Fusion Strategies for Distributed Speaker Recognition using Residual Signal Based G729 Resynthesized Speech

Dalila Yessad  
LCPTS, Speech Communication and Signal Processing Lab.  
Faculty of Electronics and Computer Sciences, USTHB, P.O. Box 32 El Alia, Bab Ezzouar, 16111, Alger  
yessad.dalila@gmail.com

Abderrahmane Amrouche  
LCPTS, Speech Communication and Signal Processing Lab.  
Faculty of Electronics and Computer Sciences, USTHB, P.O. Box 32 El Alia, Bab Ezzouar, 16111, Alger  
amrouche@usthb.dz

Abstract—With the development of VoIP (Voice over IP) service, there is an emerging need to speech compression, particularly for digital speech communication and biometric speaker recognition (SR) systems. This paper presents results issued from Universal Background Gaussian Mixture Model (GMM UBM) based SR system, that is trained and tested on clean and G729 resynthesized speech. To overcome the performance loss due to the G729 codec, residual signal extracted from clean and G729 resynthesized database is used. To get better the performance, we investigated score fusion strategies based on Logistic Regression (LR). The first fusion based on GMM UBM score using LFCC (Linear Frequency Cepstrum Coefficients) and LFCC extracted from LP (Linear Prediction) residual signal. The second used the LFCC extracted from G729 resynthesized speech and its LP residual signal. The best performance is obtained by Logistic Regression (LR) fusion. The correct rate in the first case is 95% based baseline system and 83% based G729 resynthesized speech in the second case. The obtained results, using TIMIT database, have proven the efficiency of data fusion techniques for automatic speaker recognition.

Keywords—VoIP, G729, LP residual, Score fusion, Logistic Regression (LR) fusion.

I. INTRODUCTION

Recently, with the explosive growth of the Internet and VoIP (Voice over IP) applications, the use of speech recognition technology in digital speech communication systems has lured the researchers to develop efficient speech coding techniques. Speech coding has an important role in modern voice-enabled technology, particularly for digital speech communication, where quality and complexity have a direct impact on speech recognition system. Generally, speech coding is a procedure to represent a digitized speech signal using as few bits as possible, maintaining at the same time a reasonable level of speech quality. There are many speech coding standards designed to suit the need of a given application. The commonly used VoIP codecs are G.711, G.729 and G.723.1, which are standardized by the ITU-T in its G-series recommendations.

In [1] the effect of speech coding on speaker identification (SI) was presented. In order to analyzing speaker identification system using both vocal track and excitation source features, the study is based on three kinds of coder’s as; GSM full rate (ETSI 06.10), CELP (FS-1016), and MELP (TI 2.4 kbps). The results present a significant reduction of performance in SI system due to the coding algorithms. The influence of G729 speech coding on ASR (Automatic Speaker Recognition) in VoIP applications was studied in [2], where the ITU-T G.729 speech coder is used to encode and decode the speech input for text independent speaker recognition over IP networks. Speaker recognition system was designed to use three kinds of features, the first one is the LFCC coefficients extracted from clean database, the LFCC features vector extracted from the G729 transcoded database (resynthesized speech), and LPCC (Linear Predictive Cepstral Coefficients) calculated directly from the coded parameters embedded in the ITU-T G.729 bits-stream. Experiments were performed using the TIMIT database, and the effect of the G729 on speaker recognition performance is investigated. In order to improve recognition accuracy of GMM-UBM used with the G729 codec, speaker recognition using T-norm score normalization was also examined. In [3], speaker verification system based on G729 and G723.1 coder parameters and handset mismatch compensation is studied. The adding of the LPCC extracted from the LP residual to the coder based LPCC achieves the best result and outperforms the technique that use only the coder derived LPCC. Cepstral features [4], such as MFCC (Mel Frequency Cepstral Coefficients) and LPCC, have been the dominant features for a long time in speaker recognition. These features are believed to provide pertinent cues for phonetic classification and have been successfully implemented in most existing speaker recognition systems [5]. In [6], the standard procedures for extracting MFCC and
LPCC features were applied to LP residual signals, resulting in a set of residual features for speaker recognition. In [7], we presented a study of SVM (Support Vector Machine) based GMM supervector speaker recognition using LP residual signal. MFCC and LPCC features are generated from LP residual of speech signal, and then compared with their calculation directly from speech samples. Universal Background Gaussian Mixture Models (GMM UBM) and Gaussian Superscalar (GMM SVM) based speaker modeling principles have been used. Experimental results, using ARADIGITS database, show the efficiency of the GMM SVM based approach associated with feature vectors issued from LP residual signal. From this we can conclude that the MFCC and LPCC features based LP residual contain useful speaker informations for speaker recognition applications, and the cepstral features provide additional informations in speaker recognition.

