ROBUSTNESS OF BIT-STREAM BASED FEATURES FOR SPEAKER VERIFICATION

A. Moreno-Daniel[*], *B.H. Juang*

Georgia Institute of Technology
Center for Signal and Image Processing
Atlanta, GA 30318

J.A. Nolazco-Flores[†]

Instituto Tecnológico y de Estudios Superiores
de Monterrey, Campus Monterrey
Departamento de Ciencias Computacionales
Monterrey NL, México

ABSTRACT

This contribution presents a Speaker Verification system that uses YOHO database which has been coded with ITU-T G.729 standard. A set of bitstream based features consisting of 16 LPC Cepstral coefficients and MFCC derived from the quantized line spectral pairs as well as residual information in the form of pitch was utilized to construct the speaker's models, and their robustness was studied under white noise conditions.

Results suggest that using a cohort model, MFCC are more robust under noise conditions than LPC Cepstral coefficients; the addition of pitch to the feature vector contributes from a 16% to a 29% of improvement in verification performance under different noise conditions.

1. INTRODUCTION

In the last few years, the demand of an extra security layer joint with the availability of communication tools have made biometrics an increasing important area of research to deal with many emerging scenarios. Among these demanding needs we can mention the ability to perform critical operations or retrieve confidential information remotely in a secure way, identity verification for immigration purposes, or for voting in an election process; also, several tools now allow ubiquitous access to information such as wired/wireless Internet connection, cellular phone and satellite networks, etc. Voice as a biometric measure is a non-intrusive way to validate a person's identity given that it is the main communication channel for humans. Furthermore, speech acquisition is simple and inexpensive because it doesn't require any special apparatus; additionally, the infrastructure to convey speech from one place to another has grown exponentially.

The ultimate goal of a Speaker Verification (SV) system is to correctly accept a legitimate registered user and reject impostors, who falsely claim to be a legitimate user, therefore protecting restricted information or privileges by means of estimating and verifying a set of physiological characteristics extracted from the speech waveform.

Although the task of SV has been rather well studied for over forty years [1], there has been a number of recent advances such as the use of Gaussian mixtures to model the individual's different configurations of the vocal tract [2], different score normalization schemes as presented in [3, 4], or techniques for the selection of the threshold value as studied in [5].

More recently, as a result of the growing utilization of Voice over IP (VoIP) and cellular telephony for remote operations, bitstream based features have been studied in [6] for automatic speech recognition (ASR) and in [7, 8] for SV.

Our work focuses on SV where speech waveforms have suffered a lossy compression by the ITU-T G.729 codec which is a prevalent coding standard in wireless and voice over IP applications. A set of features derived from the encoder's bit-stream is proposed, including residual information in the form of pitch, and used in an experimental system that demonstrates their effectiveness. Section 2 presents an overview of the SV system, including background information for completeness; section 3 describes the experimental setups and the studied techniques. Finally, section 4 discusses the results.

2. SPEAKER VERIFICATION SYSTEM

2.1. Background

The task of SV can be classified as *text dependent* or *text independent*, in the case when the SV system knows the transcription of the verification utterance or not, respectively. In a pseudo *text independent* SV system, like the one presented in our work, the exact transcription of the verification utterance is unknown, however it is known to belong to a closed set with fixed characteristics as described in section 2.2.

In general each registered individual 'k' has a corresponding statistical model λ_k, a Gaussian mixture in our case, which is trained with a sequence of feature vectors $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$ extracted from speech with 25-30 ms overlapping windows, every 10 ms. The labels of these features indicate the speaker's identity.

Given sequence of observed independent and identically distributed (iid) features $\mathbf{O}$, the likelihood of being generated by the k-th speaker is:

$$\mathcal{L}(\lambda_k|\mathbf{O}) = p(\mathbf{O}|\lambda_k) = \prod_{t=1}^{T} p(\mathbf{o}_t|\lambda_k), \tag{1}$$

$$p(\mathbf{o}_t|\lambda_k) = \sum_{m=1}^{M} \omega_{m,k}\,\mathcal{N}(\mathbf{o}_t; \mu_{m,k}, \mathbf{\Sigma}_{m,k}), \tag{2}$$

$$\sum_{m=1}^{M} \omega_{m,k} = 1; \qquad \forall k, \tag{3}$$

[*] Partial work was done while the author visited ITESM

[†] This work was supported by the ITESM Information Security Chair and CONACyT 2002-C01-41372

