ATVS-UAM ALBAYZIN-VL08 SYSTEM DESCRIPTION

Doroteo T. Toledano, Ismael Mateos-Garcia, Alejandro Abejon-Gonzalez, Daniel Ramos, Juan Bonillo and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

{doroteo.torre, ismael.mateos, alejandro.abejon, daniel.ramos, juan.bonillo joaquin.gonzalez} @uam.es

Abstract

ATVS submission to ALBAYZIN-VL08 will consist of different combinations of a set of acoustic and phonotactic subsystems that our group has developed during the last years. Most of these subsystems have already been evaluated on NIST LRE 07 evaluation. At the time of writing this system description some of the details of our submission are still undefined. Therefore we will briefly describe our systems and the intended combinations to be submitted, but these settings should not be taken as final in any way. As acoustic subsystems we will use a GMM SuperVectors and a GLDS-SVM subsystem, while the phonotactic subsystem will be a PhoneSVM system. We are still deciding the best fusion strategy and the best combination of subsystems at the time of writing. Output scores will be submitted in the form of log-likelihood ratio (logLR) scores in an application independent way. Open-set detection thresholds will be set to the Bayes thresholds in all cases, and the same logLR sets will probably be submitted to the closed- and open-set conditions.

1. Introduction

ATVS-UAM submission to ALBAYZIN-VL08 consists of different combinations of a set of acoustic and phonotactic subsystems that our group has developed during the last years. The two acoustic subsystems are based on two different techniques: SVM-GLDS (language recognition using SVMs with Generalized Linear Discriminant Sequence kernel) and GMM SuperVectors (also named as GMM-SVM in this document, is language recognition using SVMs that take as input the means of Gaussian Mixture Models). The phonotactic system will be a Phone-SVM subsystem (phone recognition and n-gram modeling followed by Support Vector Machine classification).

A particularity of all of ATVS subsystems is that no transcribed speech is needed to train language models. This makes them particularly useful for situations where few language resources are available or when transcription of materials for training the language models is difficult or very expensive. For this reason, our subsystems are better fitted for the restricted training condition of the evaluation. The Phone-SVM subsystem, however, requires phonetic recognizers trained on different corpora, and therefore cannot be included in our submission to the restricted training condition.

The same individual subsystems will be used to perform language recognition for test segments of 3, 10 and 30s. These subsystems will be fused together in some way. At this time we are experimenting with several fusion strategies ranging from sum fusion to anchor model fusion. The scores will be submitted as calibrated Log-Likelihood Ratios.

The rest of this system description is organized as follows: Section 2 describes the acoustic subsystems, Section 3 the phonotactic subsystems, Section 4 the fusion strategies and Section 5 the calibration process.

2. Acoustics systems

We use two different acoustic systems, both of which are based on SVMs using different features. In this section we describe these two acoustic systems and the training material used for training them.

2.1. Individual Sub-systems: SVM-GLDS

The first individual sub-system is, in fact, the fusion of two acoustic systems based on SVM [1,2] using different features. 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. Both ATVS acoustic SVM-GLDS subsystems use a polynomial expansion of degree three [3] followed by a Generalized Linear Discriminant Sequence kernel as described in [4].

2.2. Individual Sub-systems: GMM-SVM (or GMM SuperVectors)

This subsystem is based on using an SVM classifier over the GMM models space. The language model is constructed by MAP (Maximum A-Posteriori) adaptation of the means of the UBM (Universal Background). A GMM super vector is constructed by stacking the means of the adapted mixture components. Then the SVM classifier is used to train and separating hyperplane in the vector space defined by the super-vectors.

2.3. Training data used

To fulfil the restricted condition training of the ALBAYZIN-VL08 evaluation these subsystems have been trained and adjusted using exclusively the training (and development) data supplied by the organization of the evaluation. In this way, these acoustic systems can (and will) be used for the restricted training condition. We have downsampled the speech materials provided by the organization to 8 kHz because our systems were developed to work on telephone speech. This will limit the performance of our systems.

