Exploiting High-Level Information Provided by ALISP in Speaker Recognition

Asmaa El Hannani[1,⋆] and Dijana Petrovska-Delacrétaz[1,2]

[1] DIVA Group, Informatics Dept., University of Fribourg, Switzerland
[2] Institut National des Télécommunications, 91011 Evry, France
asmaa.elhannani@unifr.ch dijana.petrovska@int-evry.fr

Abstract. The best performing systems in the area of automatic speaker recognition have focused on using short-term, low-level acoustic information, such as sepstral features. Recently, various works have demonstrated that high-level features convey more speaker information and can be added to the low-level features in order to increase the robustness of the system. This paper describes a text-independent speaker recognition system exploiting high-level information provided by ALISP (Automatic Language Independent Speech Processing), a data-driven segmentation. This system, denoted here as ALISP n-gram system, captures the speaker specific information only by analyzing sequences of ALISP units. The ALISP n-gram system was fused with an acoustic ALISP-based Gaussian Mixture Models (GMM) system exploiting the speaker discriminating properties of individual speech classes. The resulting fused system reduced the error rate over the individual systems on the NIST 2004 Speaker Recognition Evaluation data.

1 Introduction

In recent years, research has expended from only using the acoustic content of speech to trying to utilise high-level information, such as linguistic content pronunciation and idiolectal word usage. Works examining the exploitation of high-level information sources have provided strong evidence that gains in speaker recognition accuracy are possible [1]. [2] explored the possibility of using word n-gram statistics for speaker verification. This technique although simple, gave encouraging results. Motivated by the work of [2], [3] applied similar techniques to phone n-gram statistics. This approach gave good results and was found to be a useful complementary features when used with short-term acoustic features. The research of [2] and [3] showed word and phone n-gram based model to be promising for speaker verification, however these techniques still based on human transcription of the speech data. The system we are proposing in this paper is inspired from the system described in [3], except that we used the automatic segmentation based on Automatic Language Independent Speech Processing (ALISP) tools [4] instead of the phonetic one. The ALISP-sequences, are

⋆ Supported by the Swiss National Fund for Scientific Research, No. 2100-067043.01/1.

automatically acquired from the output of the ALISP recognizer with no need of transcribed databases. In [5] we have built an ALISP-based GMM system exploiting the speaker discriminating properties of individual speech classes and we have shown that the ALISP segments could capture speaker information. In the ALISP n-gram system we are presenting here, speaker specific information is captured only by analyzing sequences of ALISP units. The ALISP-based GMM system and the ALISP n-gram system are combined to complement each other. The resulting fused system reduces the error rate over the individual systems.

The outline of this paper is the following: In Section 2 more details about the proposed method are given. Section 3 describes the database used and the experimental protocol. The evaluation results are reported in Section4. The conclusions and perspectives are given in Section 5.

2 System Description

2.1 ALISP segmentation

The systems described bellow use in the first stage a data-driven segmentation Automatic Language Independent Speech Processing (ALISP) tools [4]. This technique is based on units acquired during a data-driven segmentation, where no phonetic transcription of the corpus is needed. In this work we use 64 classes. The modelling of the set of data-driven speech units, denoted as ALISP units, is achieved through the following stages. After the pre-processing step for the speech data, first Temporal Decomposition is used, followed by Vector Quantization providing a symbolic transcription of the data in an unsupervised manner. Hidden Markov Modeling is further applied for a better coherence of the initial ALISP units.

2.2 ALISP N-gram System

The focus here is to capture high-level information about the speaking style of each speaker. Speaker specific information is captured by analyzing sequences of ALISP units produced by the data-driven ALISP recognizer. In this approach, only ALISP sequences are used to model speakers. For the scoring phase each ALISP-sequence is tested against a speaker specific model and a background model using a traditional likelihood ratio. The speaker specific ALISP-sequence models is generated using a simple n-gram frequency count as follows:

$$L_i(k) = \frac{C_i(k)}{\sum_{n=1}^{N} C_i(n)} \tag{1}$$

where k represents an n-gram token (ALISP sequence) and $C_i(k)$ is the frequency count of the token k in the train data of the speaker i.

