



Fusing Acoustic, Phonetic and Data-Driven Systems for Text-Independent Speaker Verification

Asmaa El Hannani^{1,2}, Dijana Petrovska-Delacrétaz²

¹DIVA Group, Informatics Dept., University of Fribourg, Switzerland

²EPH Dept., Institut National des Télécommunication, Evry, France

asmaa.elhannani@unifr.ch, dijana.petrovska@int-evry.fr

Abstract

This paper describes our recent efforts in exploring data-driven high-level features and their combination with low-level spectral features for speaker verification. In particular, we compare the phonetic and data-driven approaches and study their complementarity with short-term acoustic approach. Our objective is to show that data-driven units automatically acquired from the speech data, can be used like phonemes to extract high-level features and to bring complementary speaker-specific information that can therefore provide improvements when fused with acoustic systems. Results obtained on the NIST 2006 Speaker Recognition Evaluation data show that the combination of the phonetic, data-driven and Gaussian Mixture Models (GMM) systems brings a 27% relative reduction of the EER in comparison to the baseline GMM system.

1. Introduction

The speech signal conveys roughly two kinds of information about the speaker's identity. The first set of information reflects the spectral properties of speech (low-level) which are related to the physical structure of the vocal apparatus. These parameters are used since the beginning of the research in automatic speaker recognition. The second set of information reflects the behavioral traits (high-level) such as prosody, phonetic information, pronunciation, idiolectal word usage, conversational patterns, topics of conversations, etc...

Recently, there was a rising interest of computing and modelling high-level features, which start to be used in combination with low-level spectral features. Studies examining the exploitation of high-level information sources have provided strong evidence that gains in speaker recognition accuracy are possible [1, 2, 3, 4]. Usually these high-level features are extracted by analyzing streams produced by phonetic speech recognition systems. Two of the major problems that arise when phone based systems are being developed are the possible mismatches between the development and evaluation data and the lack of transcribed databases. In order to solve these two problems we propose to use a data-driven approach instead the phonetic one to extract such high-level features. In this way the availability of corpora is much less an issue and the training corpus can be chosen to match the working conditions as much as possible. Our data-driven approach is based on Automatic Language Independent Speech Processing (ALISP) tools [5].

The purpose of this paper is to investigate the fusion of different high-level systems with the acoustic system. Our objective is to show, through this work, that data-driven based approach provides complementary information and can be further combined with other systems to improve speaker recognition

accuracy. Specifically, we investigate and compare combination of phonetic and data-driven systems supplying high-level information with GMM systems.

The outline of this paper is the following: Section 2 and 3 describe the acoustic and high-level systems. In section 4 the evaluation results are reported. The conclusions are given in section 5.

2. Acoustic Level Systems

In this section we describe two acoustic systems: segmental and global GMM systems. The main difference between global and segmental GMM systems, is that in the former case each speaker is represented by one model. However, in the latter case each speaker is modeled via N models ($N = 64$), corresponding each to a speech class.

Same front-end processing and same modeling tools are used for both systems. The speech parameterization is done with Linear Frequency Cepstral Coefficients (LFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame a 16-element cepstral vector is computed and appended with first order deltas and the delta-energy. Bandwidth is limited to the 300-3400Hz range. The parameter vectors are normalized to fit a zero mean and a unit variance distribution. The mean and variance used for the normalization are computed file by file on all the frames kept after applying the frame removal processing. The parameterization is carried out using SPRO tools [6].

2.1. Global GMM System

The baseline system is a standard Gaussian Mixture Models (GMM) [7] system in which the multivariate probability density function of the feature vectors is modeled with a weighted sum of gaussians. Two gender-dependent background models (λ_{BM}) are trained and each speaker model (λ_S) is obtained by adaptation of the matching gender background model.

The score of the tested sequence of features $X = x_1, \dots, x_T$ is calculated using log-likelihood ratio of the speaker to the background likelihood as follows:

$$\Lambda(X) = \frac{1}{T} \sum_{t=1}^T \log p(x_t | \lambda_S) - \frac{1}{T} \sum_{t=1}^T \log p(x_t | \lambda_{BM}) \quad (1)$$

Note that average log-likelihood values are used in order to normalize out the duration effects from the log-likelihood value.

