16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AN APPLICATION CONSTRAINED FRONT END FOR SPEAKER VERIFICATION

Alexandre Preti[12], Bertrand Ravera[2], François Capman[2], and Jean-François Bonastre[1]

[1] University of Avignon, LIA
339 chemin des meinajaries, Agroparc BP 1228, 84911, Avignon Cedex 9, France
{alexandre.preti, jean-francois.bonastre}@univ-avignon.fr
[2] THALES Multimedia Processing
147 Bd Valmy, 922047, Colombes, France
{bertrand.ravera., francois.capman}@fr.thalesgroup.com

ABSTRACT

Even if the speaker recognition field is very dynamic, few studies concern the constraints linked to the use of a speaker recognition system inside a professional telecommunication network. This paper deals with this problem and proposes some adaptation of such system in the focus of a real world network monitoring application. This work is specifically dedicated to the front-end. Both real-time constraints and distributed architectures are investigated. The signal acquisition takes place on a mobile terminal, while the speaker verification process is performed on a remote server. We propose a frame-by-frame on-line processing for feature extraction, frame selection and normalization. The links between the network speech coder and the speaker recognition system are also investigated, for both the ETSI TETRA speech codec (at 4600 bit/sec) and the NATO STANAG 4591 (at 2400 bit/sec). The proposed solutions are compared with a classical unconstrained front-end (off-line processing).

1. INTRODUCTION

There is a growing interest for speaker recognition technology in the context of professional and security applications. Voice authentication can be used in various types of applications: secured access to telecom services, entrance control systems, forensic applications… In this paper, we focus on the security reinforcement of a professional communication network by adding an on-line vocal-based identity monitoring. This application allows assessing a nominal use of the mobile terminals in the network. As soon as a network intrusion is detected, a human operator can be alerted in order to check the user's identity, or to inhibit the corresponding terminal. This type of functionality is particularly well-suited for professional networks used by security and rescue forces. In the context of real scenarios, on-line processing and short-delay decision are required in order to quickly react to an impostor attack. The specificities of professional communication networks have to be taken into consideration, including the speech coding solutions and the distributed architecture of the processing (cf. figure 1). In this study, we are considering different configurations of the front-end processing. A first configuration consists in the extraction of features on a remote platform either from the decoded speech signals or through the conversion of internal speech coder parameters. An alternative solution is based on the transmission of optimised parameters directly extracted on the terminal, as inspired by the ETSI Aurora standard [1] designed for distributed speech recognition applications. In this study, we are considering the following standardised vocoders: TETRA [2] and MELP [3]. Both coders are using linear prediction of speech allowing simple extraction of Cepstrum features from linear prediction coefficients (LPCC).

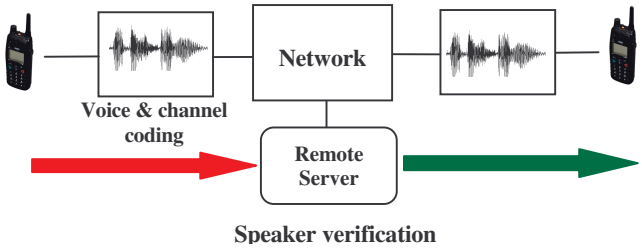

Figure 1 – Professional communication network monitoring.

For a real-time, on-line, monitoring of the communications, it is necessary to have a specific implementation of the speaker verification system when compared to a classical speaker recognition solution, designed for off-line processing. This is particularly relevant to the front-end processing and the scoring stage.

This paper introduces a complete real-time compliant front-end based on the ETSI Aurora standard. The speaker verification baseline system is described in section-2. The reference front-end is detailed in section-3. Section-4 deals with the communication network constraints. The on-line real-time compliant front-end specifications are detailed in section-5. And finally some experiments and results are provided in section-6, followed by some conclusive remarks.

