Improved Language Recognition Using Better Phonetic Decoders and Fusion with MFCC and SDC Features

Doroteo T. Toledano, Javier Gonzalez-Dominguez, Alejandro Abejon-Gonzalez, Danilo Spada, Ismael Mateos-Garcia and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autónoma de Madrid, Spain

{javier.gonzalez, doroteo.torre, joaquin.gonzalez}@uam.es

Abstract

One of the most popular and better performing approaches to language recognition (LR) is Parallel Phonetic Recognition followed by Language Modeling (PPRLM). In this paper we report several improvements in our PPRLM system that allowed us to move from an Equal Error Rate (EER) of over 15% to less than 8% on NIST LR Evaluation 2005 data still using a standard PPRLM system. The most successful improvement was the retraining of the phonetic decoders on larger and more appropriate corpora. We have also developed a new system based on Support Vector Machines (SVMs) that uses as features both Mel Frequency Cepstral Coefficients (MFCCs) and Shifted Delta Cepstra (SDC). This new SVM system alone gives an EER of 10.5% on NIST LRE 2005 data. Fusing our PPRLM system and the new SVM system we achieve an EER of 5.43% on NIST LRE 2005 data, a relative reduction of almost 66% from our baseline system.

Index Terms: Language recognition, PPRLM, SVM.

1. Introduction

Automatic Language Recognition (LR) tries to recognize the language of a particular speech segment and is usually a first step for further processing the speech segment either manually (sending the speech segment to an operator proficient in the language) or automatically (sending it to an adequate automatic dialogue manager). The last years have shown an important growth in the field, resulting in a rise in the number of sites participating in the LR evaluations organized by NIST [1].

Along the evolution of automatic LR the most widely used and successful approach to LR has been Phone Recognition followed by Language Modeling (PRLM) and Parallel PRLM (PPRLM) [2, 3]. More recently PPRLM systems have been improved further by processing the whole lattice instead of just the 1-best solution produced by the phonetic decoders [4, 5] and substituting the statistical language modeling scoring by Support Vector Machines (SVMs) taking as input vectors the n-grams [6]. In this paper we will not take into account these possibilities for improvement. Rather we will concentrate on *classical* PPRLM systems and try to improve their performance as much as possible as a first step to then make further improvements using lattice decoding and SVMs. In the process we will analyze the influence on LR results of several improvements over the baseline system [7] we submitted to NIST LRE 2005.

PPRLM systems can be complemented with other types of systems possibly operating on different features. In this paper we complement our improved PPRLM system with an SVM system operating on MFCC and SDC acoustic features.

In section 2 we describe our baseline system presenting results on data taken from NIST LRE 2003. The following sections (3, 4 and 5) will analyze the influence on PPRLM performance of the use of a different parameterization, an explicit Voice Activity Detector (VAD) and phonetic models trained on larger and more appropriate corpora. Section 6 briefly describes our new acoustic SVM system and section 7 presents results of the fusion of our PPRLM and SVM systems. Finally, section 7 presents conclusions.

2. Baseline System

Our starting point for this paper is the two PPRLM systems we submitted to NIST LRE 2005. These systems used 6 (ATVS2) or 12 (ATVS1) phonetic decoders trained on the OGI Multi-Language Telephone Speech Corpus [8] which contains roughly 1-2 hours of speech by language. These decoders are based on Hidden Markov Models (HMMs) and implemented using HTK [9]. The phonetic HMMs are three-state left-to-right models with no skips, being the output *pdf* of each state modeled as a weighted mixture of Gaussians. In ATVS2 we used 10 Gaussians per state, while in ATVS1 we used 10 and 20 Gaussians per state to have two phonetic decoders with different complexities for each language. The acoustic processing uses the Advanced Distributed Speech Recognition Standard Front-End [10], based on 12 Mel Frequency Cepstral Coefficients (MFCCs) plus a combination of energy and C0 and velocities and accelerations for a total of 39 components, computing a feature vector each 10ms. It also includes mechanisms for robustness against channel distortion (blind equalization) and additive noise (double Wiener filter).

