve hard computations. It is just a way of extracting more information from existing decodings. We are currently working on a coocurrence selection scheme which allow the use of bigram and trigram counts in the COOC-SVM approach. Future work includes the design and evaluation of a more general phonotactic approach integrating time and cross-stream dependencies.

7. REFERENCES

[1] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, January 1996.

[2] 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.

[3] 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.

[4] N. Brummer and D.A. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[5] N. Brummer, 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[6] 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 *ICASSP*, 2003, vol. IV, pp. 800–803.

[7] Petr Schwarz, *Phoneme recognition based on long temporal context*, Ph.D. thesis, Faculty of Information Technology BUT, Brno, CZ, 2008.

[8] P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, T. Mikolov, and O. Plchot, "BUT system description for NIST LRE 2007," in *Proc. 2007 NIST Language Recognition Evaluation Workshop*, Orlando, US, 2007, pp. 1–5, National Institute of Standards and Technology.

[9] P.A. Torres-Carrasquillo, E. Singer, W.M. Campbell, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, W. Shen, and D.E. Sturim, "The MITLL NIST LRE 2007 language recognition system," in *Interspeech*, 2008, pp. 719–722.

[10] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Intl. Conf. on Spoken Language Processing*, November 2002, pp. 257–286.

[11] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *ICASSP*, 2008, pp. 4145–4148.

[12] C.C. Chang and C.J. Lin, *"http://www.csie.ntu.edu.tw/∼cjlin/libsvm", LIBSVM: a library for support vector machines*, 2001.