3. Phonotactic Systems

Our phonotactic system consists of a Phone-SVM system using 7 phonetic recognizers in 7 different languages. In this section we describe this system in more detail and discuss the training material used.

3.1. Individual Sub-systems: Phone-SVM

Each of the seven different Phone-SVM subsystems is based on the following steps. First a voice activity detector segments

the test utterance into speech and non-speech segments. The speech segments are recognized with one open-loop phonetic decoder. The best decoding is used to estimate count-based 1-grams, 2-grams and 3-grams. All these parameters are reshaped as a single vector that is taken as the input of an SVM that classifies the test segment as corresponding (or not) to one language.

The process described above is repeated for the seven different open-loop phonetic recognizers used. All decoders are based on Hidden Markov Models (HMMs) trained using HTK and used for decoding with SPHINX. The phonetic HMMs are three-state left-to-right models with no skips, being the output pdf of each state modelled as a weighted mixture of Gaussians.

For each test utterance, the systems make n-grams with the transcription produced by the phonetic decoders. Support Vector Machines (SVMs) take the n-grams as input vectors [1,2].

3.2. Training data used

Most of the seven decoders were trained on SpeechDat-like corpora, containing over 10 hours of training material per language and covering hundreds of different speakers. The languages of these phonetic decoders are English, German, French, Arabic, Basque and Russian. We have also included a 7th phonetic decoder in Spanish trained on Albayzin [5] downsampled to 8 kHz, which contains about 4 hours of speech for training. We developed these phonetic recognizers for telephone speech (mainly for NIST LRE evaluations) and will use them with the test materials downsampled to 8 kHz. This will limit the performance of our systems.

Although the models that we will use for detecting the language for the ALBAYZIN-VL08 evaluation will be trained entirely on the training and development materials provided by the organization, the use of previously trained phonetic decoders in this system makes it usable only in the unrestricted training condition.

4. Fusion

In order to combine the results of the subsystems presented above different fusion techniques are currently being explored, from classical techniques, like sum fusion, to novel ones like anchor-models fusion [6-8]. Prior to any processing, the scores of each of the individual sub-systems are normalized using a test-segment dependent normalization. This normalization is also currently under study, so we have not yet decided the final configuration.

5. Calibration and decision

In order to take the actual decision we will follow a per-language detection approach in order to calibrate the output log-likelihood-ratios (logLR). Therefore, each score *for each of the target languages* will be mapped to a logLR assuming a target-language-vs.-all configuration, in the following way:

$$\log(LR) = \log\left(\frac{P(\text{score}|\text{target language})}{P(\text{score}|\text{any other non-target language})}\right)$$

After calibrating logLR values, the logarithm of the Bayes threshold will be used in order to take decisions. If the calibration process is correctly performed, this is equivalent to choosing the minimum-cost threshold for each target language detection sub-system.

6. Conclusions

For the ALBAYZIN-VL08 evaluation, we have built mainly on previously developed subsystems that we have used in NIST LRE 07, trying to adapt them for the particular task and languages proposed in the ALBAYZIN-VL08 evaluation. Our systems have been developed with the requirement of easy training for new languages, so it has been relatively straightforward to train them for the languages of the ALBAYZIN-VL08 evaluation. However, once the subsystems have been trained we still have to fine tune the fusion and the calibration. This work is still in progress at the time of writing.

Acknowledgments

We thank the organizers of the ALBAYZIN-VL08 evaluation for their hard work in preparing this evaluation and the corresponding training, development and test materials. This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

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] 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.

[3] 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.

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

[5] 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.

[6] N. Brümmer et al. “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006.” IEEE Transactions on Audio, Speech and Signal Processing, 15(7) pp. 2072-2084, 2007.

[7] Mikael Collet, Yassine Mami, Delphine Charlet, Frederic Bimbot, "Probabilistic Anchor Models Approach for Speaker Verification", in INTERSPEECH 2005.

[8] Elad Noor1, Hagai Aronowitz "Efficient Language Identification using Anchor Models and Support Vector Machines", in Odyssey 2006 ISBN: 1-4244-0472-X pp 1-6.