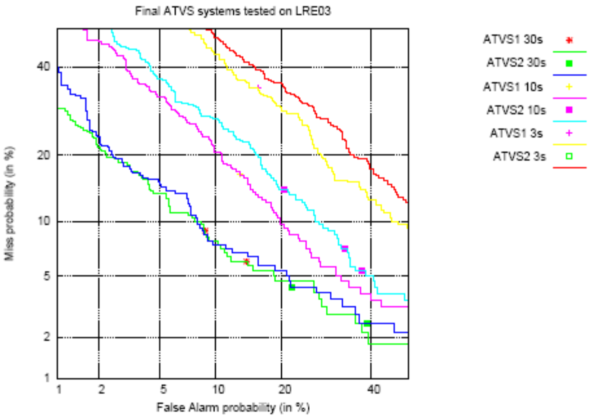

Figure 1: *Baseline system: results on a subset of NIST LRE'03 data using only the 7 languages considered in NIST LRE'05.*

The n-grams used as models for the different languages were trigrams without cut-off factor adapted from a UBM using data from one of the CallFriend (devset) database languages. The UBM n-gram was trained using transcriptions of speech segments (CallFriend devset) from the 12 CallFriend database languages. The adaptation coefficient was determined empirically and set to 0.6 for the UBM a 0.4 for the model only from the language.

Results from these two systems are shown on figure 1 (for a subset of NIST LRE 2003 data containing only test segments of the 7 languages considered in NIST LRE 2005). For NIST LRE 2003 data we attained a 9.14% EER for the 30s condition with the ATVS1 system and virtually the same with the ATVS2 system.

3. Robust vs. standard parameterization

Our baseline system used a robust front-end standardized by ETSI [10]. This front-end includes channel and noise effects compensation and has proved to produce better speech recognition results in noisy conditions. However, this front-end was less efficient than standard front-ends and was difficult to integrate with our systems. For that reason, we compared in a LR task the ETSI front-end to other simpler and more efficient. Our new front-end uses 12 MFCCs plus C0 and their velocities and accelerations for a total of 39 components, computing a feature vector each 10ms and performing Cepstral Mean Normalization (CMN).

Figures 2 shows results on NIST LRE 2003 data of 3 systems identical to the baseline systems, but with the new parameterization. The first one uses 6 phonetic decoders with 10 Gaussians/state, the second 6 with 20 Gaussians/state, and the last one all the 12 phonetic decoders. By comparing figures 1 and 2 we can conclude that the use of a robust front-end has very little influence in language recognition performance – with both front-ends results are virtually the same. By comparing the different systems in Figure 3 we can also conclude that the difference in performance achieved by using the 12 phonetic decoders (at least for the 30sec condition) does not justify the increase in computational cost required.

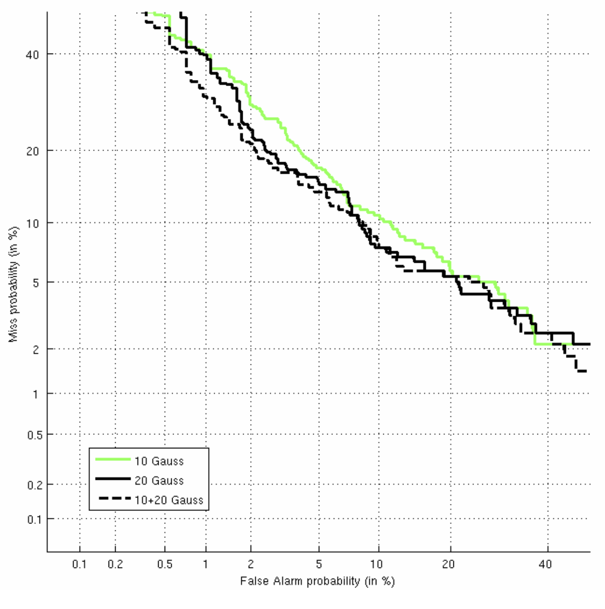

Figure 2: *Baseline system with new front-end: results on a subset of NIST LRE'03 data using only the 7 languages considered in NIST LRE'05.*

4. Adding Voice Activity Detection

One of the main differences between the NIST LR evaluations in 2003 and 2005 is that in 2003 a Voice Activity Detector (VAD) was used by NIST to remove silence areas from the recordings, while in the 2005 evaluation silence segments were kept in the recording to make conditions more realistic. Our baseline system did not include an explicit VAD. It tried to remove the effect of silence segments by removing repetitions of the silence label before training the n-grams and computing the scores. We suspected that the lack of a prior VAD to remove silences could be one of the reasons for the difference in performance between NIST LRE 03 data and NIST LRE 05 data. In order to explore this issue we have included a VAD based on energy levels and temporal restrictions and have obtained results on NIST LRE 03 data and NIST LRE 05 data, using in both cases the new parameterization and only 6 phonetic decoders with 20 Gaussians per state.

