Automatic Speaker Recognition using G729 Resyntesized Speech Over IP

D. YESSAD and A. AMROUCHE
Speech Communication and Signal Processing Laboratory,
Faculty of Electronics and Computer Sciences, USTHB,
P.O. Box 32, El Alia, Bab Ezzouar, 16111, Algiers, Algeria
Email: yessad.dalila@gmail.com
Email: namrouche@usthb.dz

***Abstract*— This paper presents speaker recognition based G729 compressed speech over IP networks, the ITU-T G729 codec is investigated to encode and decode the input speech. Two methods of speech parameterization namely Linear Predictive Coding Cepstral Coefficients (LPCC) and Linear Frequency Cepstral Coefficients (LFCC) have been used as feature extractor. Speaker recognition system was designed to use those features obtained from two kinds of database: clean and G729 resynthesized speech. Experiments were performed using TIMIT database, and the effect of the G729 on GMM-UBM speaker recognition system is studied. Finally results for GMM-UBM model based both features and databases are compared and explained. It has been observed that at clean database, the system given a recognition accuracy at 89% and 87% using LPCC and LFCC successively. However the performance of both features degrades more rapidly to 80% and 77% under G729 resynthesized speech.**

***Index Terms*— Universal Background Model (GMM-UBM), LFCC, LPCC, G729, VoIP, Resynthesized speech.**

I. INTRODUCTION

With the explosive growth of the internet and VoIP (voice over IP, or VoIP) applications, speech compression has lured the researchers to develop techniques around speech coding concept. where quality and complexity have a direct impact on speech recognition system. In general, 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. 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. The use of speech recognition technology in digital speech communication systems, especially in VoIP applications, is one of the major goals over the last years. There has been increasing interest in the performance of the automatic recognition of resynthesized coded speech [1]. Speaker verification based on GSM, G.729, and G.723.1 resynthesized speech was studied in [2]. It was shown that recognition performance generally degrades with coders bit rate. in [3] G729 Coded Parameters Under Matched and Mismatched Conditions for Distributed Speaker Recognition is studied. In [4], techniques that require knowledge of the coder parameters and coder internal structure were proposed to improve the recognition performance of G.729 coded speech. However, the performance is still poorer than that achieved by using resynthesized speech. The main goal of this work is to study the influence of G729 based the automatic speaker recognition using VoIP communications systems. We are particularly focused on the performance recognition obtained with the G729 resynthesized speech, which is the development key of VoIP application in speech technologies. In this work, the ITU-T G.729 speech coder is used to encode and decode the speech. Experiments were performed over the TIMIT corpus.

The rest of this paper is organized as follows. The G729 speech coder is explained in section 2. LPCC and LFCC features extracted from original and resynthesized database are given in section 3. The speaker recognition system used in all the experiments is presented in section 4. Section 5 present the experimental results. Finally the conclusion is drawn in section 6.

II. G729 SPEECH CODEC

The G729 codec, also known as CS-ACELP (Conjugate Structure Algebraic Code Excited Linear Prediction), is specified by the ITU (International Telecommunications Union). It compresses speech from 16 bit, 8 kHz samples (128 kbps) to 8 kbps, and was designed for cellular and networking applications. 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 [5].

III. FEATURES EXTRACTION

The details of the two parameterization methods presented in this section have been given below:

A. Linear Predictive Cepstral Coefficients (LPCC)

Linear prediction, is the process of predicting future simple values of a digital signal from a linear system. It is therefore about predicting the signal at the instant n from the p previous samples (equation 1). So the coding by linear prediction consists in determining coefficients a_k that minimize the error $e(n)$.