0-7803-8874-7/05/$20.00 ©2005 IEEE

where $\mathcal{N}(\mathbf{o}_t; \mu_{m,k}, \mathbf{\Sigma}_{m,k})$ is a multivariate Gaussian probability distribution function (pdf) with mean vector $\mu_{m,k}$ and covariance matrix $\mathbf{\Sigma}_{m,k}$; therefore the models, characterized as $\lambda_k \equiv (\omega_{m,k}, \mu_{m,k}, \mathbf{\Sigma}_{m,k}; m = 1, \ldots, M)$, are estimated using the training data (properly labeled) and the maximum likelihood criterion, via the EM algorithm.

After the speaker's models have been trained, the decision of whether to accept or reject is based on the two hypotheses H_0: true speaker; H_1: impostor. The likelihood ratio test (LRT) evaluates to a score value that suggests to accept H_0 over H_1.

$$\hat{\Theta}(\mathbf{O}) = \frac{P(H_0)}{P(H_1)} = \frac{p(\mathbf{O}|\lambda_k)}{p(\mathbf{O}|\lambda_{\bar{k}})} \tag{4}$$

$$\Theta(\mathbf{O}) = \log \hat{\Theta}(\mathbf{O}) = \log p(\mathbf{O}|\lambda_k) - \log p(\mathbf{O}|\lambda_{\bar{k}}). \tag{5}$$

Notice that while λ_k represents the model for the k-th speaker, $\lambda_{\bar{k}}$ represents the model of an impostor that provides a contrast. Depending on the chosen scope of contrast, this last model can be defined under three schemes: as a *universal background model* (*UBM*), as a *cohort* or as a *background* model. The *UBM* is a single speaker independent (SI) model which is trained with utterances from a large group of impostors; on the other hand, the *cohort* score is obtained from multiple speaker models by averaging the likelihood of the observation with a speaker dependent (SD) group of impostors $\mathrm{J_k}$ as stated in Eq. 6; finally the *background* model is a single SD model trained with data from a limited selected group of impostors.

$$P(\mathbf{O}|\lambda_{\bar{k}}) = \frac{1}{N} \sum_{j \in \mathrm{J_k}} P(\mathbf{O}|\lambda_j). \tag{6}$$

Once a verification score has been obtained, a threshold value is needed to decide if the observation was produced by an impostor or by a registered speaker. This value is selected by minimizing the cost function:

$$\mathcal{C} = P(\mathcal{I})P(\hat{\mathcal{T}}|\mathcal{I})\mathcal{C}_{FA} + P(\mathcal{T})P(\hat{\mathcal{I}}|\mathcal{T})\mathcal{C}_{FR}, \tag{7}$$

where $\mathcal{T}$ and $\mathcal{I}$ refer to a true claimant and an impostor respectively; and $\hat{\mathcal{T}}$ and $\hat{\mathcal{I}}$ refer to deciding the speaker was a true claimant or impostor respectively. $\mathcal{C}_{FA}$ and $\mathcal{C}_{FR}$ are the cost weights for false acceptance (FA) and false rejection (FR) that suit the application. FA and FR correspond to type I and type II errors in our hypothesis testing formulation.

2.2. The database

YOHO [9] is a database with 138 speakers (32 female and 106 male), including at least four speakers with mother tongue other than American English. It has two main sections: ENRollment and VERification; furthermore ENR has 4 sessions with 24 utterances each, and VER has 10 sessions with 4 utterances each; resulting in a total of 13248 and 5520 enrollment and verification utterances respectively. The transcription of utterances consists of a 'lock-combination' phrase (three two-digit numbers). The speech waveforms have been quantized with 12-bits and sampled at 8 KHz.

Although the length of each waveform is around 3-5 s, only about 2.5 s is active speech, yielding to roughly 240 s of active speech for ENR per speaker.

2.3. ITU-T G.729 codec

ITU-T G.729 [10] is a set of speech coding standards recommended for digital cellular phones, operating at the rate of 8 kb/s. This recommendation describes a "toll quality" Conjugate Structure Algebraic Code Excited Linear Prediction encoder (CS-ACELP), with a frame rate of 10 ms at 80 bits/frame. The input speech must be sampled at 8 kHz represented in 16-bit linear PCM (Pulse Code Modulation) format. A 10th order linear prediction analysis is performed on every frame of windowed speech generating parameters that characterize the signal production system. These parameters, sometimes referred to as short-term prediction or spectral envelope information, are transformed into Line Spectral Pairs (LSP) [10] parameters for quantization. The residual or excitation information consists of two components: periodic and random.