2.2. Segmental ALISP GMM System

The segmental GMM system [8] is composed of the following four steps: first the speech data is segmented using the AL-

ISP data-driven segmentation tools as described in section 3.1. Secondly ALISP class-specific background models, denoted by $\lambda_{BM,k}$ where $k = 1, \dots, 64$, are built using the feature vectors for the given ALISP class. Then speaker models are obtained via adaptation of the background models. Thus, each speaker is modeled by 64 GMMs corresponding each to an ALISP class. 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.

During the test phase, each test speech data is first segmented with the ALISP recognizer, leading to a sequence of segments $Y = y_1, \dots, y_I$. Then, background and target scores are computed for each segment $y_i = x_{1_i}, \dots, x_{T_i}$ appearing in the test utterance. The segmental scores are computed using the log-likelihood ratio of the speaker to the background likelihood using the matching ALISP class model. Assuming for example that the segment y_i has been associated to the ALISP class k , the segmental score of this segment is computed as follows:

$$\Lambda(y_i) = \frac{1}{T_i} \sum_{t_i=1}^{T_i} \log p(x_{t_i} | \lambda_{S,k}) - \frac{1}{T_i} \sum_{t_i=1}^{T_i} \log p(x_{t_i} | \lambda_{BM,k}) \quad (2)$$

Finally, and after the computation of a score for each ALISP segment found in the test utterance, the segmental scores $\Lambda(y_i)$ are combined together to generate an overall score. In this work Multilayer Perceptrons (MLP) are used to combine the individual scores for the ALISP segments.

3. High-level Systems

High-level systems include two parts. The first one consists in building the recognizer that outputs labeling of the speech data. The second part consists in performing speaker verification, based on these labelings. In this section we first introduce data-driven recognizer, then we describe the various possibilities that we have used in order to extract high-level information from the speech labelings.

3.1. Data-driven Segmentation

The data-driven speech units, denoted here as ALISP units, are automatically determined from the training corpus, with no need of phonetic transcription of the corpus. They do not require transcribed data. The steps needed to acquire and model the ALISP units are the following. After the pre-processing step, temporal decomposition is used to obtain an initial segmentation of the speech data into quasi-stationary segments. The speech segments correspond actually to spectrally stable portions of the signal. We then compute the gravity center for each segment and train a gender dependent vector quantizer to cluster these centers of gravity. The codebook size (64 in our case) defines the number of ALISP symbols. The initial labeling of the entire speech segments is achieved using minimization of the cumulated distances of all the vectors from the speech segment to the nearest centroid of the codebook. The result of this step is an initial segmentation and labeling of the training corpus. Hidden Markov Models (HMMs) are then initialized from this labeled segments to build a set of 64 ALISP units. The HMM units are then re-trained on the data set by applying a Baum-Welch re-estimation. Each ALISP unit is modeled by a left-to-right HMM having three emitting states and containing up to 8 mixtures each.

The speech parameterisation for ALISP recognizer 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. Only bands in the 300-3400 Hz frequency range are used. The parameterisation is done with HTK toolkit [9]. Cepstral mean subtraction is applied to the 15 static coefficients. The mean estimator used for the normalization is computed file by file on all the speech frames kept after applying the speech activity detector.

3.2. Speakers Modelling

3.2.1. Idiolectal System

The modelling principle of this system [10] 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. Although the ALISP units are based on the acoustic features, the speaker verification is performed only from ALISP units sequences produced by the data-driven recognizer.

The speaker model, λ_S , and the background model, λ_{BM} , are generated using a simple n-gram frequency count as follows:

$$\lambda_{BM}(k) = \frac{C_{BM}(k)}{\sum_{n=1}^{N_{BM}} C_{BM}(n)} \quad (3)$$

$$\lambda_S(k) = \frac{C_S(k)}{\sum_{n=1}^{N_S} C_S(n)} \quad (4)$$

where $C_S(k)$ and $C_{BM}(k)$ represent the frequency count of the ALISP n-gram type, k , in the speaker data and world data, respectively. N_S and N_{BM} are the number of all n-gram types in the speaker and world data, respectively.