This paper deals with the LFCC feature based on LP residual signal applied to G729 resynthesized speech over IP. Section 2 presents LP residual signal, section 3 provides feature extraction technique based on linear frequency cepstral coefficient (LFCC). Then, Sections 3 elaborates speaker feature extraction technique based on linear frequency cepstral coefficient. From this we can conclude that the MFCC and LPCC features based LP residual contain useful speaker informations for speaker recognition applications, and the cepstral features provide additional informations in speaker recognition.

II. LINEAR PREDICTION (LP) RESIDUAL

Linear Predictive Coding (LPC) is the process of predicting future sample values of a digital signal from a linear system. It is therefore about predicting the signal at instant n from p previous samples as:

$$s(n) = - \sum_{k=1}^{p} a_k s_{n-k} + G e(n)$$  \hspace{1cm} (1)

Where $$a_1, a_2, \ldots, a_p$$ are the Linear Predictive Coefficients (LPCs), $$p$$ is the model order, $$G$$ and $$e(n)$$ are respectively the excitation gain and source. The LPC features are derived adaptively for each 20-30 ms speech frame by minimization of excitation mean square energy. For simplicity, we will assume that the order of LP model is uneven, $$p = 2m - 1$$. The LPC spectrum or the transfer function of the LP filtering is defined:

$$H(z) = \frac{G}{A(z)}$$  \hspace{1cm} (2)

Where

$$A(z) = 1 - \sum_{i=1}^{2m-1} a_k z^{-1}$$  \hspace{1cm} (3)

The linear prediction residual signal, according to the LPC model, is presented by the error between the actual value $$s(n)$$ and the predicted value $$\hat{s}(n)$$. It is given by:

$$e(n) = s(n) - \hat{s}(n)$$  \hspace{1cm} (4)

Where

$$\hat{s} = - \sum_{k=1}^{p} a_k s_{n-k}$$  \hspace{1cm} (5)

LP residual signal might contain information which has not been captured by the LPC coefficients and which can be useful for the speaker recognition task.

III. LINEAR FREQUENCY CEPSTRAL COEFFICIENTS (LFCC)

Feature extraction is an important step for speaker recognition systems. In this paper, we extract LFCC from LP residual of speech signal, instead their calculation directly from speech samples.

The generation of the LFCC decomposes in six steps:

- Step 1: Cut up the signal in several overlapping windows;
- Step 2: In order to decrease the spectral distortion a Hamming windowing is applied to signal frames;
- Step 3: Apply the FFT;
- Step 4: The Linear frequency scale is then applied;
- Step 5: Apply the log after the Linear scale;
- Step 6: Finally the discrete cosine transform (DCT) of the output signal is formed.

IV. GMM UBM MODEL AND SCORE FUSION

A. Speaker Recognition using GMM UBM Model

The GMM-UBM approach (see Fig.1) is the state of the art system in text- independent speaker recognition [8]. This approach is based on a statistical modelling paradigm, where a hypothesis is modelled by a GMM model:

$$p(x/\lambda) = \sum_{i=1}^{m} w_i N(x|\mu_i, \Sigma_i)$$  \hspace{1cm} (6)

Where, $$w_i, \mu_i, \Sigma_i$$ are, respectively the weights, the mean vectors, and the covariance matrices (generally diagonal) of the mixture components.

During the test, the system has to determine whether the recording $$y$$ was pronounced by a given speaker $$S$$. This question is modelled by the likelihood ratio:

$$\frac{p(y|\lambda_{hyp})}{p(y|\lambda_{S})} \geq r$$  \hspace{1cm} (7)

Where $$y$$ is the test speech recording, $$\lambda_{hyp}$$ is the model of the hypothesis where $$S$$ pronounced $$y$$, $$\lambda_{S}$$ is the model of the negated hypothesis ($$S$$ did not articulate $$y$$), $$p(x|m)$$ is the GMM likelihood function, and $$r$$ is the decision threshold.