Every frame, 18 bits are allocated for the short-term predictor in the form of LSP parameters, while 62 bits are used for the residual (20 and 42 bits for the periodic and random components respectively). The average spectral distortion due to quantization is approximately 1.5-2dB [11].

The periodic part of the residual consists of pitch estimates $\mathrm{P_s}$, which provides an index pointer for a position in the adaptive codebook (CB) to facilitate "long term" prediction spanning over a pitch period; and pitch gains $\mathrm{G_p}$'s, which is the corresponding scaling factor to produce the best match between the input speech and its delayed version as encapsulated in the adaptive CB. Notice that the gain is also a measure of correlation between the input and its delayed version; the magnitude of such a long span correlation is nearly one for a periodic signal and nearly zero if the signal lacks of periodicity. It can thus be considered a crude measure of vocality.

The random part consists of the algebraic CB indices and signs (I_c's and S's) and the fixed (algebraic) CB gains G_a's. This component is related to the excitation function that cannot be properly represented with both the long and short term predictors.

3. EXPERIMENTAL SETUP

Two sets (A and B) of experiments are presented in this section. Set A obtains a sequence of feature vectors from the speech waveform, as it is conventionally done, by means of a mel-scaled filter bank, transforming to cepstral domain and obtaining 13 MFCC (mel-frequency cepstral coefficients). Input speech waveforms consist of clean and noisy speech, where the noise can be either Gaussian white noise (for SNR= 20dB, 15dB, 10dB), coding distortion, or both.

Set of experiments B computes a sequence of bit-stream based feature vectors (without waveform synthesis). We study two types of LSP-derived features in the form of LPC-Cepstra and MFCC (from the LPC estimated spectrum). Additionally, we attempt to incorporate residual information to the feature vectors by using the pitch estimate as explained in 3.2.

In both sets, Gaussian mixture models were used with $M = 64$ mixtures. 138 speaker's models were trained. Previous work [7] has studied SV from bit-stream using a *UBM* for H_1; in our work we use *cohort* to calculated the verification score using Eq. 5 and 6; where $\mathrm{J_k}$ is the set of all speakers excluding k.

3.1. Set A (waveform)

This experiment set is our baseline. A total of 12 MFCCs plus energy is extracted every 10 ms; then Δ and Δ^2 are appended to the vector, forming a 39 dimension feature vector.

Training and testing data for experiment A.1 was extracted from clean and noisy waveforms (white noise) for SNR values of 20dB, 15dB and 10dB. Both matched and mismatched conditions are studied, as it is in the rest of the experiments. In experiment A.2, we trans-coded the waveforms used in A.1 with G729 codec (coding followed by decoding). Notice that white noise was added to the clean input before trans-coding.

3.2. Set B (bit-stream)

Three types of features were extracted from G.729 bit-stream. The first one considers the quantized LSP parameters, which have a one to one correspondence to linear prediction coefficients (LPC), and further transforms them to cepstral domain using the recursion:

$$c[n] = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c[k] a_{n-k}, \tag{8}$$

where the convention of $1 - A(z)$ was used for the inverse filter, $a_0 = 1$ and $a_n = 0$ for $n > p$. Although knowing $c[1], \ldots, c[p]$ is sufficient to recover back the LPC coefficients [8]; the effect of truncating a LPC-Cepstral sequence (also referred as rec-cepstrum) or multiplying the sequence by a rectangular window, is the convolution of the log power spectrum with the Fourier transform of a rectangular window (i.e., a *sinc* function), causing the smooth of the power spectrum estimate from the LPC coefficients, and reducing therefore the sharpness of the formant peaks. This effect is not necessarily bad, since formant sharpness are artifacts themselves. In our experiments 16 LPC-Cepstral coefficients were considered.

The second type of feature we study is MFCC based on the LPC spectrum estimate, which is obtained by sampling around the unit circle as in [7]; then a mel-scaled triangular shaped filter bank is used and finally the output is converted to cepstral domain.