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n) \tag{1}$$

LPC are believed to give very accurate formant information of acoustic signals. Both LPC and cepstrum coefficients are widelyused in speech and speaker recognition applications. In this work the cepstrum coefficients $\{ceps_q\}_{q=0}^{Q}$ can be estimated from the LPC coefficients $\{a_q\}_{q=1}^{p}$ using a recursion procedure:

$$ceps_q = \begin{cases} ln(G), & q = 0 \\ a_q + \sum_{k=1}^{q-1} \frac{k-q}{q} a_k ceps_{q-k}, & \text{1qp} \\ \sum_{k=1}^{p} \frac{k-q}{q} a_k ceps_{q-k}, & p < \text{qQ} \end{cases} \tag{2}$$

Where G is the gain term in the LPC model, p the LPC model order, and $Q+1$ the number of cepstrum coefficients.

B. Linear Frequency Cepstral Coefficient (LFCC)

The generation of Linear Frequency Cepstral Coefficients (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 aapplyed 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 then formed.

IV. Gaussian Mixture Model Universal Background (GMM-UBM)

The speaker recognition system is a Gaussian mixture model-universal background. The GMM-UBM (see figure 1) approach is the state of the art system in text- independent speaker recognition [6]. This approach is based on a statistical modeling paradigm, where a hypothesis is modeled by a GMM model:

$$p(x|\lambda) = \sum_{i=1}^{i=m} \alpha_i N(x|\mu_i, \sum_i) \tag{3}$$

Where α_i, μ_i and $\sum_i$ respectively, the weights, the mean vectors, and the covariance matrices (generally diagonal) of the mixture components. During a test, the system has to determine whether the recording Y was pronounced by a given speaker S. This question is modeled by the likelihood ratio;

$$\frac{p(x|\lambda_{hyp})}{p(x|\lambda_{\overline{hyp}})} \geq \tau \tag{4}$$

Where Y is the test speech recording, λ_{hyp} is the model of the hypothesis where S pronounced Y, $\lambda_{\overline{hyp}}$ corresponds to the model of the negated hypothesis (S did not pronounce Y), $p(y|m)$ is the GMM likelihood function, and τ is the decision threshold. The model $\lambda_{\overline{hyp}}$ 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 λ_{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.

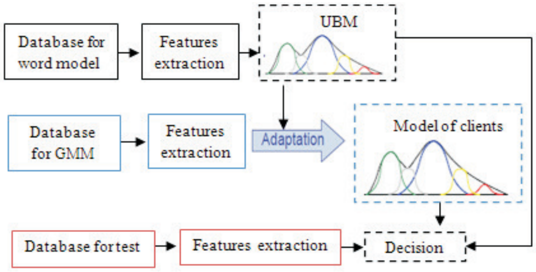

Fig. 1. GMM-UBM speaker recognition system.

In this study The GMM-UBM system is the LIA SpkDet system [7] based on the ALIZE platform3 and distributed under an open source license. This system produces speaker models using MAP adaptation by adapting only the means from a UBM. 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 plots and measured in terms of equal error rate (EER).

V. Results and Discussions

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 men 56) is 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 encodec by G729 (8kbits/s) in the client part, then transmitted to the server, to be resyntesized by G729 decoder. For the rest of the this paper we use the designation G729TIMIT to described the resyntesized TIMIT data base, which is encoded G729 and saved in bit-stream format, to be transmitted under UDP (User Datagram Protocol) protocol from the client to the server part.

C. Features Extraction

Speaker utterances issued from TIMIT or G729TIMIT database, were represented by 19 LPCC or LFCC, with

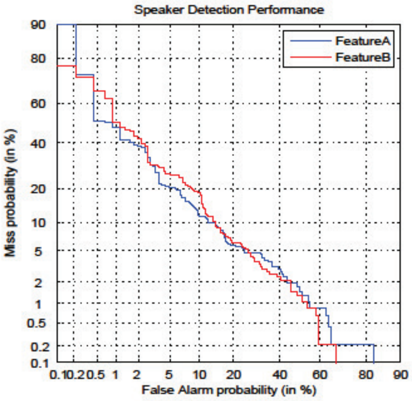

Fig. 2. The performance of GMM-UBM system based LPCC and LFCC features extracted from TIMIT database.