2. THE SPEAKER VERIFICATION SYSTEM

The speaker recognition system is developed using ALIZE/SpeakerDet[1] toolkit developed by the LIA [4,5]. It is based on statistical modeling using Gaussian Mixture Model (GMM) [6]. A Universal Background Model (UBM) is estimated via the EM algorithm, using several hours of recorded

[1] http://mistral.univ-avignon.fr

data. As it is done off-line, there is no need for a real-time front-end processing. Speaker models are adapted from the UBM via a Maximum A-Posteriori (MAP) adaptation procedure [6]. The UBM and speaker models are composed of 512 Gaussian components with diagonal covariance matrices. No score normalization is performed in the presented results.

3. REFERENCE FRONT END PROCESSING

As in most of state-of-the-art speaker recognition systems, we are using cepstral analysis of the speech signal for the feature extraction step, as depicted in figure 2. A preliminary frame selection step based on Voice Activity Detection (VAD) is used to discard useless frames when detected as silence or noise [7]. The VAD is based on statistical modelling of the energy distribution. Finally a feature normalization step based on Cepstral Mean and Variance Normalization (CMVN) is used to remove some channel effects.

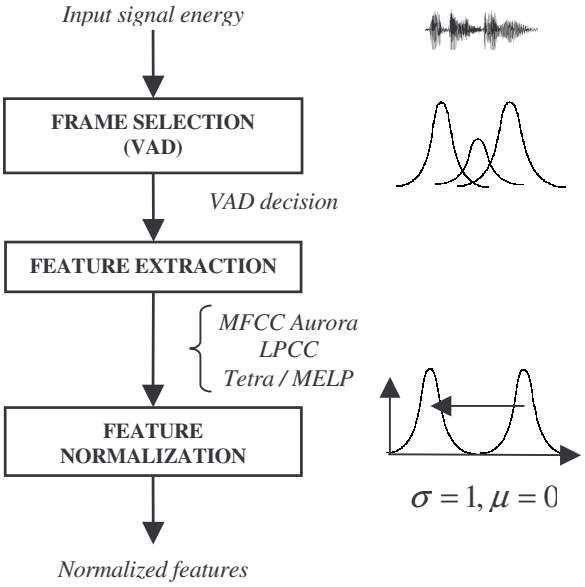

Figure 2 - Structure of speaker recognition front-end

These three steps are most of the time based on a file entry and use the whole utterance to estimate the needed parameters.

3.1 The ETSI Aurora standard

The ETSI Aurora standard was originally designed for Distributed Speech Recognition (DSR) systems, meaning that the processing is distributed between the terminal and the network. The terminal performs the feature extraction and the associated compression processes. The resulting compressed features are then transmitted through the network to a remote back-end recognizer. Performance degradations resulting from the transcoding on the voice channel are therefore removed. The Aurora features consist of 13 static Mel-scaled filter-bank derived cepstral coefficients and a log-energy coefficient, computed every 10-ms frame. The feature vector is then extended by adding first and second derivatives of cepstral coefficients. The Aurora quantization stage is then applied to the cepstral parameters.

3.2 Frame Selection

The frame selection step discards useless frames - such as noise or weak level speech - since they significantly decrease the speaker recognition performance [7]. A classical energy-based frame pruning is applied as a VAD. It is based on a 3-Gaussian modelling of the frame energy distribution (this modelling is done file by file), and the selected frames are the one corresponding to the Gaussian with the highest mean value.

3.3 Feature Normalization

The parameters are finally normalized, file by file, using a standard cepstral mean subtraction and variance normalization. It reduces the stationary convolution noises due to the channel and the mismatch between training and testing conditions. The best recognition results are obtained when mean and variance parameters are averaged along a whole utterance.

4. COMMUNICATION NETWORK CONSTRAINTS

Within the framework of professional telecommunication networks it is necessary to take into account the whole transmission chain. The architecture of a typical network is described in figure 3.