For the scoring phase each ALISP-sequence is tested against the speaker specific model and the background model using a traditional likelihood ratio.

$$\Lambda(X) = \frac{\sum_{k=1}^K C_X(k) \cdot \log \left[\frac{\lambda_S(k)}{\lambda_{BM}(k)} \right]}{\sum_{k=1}^K C_X(k)} \quad (5)$$

where $C_X(k)$ represents the number of occurrences of the ALISP n-gram type, k , in the test utterance X . The sums are over all of the ALISP n-gram types in the test segment. 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.

3.2.2. Language Model System

In this system [11], the label sequences produced by the ALISP recognizer are used to train ALISP n-gram models using the HTK Language Model (LM) tools (see 14th chapter of the HTK book [9] for more details). The main idea of language models is to predict each symbol in the sequence given its $n-1$ predecessors. It is based on the assumption that the probability of a specific n-gram occurring in some unknown test text can be estimated from the frequency of its occurrence in the training text.

The n-gram probabilities are estimated from speakers and background data. $P_S(s_i | s_{i-n+1} \dots s_{i-1})$ and $P_{BM}(s_i | s_{i-n+1} \dots s_{i-1})$ the probabilities that the segment s_i follows the sequence $s_{i-n+1} \dots s_{i-1}$ in the speaker and background data, respectively, are calculated as follows:

$$P_S(s_i | s_{i-n+1} \dots s_{i-1}) = \frac{C_S(s_{i-n+1} \dots s_i)}{C_S(s_{i-n+1} \dots s_{i-1})} \quad (6)$$

$$P_{BM}(s_i | s_{i-n+1} \dots s_{i-1}) = \frac{C_{BM}(s_{i-n+1} \dots s_i)}{C_{BM}(s_{i-n+1} \dots s_{i-1})} \quad (7)$$

where $C_S(k)$ and $C_{BM}(k)$ represent the number of time the sequence k occurred in the speaker and world data, respectively.

The speaker specific language models are adapted from the background models.

$$\tilde{P}_S = (1 - \alpha) \cdot P_S + \alpha \cdot P_{BM} \quad (8)$$

where α is an adaptation factor ranging from 0 to 1 and set in this work to 0.7

Given a test utterance X , we first produce its labelling $\{s_1, \dots, s_I\}$ using the ALISP recognizer. The sequence of labels s_i are then used to compute the likelihoods with the statistical models computed previously as follows:

$$P_S(X) = \prod_{i=1}^I \tilde{P}_S(s_i | s_{i-n+1} \dots s_{i-1}) \quad (9)$$

$$P_{BM}(X) = \prod_{i=1}^I P_{BM}(s_i | s_{i-n+1} \dots s_{i-1}) \quad (10)$$

The recognition score is a log-likelihood ratio computed with:

$$\Lambda(X) = \frac{1}{I} \log \left[\frac{P_S(X)}{P_{BM}(X)} \right] \quad (11)$$

where I is the number of ALISP segments in the test utterance. As for the idiolectal systems, we built 1-, 2- and 3-gram systems.

3.2.3. Duration System

This system is inspired from the system described in [4], except that we used data-driven units instead of the phonetic ones. In this work two types of duration vectors are extracted: (1) The ALISP unit-level vectors which are composed of the duration of the ALISP units and are one-dimensional vectors and (2) the ALISP state-level vectors, which are composed of the HMM states durations in the ALISP units and are three-dimensional features. Two GMMs system are used to model each type of features. The speaker specific 64 models are adapted from the 64 ALISP class dependent background models. Eight gaussians are used for each model.

During the test phase, the duration vectors of the test speech data are first built. Then, the test duration vectors are compared to the hypothesized speaker model and to the background model of the corresponding ALISP class. The scores are then calculated using the log-likelihood ratio of the speaker likelihood to the background likelihood. The final score is obtained as the sum of the log-likelihoods normalized by the number of the ALISP classes scored.