The third type of features consists of a pitch estimate: $\log f_0$ derived from the long term predictor. A moving median filter was applied to remove any glitches.

Experiment B.1 trains Gaussian mixture models as in experiments A, using 16 LPC-Cepstral coefficients plus Δ and Δ^2, forming a 48 dimensional feature vector. Training and testing conditions are consistent with those for experiment A.2.

Experiment B.2 is similar to experiment B.1, except that MFCC (from LPC spectrum) were used. The feature vector is 39 dimensional as in experiments A.

Experiment B.3 consists on appending the residual feature: $\log f_0$ to experiment B.2.

The objective of experiment B.3 is to observe the impact of the residual features on the final SV performance.

4. RESULTS

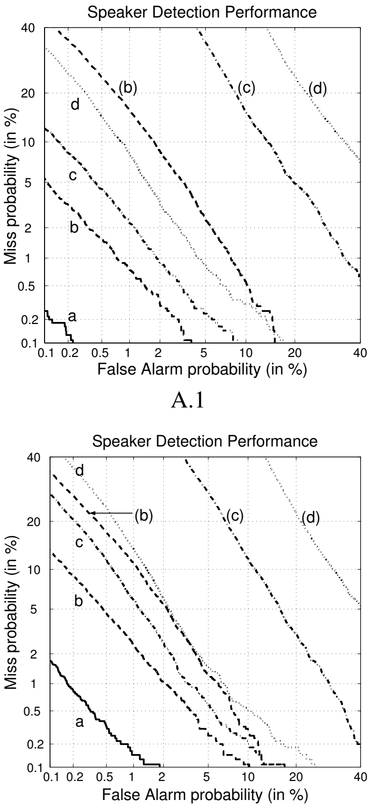
As described in section 3, we used the conventional feature vectors based on MFCC obtained from speech waveforms in experiment set A, while in experiment set B we used features based on the bit-stream. A cohort approach was used for H_1.

Results are presented in the form of equal error rate (EER) as in table 1 or in the form of detection error trade-off (DET) plots. The EER operation point of a SV system occurs when the threshold is adjusted so that FR and FA have the same value. A DET plot computes the two types of erros while sweeping the threshold, using normal deviate scale in both axis.

Exp /dB	Cln	20	15	10	(20)	(15)	(10)
A.1	0.18	1.0	1.4	2.4	3.8	12.0	22.0
A.2	0.41	1.5	2.2	3.1	3.1	11.0	21.0
B.1	0.34	-	1.6	2.8	15.0	28.0	-
B.2	0.45	1.2	1.9	3.3	3.8	10.0	21.0
B.3	0.34	1.0	1.5	2.5	2.7	7.2	17.0

Table 1. EER in % for experiment sets A and B. Column header indicates SNR in dB, the lack and presence of parenthesis denote matched and mismatched conditions respectively.

Figure 1 shows that although coding noise deteriorates the SV performance under clean conditions by 0.23%, the effect of noise presence masks the effect of coding distortion, causing both performances to be comparable under noisy matched conditions. On the other hand, speaker models trained with clean trans-coded speech showed more robustness under mismatched noisy conditions.



A.1

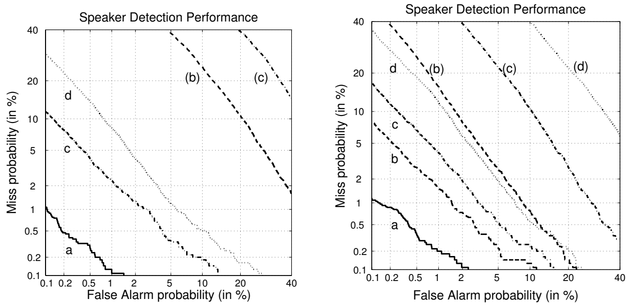
A.2

Fig. 1. DET plots from experiment set A. Labels {a, b, c, d} denote {clean, 20dB, 15dB, 10dB} respectively. Parenthesis indicate mismatch condition.

Experiment sets B.1 and B.2 use bit-stream based feature vectors: LPC-Cepstra coefficients and MFCC, both derived from the quantized LSP. Figure 2 depicts the performance of both experiments, where it can be observed that experiment B.1 outperforms

B.2 under clean conditions by 0.11%, however MFCC showed more robustness under mismatched condition.