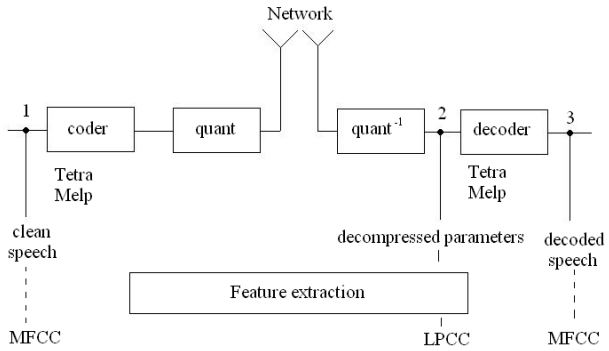

Figure 3 – Input signal taken from different points inside the network architecture

The feature extraction process can take place at three different locations in the network. They can be extracted on the terminal (point 1 in figure 3). But in this case, they must be encoded and transmitted as in the Aurora standard, which is bandwidth consuming since simultaneous encoded speech for communication should also be transmitted. However, this configuration will be considered as the reference configuration in terms of performance, since the original speech signal is used. The standard configuration (point 3 in figure 3) will consist in extracting the front-end parameters from the re-synthesized speech at the output of the low-bit rate speech coded used for communication. And finally, an optimised configuration (point 2 in figure 3), which does not require the re-synthesized speech, will perform the feature extraction in the compressed domain from the transmitted bit stream. As the considered low bit-rate speech coders (TETRA and

MELP) are based on Linear Prediction Analysis, it is possible to use directly the prediction coefficients for speaker recognition using a modified measure, like proposed in [8]. In our work, we simply take advantage of the available LPC-related coefficients (LSP/LSF) to extract LPC-based Cepstral Coefficients (LPCC) in order to minimize the differences with a classical state-of-the-art speaker recognition front-end. The following equation (1) gives the computation of LPCC from the Linear Prediction coefficients resulting from the dequantization stage of the speech decoder.

$$\hat{c}_1 = -\alpha_1$$
$$\hat{c}_n = -\alpha_n - \sum_{m=1}^{n-1}(1-\frac{m}{n})\alpha_m \hat{c}_{n-m} \qquad (1 < n \le p) \qquad (1)$$

5. APPLICATION CONSTRAINED FRONT-END

This section describes the proposed modifications in order to achieve on-line processing. It mainly concerns the Voice Activity Detection, the Feature Normalization process and the Feature Extraction itself. The work is done in the focus of a professional telephonic network, using the reference ETSI Aurora standard, and the TETRA and MELP coders.

5.1 Voice Activity Detection

As seen in section-3, the reference speaker recognition system relies on an energy-based frame selection process, which shows a drastic impact on the system performance [7]. This process is usually based on a file by file GMM modelling of the frame-energy distribution, associated with a decision process. In our application, this file-based processing is no more possible and the energy-based frame selection should be done on a frame-by-frame basis. We propose to use the VAD which is part of the ETSI Aurora standard, since it is easily accessible as a reference VAD, and also provides an on-line decision for each frame, within a reasonable delay (6-frame buffer) for the targeted applications.

The Aurora VAD is based on the energy acceleration measured on different subbands of the spectrum: across the whole spectrum and over a sub-region most likely to contain the fundamental pitch. Spectral variance measure is also computed on the lower half of the spectrum. The class information decision uses Zero Crossing Rate, VAD decision and threshold on upperband energy to determine voicing class. More details can be found in [1].

A preliminary experiment shows that the Aurora VAD selects more frames (80 %) than the classical off-line selection process (50 %). Since the speaker recognition performances are significantly affected by the inclusion of non-speech or low energized frames, we use the voicing level information (unvoiced, mixed-voiced, and fully-voiced) provided by the Aurora standard in order to improve the selection process. This is in accordance with previous results showing the predominance of voiced frames on unvoiced frames for the speaker recognition task. This voicing information is used in order to reduce the VAD selected frames to a fully-voiced subset.

5.2 Feature Cepstral Mean and Variance Normalization

For each dimension of the feature space, the coefficients are normalized using a CMVN function to obtain a zero-mean and a unit-variance distribution of the parameters. This normalization requires both mean and variance estimators which need to be robustly estimated. In the reference front-end, as the recordings are long enough, these estimators are obtained from the entire file. This file-based solution does not fit the targeted on-line constraints. Some sliding window-based solutions, have been shown to perform as well or even slightly better than the file-based CMVN [9, 10].