The comparison of results obtained for the systems with and without VAD on NIST LRE 03 data (figure 3) and NIST LRE 05 data (figure 4) shows that results are almost the same with and without VAD. This means that the removal of repetitions of the silence label seems to be an adequate way of removing the influence of the silent segments. Computational efficiency, however, is higher with the inclusion of an external VAD that avoids further processing of silences.

5. Using better phonetic decoders

Quality of the phonetic decoders has been recently proposed as a crucial factor in PPRLM performance for language recognition [11]. However, the experiments in [11] were performed using a very special phonetic decoder using artificial neural networks. Here we will extend the work in [11] by checking whether the same conclusions stand for more conventional HMM-based phonetic decoders. Towards this end, we have substituted the phonetic decoders trained on OGI Multi-Language Telephone Speech Corpus, which contained around 1-2 hours of speech by language, by new phonetic decoders trained on SpeechDat-like corpora, all of which contain over 10 hours of training material covering hundreds of different speakers. In particular, we have trained 6 new phonetic decoders in English, German, French, Arabic, Basque and Russian using SpeechDat-like corpora. We have also included a 7th phonetic decoder in Spanish trained on Albayzin [12] downsampled to 8 kHz, which contains about 4 hours of speech for training, but we report results separately for the system with the 6 and 7 recognizers. All the phonetic decoders share the same HMM structure – identical to the baseline systems, with 20 Gaussians/state. Also, the front-end is the same used in former sections and the systems include the external VAD.

With the new phonetic decoders important improvements are obtained. For the NIST LRE 2003 data (figure 3) just by changing the 6 phonetic decoders trained on OGI by 6 phonetic decoders trained on SpeechDat-like corpora language recognition results improve very significantly moving from 10.04% EER to 6.45% EER. Adding the Spanish recognizer the EER reduces to only 5.08%. This improvement is even more noticeable on NIST LRE 2005 data (figure 4). Here we move from a 16.38% EER to an 8.37% EER – a relative reduction of almost 50%. Adding the phonetic decoder for Spanish we get a 7.94% EER. These results stress the importance of having good quality phonetic decoders for language recognition based on PPRLM.

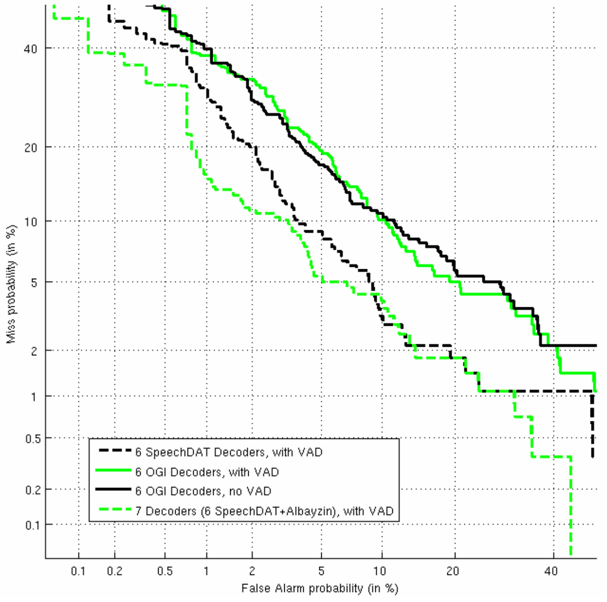

Figure 3: *The effect of VAD and better phonetic models: Comparison of results using models trained with OGI (with and without VAD) and models trained on SpeechDAT-like corpora on a NIST LRE'03 subset using only the 7 languages of NIST LRE'05.*

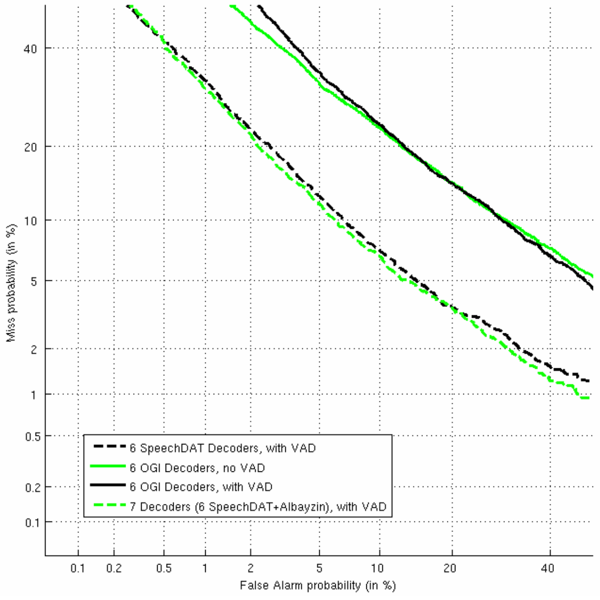

Figure 4: *The effect of VAD and better phonetic models: Comparison of results using models trained with OGI (with and without VAD) and models trained on SpeechDAT-like corpora on NIST LRE'05 data.*