4. Results

In this section we report the results of the fusion of high-level systems described in section 3 with the GMM systems described in section 2. The scores from the different systems are fused using the SVMlight [12] software. The SVM combiner

uses a Radial Basis Function kernel and is trained on NIST 2004 SRE trials. The systems are evaluated on English trials of the 8conv4w-1conv4w task of NIST 2006 SRE.

Table 1 shows the EER and minimum DCF of the different data-driven systems and their fusion with the GMM system. Many observation can be made from these results. In the first part of the Table the performance of each system alone is given. One can observe clearly from these results that none of the ALISP high-level systems alone is competitive with the GMM systems. This is not a surprising result and is not specific to our data-driven approach. Effectively, high-level systems are usually used as a boosting factor for the performance of the acoustic systems and not as isolated systems.

Systems	EER (%)	min DCF
Segmental GMM (Seg. GMM)	5.09	0.0272
<i>Baseline GMM (GMM)</i>	5.70	0.0290
Segments duration (1)	17.06	0.0842
States duration (2)	16.97	0.0770
Idiolectal: 1gram (3)	20.20	0.0997
Idiolectal: 2gram (4)	16.64	0.0887
Idiolectal: 3gram (5)	15.43	0.0885
Language models: 1gram (6)	21.55	0.0998
Language models: 2gram (7)	15.57	0.0896
Language models: 3gram (8)	15.70	0.0812
Seg. GMM + GMM	4.63	0.0256
Seg. GMM + (1)	4.53	0.0264
Seg. GMM + (2)	4.81	0.0269
Seg. GMM + (3)	5.03	0.0274
Seg. GMM + (4)	5.02	0.0276
Seg. GMM + (5)	5.01	0.0273
Seg. GMM + (6)	5.00	0.0275
Seg. GMM + (7)	4.90	0.0278
Seg. GMM + (8)	5.06	0.0271
ALL	4.30	0.0242
ALL minus (GMM & Seg. GMM)	10.99	0.0539

Table 1: EER and minimum DCF of the different data-driven systems and their fusion with the GMM system.

The second part of the Table presents the performance of the combination of each high-level system with the best GMM system. This latter is a segmental system that exploits the speaker discriminating properties of individual ALISP classes, by modelling them separately using GMMs. The results show that the single best system to fuse with the segmental GMM system is the segments duration system yielding an EER of 4.53%. The minimum DCF is also reduced. Note also that even the fusion of both GMM systems is improving the accuracy.

Finally we conducted experiments examining the fusion of subsets of high-level systems with the GMM systems. For sake of readability only the best result is reported. The best performance was obtained when fusing ALL systems. A 15% relative reduction of the EER is obtained. This result confirms clearly that our data-driven high-level systems are supplying complementary information to the acoustic system and that every system is contributing to improve the fusion result. Fusing only high-level systems gives an EER of 10.99%. This result could certainly be improved by improving the individual performances of high-level systems.

The next set of experiments concerns the comparison of the performances of high-level systems using both phonetic and data-driven approaches and their fusion with the acoustic systems. The phonetic recognizer is based on HMMs and use the

same front-end processing as the ALISP recognizer. Phonetic HMMs are trained using the NTIMIT corpus [13], which is phonetically transcribed using the ARPAbet alphabet. This alphabet contains 61 phonemes. For each ALISP high-level system described in section 3, we built an equivalent phonetic system.

Table 2 shows the comparison of performances of the ALISP data-driven and phonetic high-level systems. Results show that the data-driven based approach outperforms the phonetic one for all systems. One possible explanation to this result is the mismatch between the training data of the phonetic decoders and the evaluation data. These results suggest that better speaker recognition results can be achieved using data-driven units than phonemes, at least if there is little or no transcribed data available recorded in similar conditions as the evaluation data.