Following the similar approach, we have implemented a frame normalization based on a first-order forgetting process [10, 11]. This procedure is using a parameter initialization window of N frames (only the frames selected by the VAD are taken into account). The frames inside this window are normalized only when the window is full, after this step the normalization is done frame by frame, without any delay. The normalization parameters are then continuously updated during the normalization process using the following equations (2, 3):

$$\hat{\sigma}_i^2 = \beta\hat{\sigma}_{i-1}^2 + (1-\beta)\sigma_i^2 \quad (2)$$
$$\hat{\mu}_i = \beta\hat{\mu}_{i-1} + (1-\beta)\mu_i \quad (3)$$

$$\begin{cases} \beta = \mathbf{1} \text{ for frame index i: [0;N]} \\ \beta = \textbf{(window size – 1) / window size} \text{ otherwise} \end{cases}$$

5.3 TETRA and MELP parameters

In this section, we evaluate the LPCC parameters directly extracted from internal speech coder parameters. We are considering for this evaluation the ETSI TETRA speech coder and the NATO STANAG-4591 MELP coder, since they are widely used in professional and military coders. We do not consider the channel effects in this study.

5.3.1 TETRA coder

The Terrestrial Trunked Radio (TETRA) coder is a widely used speech coder in Professional and Private Mobile Radio networks (PMR). The bit-rate is 4.6 kbit/s. It is based on a 10^{th}-order LPC-based analysis. Each frame is computed every 35 ms, and linear interpolation is used to generate a sub-frame every 7.5 ms.

5.3.2 MELP coder

The NATO STANAG-4591 Mixed Excitation Linear Prediction (MELP) coder is a 2.4 kbit/s vocoder. It is considered as a state-of-the-art coder for low-bit rate applications and therefore addresses the requirements of professional and military networks. The MELP coder is also based on the traditional LPC model. The analysis is based on 22.5 ms frames, and LPC interpolation is performed for each pitch period.

For both TETRA and MELP coders, the LPCC cepstral parameters are extracted directly from the encoded LPC prediction coefficients as described in section 4 (equation 1).

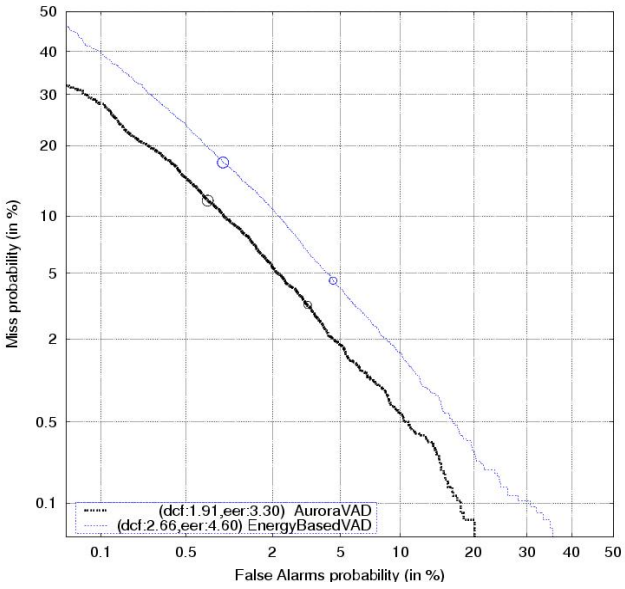

Figure 4 – DET curves for Aurora (DCF: 1.91, EER: 3.30) and Energy-based VAD (DCF: 2.66 , EER: 4.60).

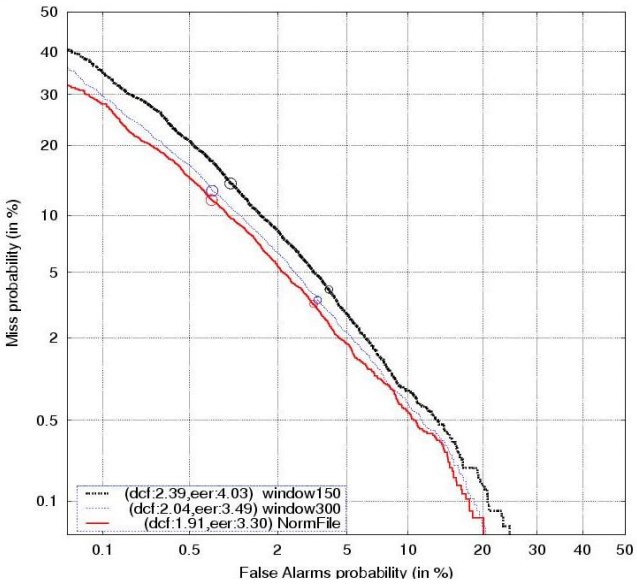

Figure 5 – DET curves for 150 and 300 frames initialization window size and the reference (file processing mode)