The model $$\lambda_{hyp}$$ is a generic background model, the so-called UBM, and is usually trained during the development phase using a large set of recordings coming from a large set of speakers. The model $$\lambda_{hyp}$$ is trained using a speech record obtained from the speaker $$S$$. It is generally derived from the UBM by moving only the mean parameters of the UBM, using a Bayesian adaptation function.
In this study, the GMM-UBM system is the LIA SpkDet system [9] based on the ALIZE platform3 and distributed under an open source license. This system produces speaker models using MAP (Maximum A Posteriori) adaptation by adapting only the means from a UBM with a relevance factor of 14. The UBM component was trained on a selection of 60 corpus. For all the experiments, the model size is 128 and the performances are assessed using DET (Detection Error Trade-off) plots and measured in terms of Equal Error Rate (EER).

B. Logistic Regression (LR) fusion

To increase at more accurate and reliable decision of speaker recognition system, fusion is typically implemented as a linear combination of the subsystem scores. Parameters of the linear model are commonly estimated using the logistic regression method. In this work, we investigated the logistic regression fusion method to combine scores used from GMM UBM model based TIMIT and G729 TIMIT database.

Logistic regression [10] is a probabilistic linear model, which start with the realization that target class posterior can be modeled as:

\[
p(y_t | s) = \left( 1 + \exp\left( - (w^T s + w_0) \right) \right)^{-1}
\]

(9)

Where \( w \) is a vector of the linear model weights. The fused target and non-target scores are

\[
C_{w, r}(w, s) = \frac{p}{N_t} \sum_{i=1}^{N_t} \log \left( 1 + e^{-K_i \logit P} \right) + \frac{1-p}{N_f} \sum_{j=1}^{N_f} \log \left( 1 + e^{R_i \logit P} \right)
\]

(11)

Where the fused target and non-target scores are respectively:

\[
K_i = \beta_0 + \sum_{i=1}^{N_t} \beta_i Y_i
\]

(12)

\[
R_i = \beta_0 + \sum_{j=1}^{N_f} \beta_j U_i
\]

(13)

\[
\logit P = \log \frac{p}{1-p}
\]

(14)

Where \( V \) and \( U \) are vectors of target and non target scores of the \( N \) components for \( N_t \) target and \( N_f \) non target trials, \( \beta \) are weights of fusion.

V. PERFORMANCE EVALUATION

A. Speech Database

In this work we investigate TIMIT database to corroborate our experiences. The waves corresponding to the SI sentences are used for training each speaker model. 504 speakers of the database (168 women and 336 men) are explored to building speaker models. In the test step, five SX sentences of every speaker (112 women and 56 men) are tested separately (112x5+56x5=840 test patterns of second each, in average). The experiments are totally text independent. The remaining 60 speakers of the database are used to train the world model needed for the speaker verification experiments. 840 client accesses and 840 impostor accesses are prepared (for each client access, an impostor speaker is randomly chosen).

B. G729 Resynthesized database

TIMIT database is treated under G729 codec, the signal waveforms of each speaker is resynthesized by G729 (8kbits/s) speech codec, the database was encoded by G729 coder in the client part. The G729 encoder operates on speech frames of 10ms corresponding to 80 samples of 16 bits at a sampling frequency of 8 KHz. The speech signal is analyzed in each frame to extract the coefficients of Linear Prediction (LP) of the 10th order, which are converted into Line Spectral Pairs (LSP) digitized at 18 bits per predictive quantification vector. By following, in attendance other parameters are estimated from the residual error signal of linear prediction on the basis of sub-frames with 40 samples, or 5ms. The CELP model are encoded and transmitted in bit-stream to the server, were the G729 decoder used to reconstruct the speech by filtering the excitation through the short term synthesis filter based on a 10th order Linear Prediction (LP) filter. The reconstructed speech is enhanced by a post filter [11]. For the rest of the this paper we use the designation G729TIMT to described TIMT database used in bit-stream after G729 encoded in the client
part, then transmitted under UDP (User Datagram Protocol) protocol to the server, to be resynthesized by G729 decoder over IP.