Systems	EER (%)	
	ALISP	Phonetic
Segments duration	17.06	23.41
States duration	16.97	21.68
Idiolectal: 1gram	20.20	21.98
Idiolectal: 2gram	16.64	17.62
Idiolectal: 3gram	15.43	17.72
Language models: 1gram	21.55	21.74
Language models: 2gram	15.57	17.62
Language models: 3gram	15.70	18.05

Table 2: Comparison of the phonetic and data-driven high-level systems on the 8conv4w-1conv4w task of NIST 2006 SRE.

Performances of the fusion are reported in Table 3. Systems denoted as "ALISP systems" and "Phonetic systems" correspond to the fusion of all data-driven and all phonetic high-level systems, respectively. As it can be seen in Table 3, the ALISP approach outperforms significantly the phonetic one. However, when fusing both approaches with GMM systems, there is only a slight difference favoring the ALISP approach. A possible explanation to this results is that some of the errors made by the phonetic systems are may be already covered by the GMM systems. Combining the three approaches (GMM, Phonetic and ALISP) further improves the results and again confirms that the systems are indeed providing complementary information.

Systems	EER (%)	min DCF
ALISP systems (A)	10.99	0.0539
Phonetic systems (P)	14.06	0.0625
GMM systems + (A)	4.30	0.0242
GMM systems + (P)	4.44	0.0242
GMM systems + (A) + (P)	4.17	0.0237

Table 3: Comparison of the fusion of data-driven and phonetic systems with the GMM systems on the 8conv4w-1conv4w task of NIST 2006 SRE.

5. Conclusions

In this paper we have shown that like phonetic systems, the data-driven systems can complement short-term acoustic systems to reach better speaker recognition performances. Indeed, the fusion of the acoustic system with the ALISP systems supplying high-level information, provided by units automatically acquired from the speech data, improves the speaker verification accuracy. We have also shown that the three approaches (GMM, Phonetic and ALISP) are providing complementary information.

6. Acknowledgment

This work was supported by the Swiss National Fund for Scientific Research (No. 200020-108024) and by the BioSecure Network of Excellence (IST-2002-507634). Our thanks go also to J. Černocký for the ALISP tools and to F. Bimbot for the temporal decomposition package, which were the basis for the data-driven segmentation tools.

7. References

- [1] D. R. et al., "The supersed project: Exploiting high-level information for high-accuracy speaker recognition," *In Proceedings of ICASSP*, April 2003.
- [2] J. Campbell, D. Reynolds, and R. Dunn, "Fusing high- and low-level features for speaker recognition," *In Proceedings of Eurospeech*, September 2003.
- [3] D. Garcia-Romero, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Support vector machine fusion of idiolectal and acoustic speaker information in spanish conversational speech," *In Proceedings of ICASSP*, April 2003.
- [4] L. Ferrer, H. Bratt, V. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," *In Proceedings of Eurospeech*, pp. 2017–2020, September 2003.
- [5] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot, "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.
- [6] SPRO:Speech Signal Processing Toolkit, "<http://www.irisa.fr/metiss/guig/spro/spro-4.0.1/spro.html>."
- [7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10(1-3), pp. 19–41, January/April/July 2000.
- [8] A. El Hannani and D. Petrovska-Delacrétaz, "Improving speaker verification system using alisp-based specific gmms," *In Proceedings of AVBPA*, July 2005.
- [9] Cambridge University Engineering Department. HTK: Hidden Markov Model Toolkit, "<http://htk.eng.cam.ac.uk>."
- [10] A. El Hannani and D. Petrovska-Delacrétaz, "Exploiting high-level information provided by alisp in speaker recognition," *In Proceedings of the Non Linear Speech Processing Workshop (NOLISP)*, April 2005.
- [11] A. El Hannani, D. Toledano, D. Petrovska-Delacrétaz, A. Montero-Asenjo, and J. Hennebert, "Using data-driven and phonetic units for speaker verification," *In Proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June 2006.
- [12] T. Joachims. SVMlight: Support Vector Machine, "<http://svmlight.joachims.org>."
- [13] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database," *In Proceedings of ICASSP*, April 1990.