6. DATABASE AND EXPERIMENTS

This section presents the database used for setting up all the experiments presented bellow.

6.1 BREF database

The French database BREF [12] is used for the experiments. BREF is composed of read sentences, recorded in a quiet environment. The original 16 KHz recordings are down-sampled at 8 kHz in order to be compliant with the telephone bandwidth. The UBM training is done on a first subset of 40 speakers which consist in a 8 hours dataset before the frame selection step. A second set of 40 speakers (20 male and 20 female) is used for the target speakers. Finally, a third set composed by 35 other speakers is used for the impostor tests. No cross gender tests are performed, but one should note that gender detection could easily be included. This setup gives a total of about 8 thousand true target trials and about 90 thousands impostor trials. In order to be as close as possible to the application constraints, the training speech segment have a duration of 1 minute and 8 seconds duration files are used for the tests. The performance is evaluated through classical DET performance curves. Equal Error Rate values (False Acceptance equals False Rejection) and scores in terms of NIST Decision Cost Function (DCF) are also provided.

6.2 VAD experiments

This section evaluates the proposed Aurora VAD, using fully-voiced class information. and. We use the feature extraction process as described in section 3. The performance of the proposed VAD are provided in figure 4 together with the results of the reference file-based VAD (denoted energy-based VAD), for comparison purpose.

The best performance is obtained with the Aurora-based frame-by-frame VAD, which allows on-line processing. Using the "fully-voiced class" Aurora VAD, a 28 % relative improvement is achieved for both EER and DCF evaluation measures.

6.3 Normalization experiments

In this section, we use the on-line normalization procedure presented in section 5.2 to process both the speaker training and the test data, while an off-line processing is still used for the UBM training. Two different experiments with two different initialization window sizes (150 and 300 frames) are performed.

Figure 5 presents the results for the file-based normalization and for the frame-by-frame normalization, with 150 and 300 frames initialization windows. The results obtained with a 300 frames are very close to the ones obtained with the file-based normalization mode (an EER of 3.49% for the one-line processing to be compared with 3.30% for the file-based solution).

Howether, as the test segment duration is only 8 seconds, which gives in average about 500 VAD-selected frames, the 300 initialization window contains 60% of the total of selected frames.

6.4 TETRA/MELP compressed domain experiments

In this section, we evaluate the potential use of acoustic parameters extracted from the compressed domain, i.e. directly from the coder parameters. For all the experiments based on internal speech coder parameters, all the audio recordings (UBM and speaker training data, and testing data) are encoded using the TETRA and MELP coders. The experiments are done using LPCC cepstral parameters as described in section 4 and the reference VAD described in section 3. It should be noted that there is no Mel-frequency scaling in the LPCC extraction process. For comparison purpose, we also propose several experiments where the acoustic parameters are extracted from the coder bit-stream after decoding the speech (point 3 in figure 3). In this case the cepstral analysis is performed on the re-synthesized speech using the reference

front-end described in section 2.1. In order to evaluate the impact of the quantization losses for both coders, we propose an experiment without including the quantization stage for both coders. Finally, the baseline performance is provided for comparison, using directly the clean speech and the reference front-end (point 1 in figure 3). The results are presented in Table 1.

First we could notice that the MELP coder outperforms the TETRA coder in terms of DCF and EER when the features are extracted from the re-synthesized speech at the output of the decoders. These results are in discordance with the performance obtained in terms of speech quality restitution. In fact, the PESQ measure (Perceptual Evaluation of Speech Quality) averaged on the whole test database is superior for the TETRA coder (2.953) than for the MELP (2.945). As the internal LPC analysis is the same for both coders results are equivalent for MELP and TETRA coders when extracting LPCC directly from the LPC parameters. Looking to both results, we could make the assumption that the MELP re-synthesis adds useful information for cepstral analysis.