The background model is estimated using

$$L_{Bm}(k) = \frac{C_{Bm}(k)}{\sum_{n=1}^{N} C_{Bm}(n)} \tag{2}$$

where $C_{Bm}(k)$ is the frequency count of the token k in the world data.

Then for each ALISP n-gram found in the test utterance a score is calculated using the log-likelihood ratio of the speaker likelihood to the background likelihood

$$S_{ti} = \frac{\sum_{n=1}^{M} (C_t(n).log\,[L_i(n) - L_{Bm}(n)])}{\sum_{n=1}^{M} C_t(n)} \tag{3}$$

where $C_t(n)$ is the count of the the token n in the test utterance t.

Finally, the ALISP n-gram scores are fused together to generate an overall score for the test segment.

In this work three n-gram (1-gram, 2-gram and 3-gram) systems are built. The evaluation of their individual performances and their fusion is presented in section 4.

2.3 ALISP-based GMM System

This system uses GMMs on a segmental level in order to exploit the different amount of discrimination provided by the ALISP classes [5]. In this segmental approach we represent each speaker by 64 GMMs each of them models an ALISP class. The speaker specific 64 models were adapted from the 64 gender and ALISP class dependent background models.

During the test phase, each test speech data is first segmented with the 64 ALISP HMM models. Then, each ALISP segment found in the test utterance is compared to the hypothesized speaker models and to the background model of the specific ALISP class.

Finally, and after the computation of a score for each ALISP segment, the segmental scores are combined together to form a single recognition score for the test utterance. A Multi-Layer Perceptrons (MLP) [6] is used to combine the individual scores for the ALISP segments.

2.4 Fusion

There are several scenarios for combining the decisions of multiple systems [7]. In [8] we have compared three fusion methods of speaker verification systems: the linear summation, the Logistic Regression (LR) and the Multi-Layer Perceptron (MLP). In this work we choose a Multi-Layer Perceptron to fuse the scores from the various systems. This perceptron has a layer consisting of inputs for each system, a hidden layer with 5 neurons, and an output layer using sigmoid as activation function.

3 Experimental Setup

All experiments are done on the NIST'2004 data which is split into two different subsets: the *Development-set* and the *Evaluation-set*, used to test the performance of the proposed system.

The speech parametrization is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame a 15-element cepstral vector is computed and appended with first order deltas. Cepstral mean substraction is applied to the 15 static coefficients and only bands in the 300-3400 Hz frequency range are used. The energy and delta-energy are used in addition during the ALISP units recognition.

During the preprocessing step, after the speech parametrization, we separated the speech from the non-speech data. The speech activity detector is based on a bi-Gaussian modeling of the energy of the speech data [9]. Only frames higher than a certain threshold are chosen for further processing. Using this method, 56% of the original NIST 2004 data are removed.

In the ALISP-based GMM system [1], 64 ALISP-specific gender-dependent background models (with 32 Gaussians) are built and for each target speaker, 64 specific GMM with diagonal covariance matrices is trained via maximum a posteriori (MAP) adaptation of the Gaussian means of the matching gender background models. If an ALISP class does not occur in the training data for a target, the background model of this class becomes that target's model.

The gender dependent background models for the GMMs and the gender dependent ALISP recognizers, are trained on a total of about 6 hours of data from (1999 and 2001) NIST data sets. The MLP is trained on the development set.