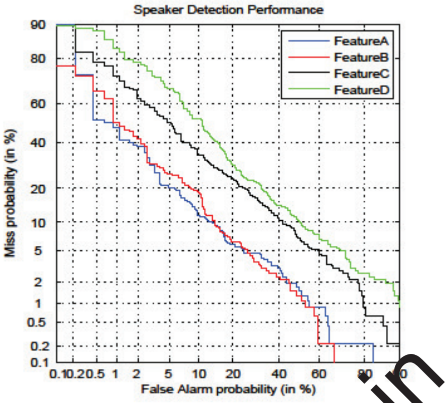

Fig. 3. The performance of GMM-UBM system based LPCC and LFCC features extracted from G729TIMIT database.

their first derivatives and the delta energy. Altogether, a 40 coefficients vector is extracted from clean (TIMIT) and G729TIMIT (G729 resynthesized speech). Mean subtraction and variance normalization were applied to all features. In this work the LPCC and LFCC features are derived under different conditions, four experiments are adopted and named as; FeatureA, FeatureB, FeatureC and FeatureD:

- Feature A: LPCC features obtained from the clean TIMIT corpus;
- Feature B: LFCC extracted from TIMIT database;
- Feature C: LPCC computed from the G729TIMIT database;
- Feature D: LFCC features derived from G729TIMIT database.

D. Experimental Results

Four different experiments are presented. In the first experiment, we obtained the best performance with GMM-UBM model based feature A, correct rate to 89%. The second experiment, with features B , afford a correct rate to 87% in average the same result in experiment based feature A. , the performance of the system degrades considerably due to the utilization of the G729 codec speech, the correct rate is found at 80% In the experiment under featuresC and to 77% in the last experiment based Feature D. From this experiments, it has been observed that both LPCC and LFCC along with its 1st order derivatives can work as efficient parameterization of the speech signal for GMM-UBM recognition system, however LPCC shows relative robustness compared to LFCC features. A plot of GMM-UBM scores system using the four kinds of features extractors is shown in figure 2 and 3.

VI. Conclusion

This work reflects the results obtained in the evaluation of a LPCC and LFCC features on clean and G729 resynthesized speech. It has been observed that under sclean speech, LPCC and LFCC feature vector gives a bast accuracy recognition. However The performance of the GMM-UBM system degrades considerably with the resynthesized speech, and LPCC features achieves the best result and shows relative robustness compared to LFCC features. From the above experiments, it has been observed that both LPCC and LFCC can work as efficient features extracted from G729 synthesized speech system.

References

[1] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, J.M. Huerta and R.M. Stern, "Speech recognition from GSM coder parameters,". in Proc. 5th Int. Conf. On Spoken Language Processing, vol. 4, 1998, pp.1463-1466.

[2] W.M. Campbell, "Speaker and language recognition using speech codec parameters,". in Peoc. Eurospeech99, vol.2, 1999, pp.787-790.

[3] T.F. Quatieri, R.B. Dunn, D.A. Reynolds, J.P. Campdell and E. Singer, "Speaker Recognition Using G.729 Codec Parameters,". in Proc. ICASSP. 2000, pp. 89-9.

[4] D. Yessad and A. Amrouche, "G729 Coded Parameters Under Matched and Mismatched Conditions for Distributed Speaker Recognition,". ICTA12 Bejaia, Algeria, 2012.

[5] ITU-T Recommendation G.729.:Coding of Speech at 8 Kbit/s Using Conjugate-Strucuture Algebraic Code-Excited Linear-Prediction (CS-ACELP), 1996.

[6] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker verification using adapted gaussian mixture models,". in Digital Signal Processing, vol. 10, no. 1-3, 2000, pp. 19-41.

[7] http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA RAL.