Secondly, the quantization loss is of 18 % relative for both DCF and EER (about 4.5% of EER before the quantization step to be compared with 5.5% of EER after the quantization).

Lastly, when using the quantized coder parameters (level 2 in the network architecture), the relative loss compared to the clean speech is abut 20% in terms of EER for both coders (4.58% of EER for the clean speech and about 5.5% for the coder-based acoustic parameters).

Experiment	DCF	EER
LPCC Tetra (2)	3.55	5.57
LPCC Melp (2)	3.55	5.58
Aurora front-end on Tetra decoded speech (3)	3.84	7.29
Aurora front-end on Melp decoded speech (3)	2.94	5.20
Aurora front-end clean speech (1)	2.78	4.58
LPCC Tetra without quantization	2.9	4.55
LPPC Melp without quantization	3	4.56

Table 1 – Results in terms of DCF and EER for recognition experiments using different feature extraction methods (number in brackets refers to figure 3).

7. DISCUSSION AND FUTURE WORK

In this work, we focused on the constraints linked the use of speaker recognition inside a professional telecommunication network. Our main interest was the front-end processing under the real-time online constraints. We have proposed a complete speaker verification front-end for on-line processing. Three points were investigated: the VAD, the frame normalization process and the acoustic parameter extraction inside the telecommunication network architecture.

The proposed Aurora-based frame-by-frame VAD gives a 28 % improvement on recognition results compared to a standard energy-based VAD. The introduced normalization procedure performs nearly the same as file-based normalization.

Moreover we have analysed the impact of different feature extraction on decoded speech and in the coder compressed domain. The cepstral feature extraction based on the LPC analysis of the TETRA coder performs the best on TETRA coded speech, whereas extracting feature on decoded speech performs better for MELP coded speech. These results encourage the set up of a real-time on-line speaker recognition process with a negligible lost in terms of speaker recognition performance.

Future works would be more focused on a decision step suitable for a telecommunication network monitoring application. Particularly, determining when a robust decision could be done, taking into account the amount of buffered frames as well as the quality of the accumulated data (for example in terms on SNR) is of a critical importance.

REFERENCES

[1] Aurora ETSI ES 202 212 V1.1.2 (2005-11).

[2] Tetra ETSI EN 300 395-1 v1.3.1 (2005-06)

[3] L.M Supplee, R.P. Cohn, J.S. Collura, and A.V. McCree, "MELP: the new federal standard at 2400 bps," *in Proc IEEE ICASSP*, Munich, Germany, April 1997, vol.2, pp. 1591-1594.

[4] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve and J. Mason, " ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," *in Speaker Odyssey*, South Africa, January 2008.

[5] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition, " *in ICASSP*, 2005.

[6] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification, " *EURASIP Journal on Applied Signal Processing,* vol. 4, pp. 430–451, 2004.

[7] L. Besacier, J-F. Bonastre, C. Fredouille, " Localization and selection of speaker-specific information with statistical modelling," *Speech Communication*, vol. 31, pp.89-106, 2000.

[8] M. Petracca, A. Servetti, J.C De Martin, "Performance analysis of compressed-domain automatic speaker recogni tion as a function of speech coding technique and bit rate," *in ICME* 2006.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification, " *in Proc. ISCA Workshop on Speaker Recognition : A Speaker Odyssey*, June 2001.

[10] P. Pujol, D. Macho, C. Nadeu, "On Real-Time Mean-and-Variance Normalization of Speech Recognition Features, " *in ICASSP 2006 Proceedings*. Toulouse France, 2006.

[11] D. Mauler & R. Martin, "Noise power spectral density estimation on highly correlated data," *in IWAENC* 2006, Paris, 2006.

[12] L. Lamel, J.-L. Gauvain, M. Eskenazi, "BREF, a large vocabulary spoken corpus for French", i*n EUROSPEECH,* 1991, 505-508.