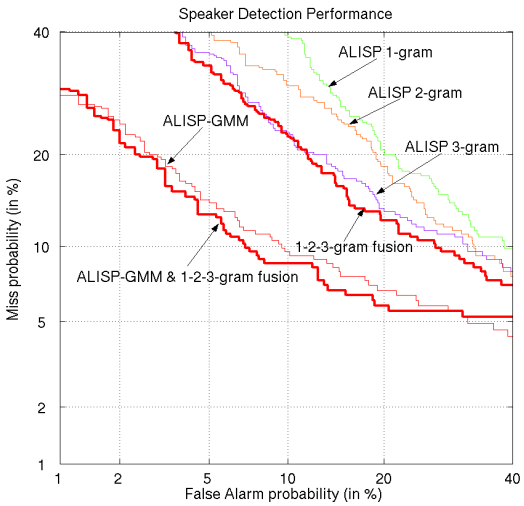

Fig. 1. Speaker verification results for the ALISP-based GMM system, the n-gram systems and their fusion on the evaluation data set (subset of NIST'04).

[1] based on the BECARS package [10].

4 Experimental Results

We present in this section results for "8sides-1side" NIST 2004 task on the evaluation data set, as defined in section 3. For this task we dispose of 40 minutes to build the speaker model and 5 minutes for the test data (including silences). Performance is reported in term of the Detection Error Tradeoff (DET) curve [11]. Results are compared via Equal Error Rates (EER): the error at the threshold which gives equal miss and false alarm probabilities.

The Figure 1 shows DET curves of the fusion results. For reference, the four individual systems are also shown. Fusing the three ALISP n-gram (1-gram, 2-gram, 3-gram) systems lead to an improvement over the individual n-gram systems. These systems although worse compared to the ALISP-based GMM system, gave encouraging results.

In the next set of experiments, we fused the ALISP n-gram systems with the ALISP-based GMM system. Results are clearly showing that the new systems (ALISP n-gram) are supplying complementary information to the acoustic system (ALISP-based GMM).

5 Conclusions

In this paper we have presented a speaker verification system based on data-driven speech segmentation and exploiting high-level information. We have shown that the fusion of the acoustic ALISP-based GMM system with the n-gram systems treating high-level information (provided by the ALISP sequence), improve the speaker recognition accuracy. The great advantage of the proposed method is that it is not grounded on the usage of transcribed speech data. In other hand the ALISP data-driven segmentation can be used in different levels in speaker verification systems in order to extract complementary types of information.

6 Acknowledgement

Our thanks go to Jean Hennebert for his helpful discussions and MLP software.

References

1. Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, J., Xiang, B.: The supersid project: Exploiting high-level information for high-accuracy speaker recognition. In Proc. ICASSP (2003)
2. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. Eurospeech **vol. 4** (2001) 2517–2520
3. Andrews, W., Kohler, M., Campbell, J., Godfrey, J.: Phonetic, idiolectal, and acoustic speaker recognition. Speaker Odyssey Workshop (2001)

4. Chollet, G., Černocký, J., Constantinescu, A., Deligne, S., Bimbot, F.: Towards ALISP: a proposal for Automatic Language Independent Speech Processing. In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag (1999)
5. El-Hannani, A., Petrovska-Delacrétaz, D.: Improving speaker verification system using alisp-based specific GMMs. submitted to AVBPA (2005)
6. Haykin, S.: Neural Networks: A Comprehensive Foundation. IEEE Computer society Press (1994)
7. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **vol. 20** (1998) 226–239
8. El-Hannani, A., Petrovska-Delacrétaz, D., Chollet, G.: Linear and non-linear fusion of alisp-based and GMM systems for text-independent speaker verification. In proc. of ODYSSEY04, The Speaker and Language Recognition Workshop (2004)
9. Magrin-Chagnolleau, I., Gravier, G., Blouet, R.: Overview of the 2000-2001 elisa consortium research activities. Speaker Odyssey Workshop (2001)
10. Blouet, R., Mokbel, C., Mokbel, H., Sanchez, E., Chollet, G., Greige, H.: Becars: A free software for speaker verification. Proc. Odyssey (2004)
11. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The det curve in assessment of detection task performance. Proc. Eurospeech'97 **vol. 4** (1997) 1895–1898