6. SVM Sytems with MFCC and SDC-MFCC features

Besides PPRLM systems, which tend to be the best performing individual systems for LR [5], other systems very used for LR are acoustic systems that model the acoustic features for each particular language, typically using Shifted Delta Cepstra (SDC) features.

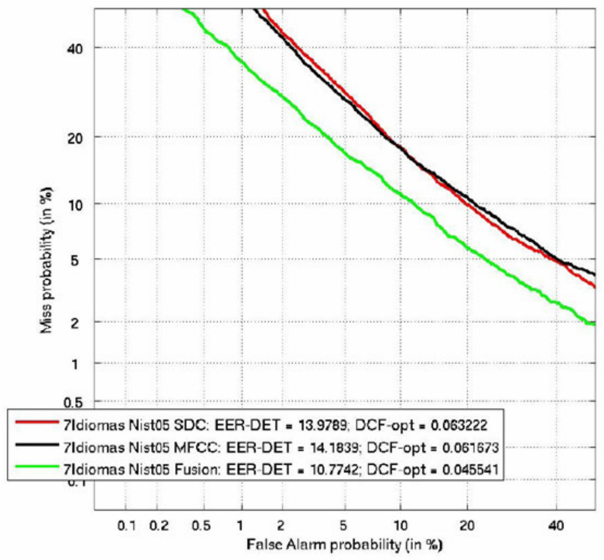

Figure 4: *Acoustic SVM systems using MFCC and SDC features, and the fusion of both. Results on NIST LRE'05 data.*

We have developed an acoustic system based on Support Vector Machines (SVM) [13]. Actually the system is the sum fusion of two SVM systems, one using 19 MFCC coefficients plus deltas and the other using SDC-MFCCs (7-2-3-7) [14]. In order to avoid channel mismatch effects, Cepstral Mean Normalization is applied, followed by RASTA filtering and feature mapping [15]. Both systems use a kernel expansion on the whole observation sequence, and a separating hyperplane is computed between the target language features and the background model. ATVS acoustic SVM-GLDS system uses a polynomial expansion of degree three [16] followed by a Generalized Linear Discriminant Sequence kernel (GLDS) as described in [17]. Finally, Tnorm score normalization technique is performed in order to scale the scores distribution.

The system has been trained using data from CallFriend, NIST LRE 1996, NIST LRE 2003 and has been evaluated on NIST LRE 2005 data (figure 4). The SVM system using MFCC features achieved a 14% EER and the SVM system using SDC-MFCC features achieved a 13.2% EER on NIST LRE 2005 data. When these two SVM systems were fused together with sum fusion we achieved an EER (figure 5) of only 10.5%.

7. Fusion with acoustic systems

Systems submitted to NIST LR Evaluations are rarely based on a single methodology. Rather they are usually the fusion of several systems using different approaches to the problem of LR. Even if the other systems are worse in terms of LR performance than the PPRLM system, the fusion of different systems tend to improve overall LR performance.

We have fused the results of our improved PPRLM system and our new SVM acoustic system with a simple sum fusion followed by Tnorm. This fusion has produced the best result we have achieved so far on NIST LRE 2005 data (figure 5), a 5.43% EER, which implies a relative reduction of almost 66% from our baseline system.