Contrasting these results with the ones obtained from conventional MFCC features from waveform (experiment A.2), both bit-stream based experiments, B.1 and B.2, perform similarly to A.2 under matched condition, while only B.2 yields to comparable results under mismatched condition.



B.1 B.2

Fig. 2. DET plots from experiment sets B.1 and B.2: LPC-Cepstra and MFCC from quantized LSP.

Clearly, residual information is uncorrelated to spectral information, becoming a good candidate to extend our LSP-based feature vector. Experiment B.3 appends the third bit-stream feature: $logf_o$ to the MFCC vector obtained from LSP, because of its robustness under mismatched conditions found in experiment B.2. Figure 3 shows the SV performance for this augmented feature vector. By comparing Fig. 3 with Fig. 2, we can appreciate the amount of speaker's information provided by the pitch.

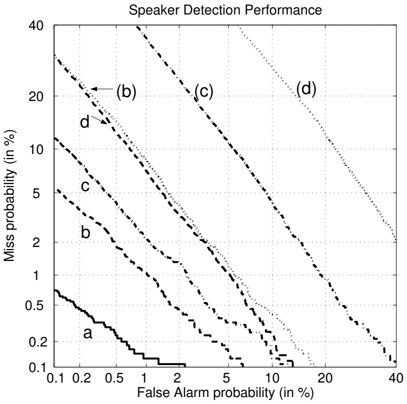


Fig. 3. Experiment B.3, augmented feature vector: MFCC from LSP + $\log f_0$

From the EER presented in table 1, we observe that the use of residual information (pitch) brings an improvement to the SV performance in the range of 16% to 29%, depending on the scenario.

Comparing the results of the bit-stream based SV system presented in experiment B.3 with the results in experiment A.2 (MFCC from trans-coded waveform); bit-stream features outperform trans-coded waveform approach in all scenarios. Additionally, the performance of B.3 under noisy matched conditions is comparable to experiment A.1 (waveform without coding distortion); while B.3 outperforms A.1 under mismatch conditions.

5. CONCLUSIONS

The robustness of bit-stream features was studied in this paper for features extracted from G.729 bit-stream, using a cohort approach for the H_1. These features were derived from the quantized LSP coefficients and the residual in the form of pitch. It was found that a MFCC-like feature vector obtained from LSP coefficients was more robust than LPC-Cepstral feature vectors. Augmenting the MFCC feature vector to include pitch information improved the performance in some cases up to 29%.

The augmented MFCC + $\log f_0$ feature vector outperforms the conventional approach of extracting MFCC from trans-coded waveforms under all scenarios, by using a cohort score for H_1.

6. REFERENCES

[1] D.A. Reynolds, "An overview of automatic speaker recognition technology," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, pp. 4072–4075, 2002.

[2] D.A. Reynolds, *A Gaussian mixture modeling approach to text-independent speaker indentification*, Ph.D. dissertation, Georgia Institute of Technology, 1992.

[3] O. Siohan, C.H. Lee, A.C. Surendran, and Q. Li, "Background model design for flexible and portable speaker verification systems," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 825–828, 1999.

[4] A.E. Rosenberg, J. Delong, C.H. Lee, B.H. Juang, and F.K. Soong, "The use of cohort normalized scores for speaker recognition," *Int. Conf. Spoken Lang. Proc.*, pp. 599–602, 1992.

[5] K. Chen, "Towards better making a decision in speaker verification," *Pattern Recognition*, vol. 36 (2), pp. 329–349, 2003.

[6] H.K. Kim and R. Cox, "A bitstream-based front-end for wireless speech recognition on is-136 communications system," *IEEE Trans. Speech and Audio Proc.*, vol. 9, pp. 558–568, 2001.

[7] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, and J.P. Campbell, "Speaker and language recognition using speech codec parameters," *Proc. Europ. Conf. Speech Comm. Technology*, vol. 2, pp. 787–790, 1999.

[8] T.F. Quatieri, R.B. Dunn, D.A. Reynolds, J.P. Campbell, and E. Singer, "Speaker recognition using g.729 codec parameters," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 89–92, 2000.

[9] J.P. Campbell Jr, "Testing with the yoho cd-rom voice verification corpus," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 341–344, 1995.

[10] ITU-T, "Recommendation G.729 - coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear-prediction (CS-ACELP)," 1996.

[11] K. Paliwal and B. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Proc.*, vol. 1, pp. 3, 1993.