C. Features Extraction

Speaker utterances issued from TIMIT or G729TIMIT database, were represented by 19 Linear Frequency Cepstral Coefficients (LFCC) determined through linear filterbank analysis, with their first derivatives and the delta energy. Altogether, a 40 coefficients vector is extracted from clean (TIMIT), G729TIMIT (G729 resynthesized speech) and LP residual based both clean and G729TIMIT database. Mean subtraction and variance normalization were applied to all features. In this work, the LFCC features are derived under different conditions, four experiments are adopted and named as; expA, expB, expC and expD:

- expA: LFCC features obtained from the front clean TIMIT corpus.
- expB: LFCC computed from the G729TIMIT database (G729 resynthesized speech).
- expC: LFCC derived from residual of TIMIT database.
- expD: LFCC features derived from residual used G729TIMIT database.

Figure 2 show the speech waveforms and the corresponding LP residual signals, of the vowel /a/ from the sound, expressed by two different male speakers from TIMIT database.

We can see the differences between the two segments of residual signals. In addition to the difference between their pitch periods, the magnitudes of the secondary pulses of speaker B are relatively high.

D. Experimental Results

We evaluate the speaker recognition performances of LFCC based clean speech (expA) like baseline system, using GMM-UBM recognizer. In order to test the robustness of the system to speech coding distortion, we describe the experiment B (expB), then, we extract the LFCC from LP residual signal in experiment (expC) using TIMIT database. G729TIMIT database is also investigated and used to extract the LP residual signal in experiment D (expD). The obtained performances in expA show a high rate correct recognition to 91%, but in expB we observed a significant rate reduction to 74%, compared with the baseline system (expA), due to G729 effect (see figure 4).
In order to reduce the losses due to G729 codec, we exploit LP residual in expC and expD, this last obtained a correct rate to 73%. In average, it shows the same result compared with expB to 74%. The investigation of LP residual signal based TIMIT speech, in expC, obtained correct rate to 90.8%. The results are shown in figure 5.

These results confirmed the efficiency of LP residual signal, which contains useful information about the source excitation, and vocal tract system represented by an all-pole filter. However, the performance recognition based resynthesized database is poorer than that achieved by features derived from clean dataset. In order to perform recognition performance of GMM-UBM with the G729 transcoded databases and overcome the drawback of G729 codec, we exploit the Logistic Regression (LR) as fusion strategy. The scores obtained from GMM-UBM model are fused and presented in experiment E (expE) and experiment F (expF):

- expE: Present the result of LR score fusion of both scores obtained in expA and expC.
- expF: Represent the result of LR score fusion based scores resulted from expB and expD.

The results are presented in figure 6.

The correct recognition rate achieves by expE is 94% and 83% by expF. It presents a best results and we can conclude that the fusion strategy achieves the best performances, and outperform expA, expB, expC and expD. Here, we investigate LFCC extracted from LP residual to characterize the time-frequency characteristics of the pitch pulses, because in frequency domain, the useful temporal information, the amplitudes and the time locations of pitch pulses, are not represented in the Fourier spectra of LP residual. Then the LR fusion strategy is adopted to overcome the drawback of G729 codec.

VI. CONCLUSION

In this paper, we investigated the logistic regression fusion technique of scores issued from GMM-UBM based speaker recognition. This technique is applied to scores based GMM-UBM model using both LFCC and LFCC extracted from LP residual of TIMIT or G729 TIMIT over IP.

In order to evaluate the degradation in performance introduced by some aspects of the codec, the recognition performance when extracting LFCC features from G729TIMIT speech was measured, and it was compared with LFCC extracted from original speech. Thus, experiments were carried out, using the LFCC based LP residual speech,

The performance derived from resynthesized speech is still poorer than that achieved by features from clean database. In order to perform the score of GMM-UBM with the G729 resynthesized database, logistic regression fusion strategy has been contributed. The fusion method improved a correct rate from 74% to 83%, and achieved best results compared to speaker recognition over IP using resynthesised speech without LR fusion.
REFERENCES