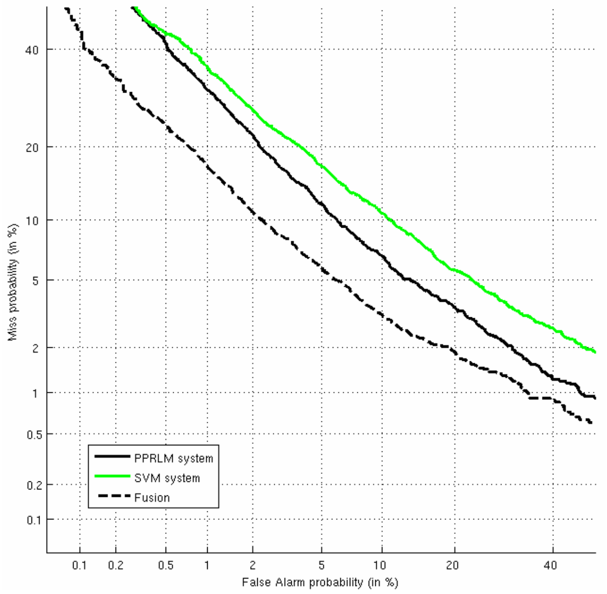

Figure 5: *Fusion of PPRLM system and acoustic SVM system on NIST LRE'05 data.*

8. Conclusions

In this paper we have improved our baseline PPRLM system achieving an EER reduction of almost 50% (from 16.38 to 8.37%). This improvement was mainly achieved by changing the phonetic decoders by other better trained (on more and more adequate data). We have also improved our PPRLM system by adding an explicit Voice Activity Detector (VAD) and a simpler front-end. While the influence of these changes on LR performance is very limited, they improve substantially the computational efficiency of the PPRLM system.

We have also developed a new acoustic system based on the fusion of two SVM systems, on using standard MFCC features and other using SDC features. Each of these systems achieves a LR performance of 13-14% EER by itself, but the fusion of both achieves an EER of only 10.5%.

By fusing our improved PPRLM system with our new acoustic SVM system we obtain a remarkable 5.43% EER on NIST LRE 05 data, which represents an EER relative reduction of around 66% from our baseline system.

9. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

10. References

[1] "National institute of standard and technology. Language Recognition Evaluation Main Page," http://www.nist.gov/speech/tests/lang/index.htm.

[2] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech.," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[3] Gleason T.P. Campbell W.M. Reynolds D.A. Singer E., Torres-Carrasquillo P.A., "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Proc. Eurospeech 2003*, Sept. 2003, pp. 1345–1348.

[4] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition using Phone Lattices", in Proc. ICSLP 2004.

[5] Shen, W., Campbell, W., Gleason, T., Reynolds, D., Singer, E., "Experiments with Lattice-based PPRLM Language Identification", in Proc. IEEE Odyssey 2006, Puerto Rico, June 2006.

[6] A. 0. Hatch, B. Peskin, & A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding", In Proc. ICASSP 2005, Vol. 1, pp. 169-172.

[7] A. Montero-Asenjo, D. T. Toledano, J. González-Domínguez, J. González-Rodríguez, J. Ortega-García, "Exploring PPRLM performance for NIST 2005 Language Recognition Evaluation", in Proc. IEEE Odyssey 2006, Puerto Rico, June 2006.

[8] "OGI multi language telephone speech," http://www.cslu.ogi.edu/corpora/mlts/.

[9] Hidden Markov Model ToolKit (HTK), available on http://htk.eng.cam.ac.uk/ .

[10] ETSI ES 202 050 (v1.1.3): "Speech processing, transmisión and quality aspects (STQ); Distributed speech recognition; Advanced front-end features extraction algorithm; Compression algorithms."

[11] Matejka Pavel, Schwarz Petr, Cernocký Jan, Chytil Pavel, "Phonotactic Language Identification using High Quality Phoneme Recognition", In: Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, Lisbon, PT, 2005, p. 2237-2240.

[12] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, C. Nadeu, "ALBAYZÍN Speech Database: Design of the Phonetic Corpus," in *proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH).* Berlin, Germany, 21-23 September 1993. Vol. 1. pp. 175-178.

[13] 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.

[14] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds and J.R. Deller Jr, "Approaches to Language Identification using Gaussian Mixture. Models and Shifted Delta Cepstrum", ICSLP, 2002.

[15] Gonzalez-Rodriguez, J., Ramos-Castro, D., Torre-Toledano, D., Montero-Asenjo, D., Gonzalez-Dominguez, J., Lopez-Moreno, I., Fierrez-Aguilar, J., Garcia-Romero, D. and Ortega-Garcia, J., "On the Use of High-level Information for Speaker Recognition: the ATVS-UAM system at NIST SRE 2005", to appear in IEEE Aerospace and Electronic Systems Magazine, 2007.

[16] W. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in Proc. of IEEE International Workshop on Neural Networks for Signal Processing, 2000, pp. 775–784.

[17] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in Proc. of ICASSP, 2002, pp. 161–164.