Speaker Verification Using Coded Speech

Antonio Moreno-Daniel[1,2], Biing-Hwang Juang[1], and Juan A. Nolazco-Flores[2,*]

[1] Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta GA, USA
{antonio,juang}@ece.gatech.edu
[2] Departamento de Ciencias Computacionales, Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey NL, México
jnolazco@itesm.mx

Abstract. The implementation of a pseudo text-independent Speaker Verification system is described. This system was designed to use only information extracted directly from the coded parameters embedded in the ITU-T G.729 bitstream. Experiments were performed over the YOHO database [1]. The feature vector as a short-time representation of speech consists of 16 LPC-Cepstral coefficients, as well as residual information appended in the form of a pitch estimate and a measure of vocality of the speech. The robustness in verification accuracy is also studied. The results show that while speech coders, G.729 in particular, introduce coding distortions that lead to verification performance degradation, proper augmented use of unconventional information nevertheless leads to a competitive performance on par with that of a well-studied traditional system which does not involve signal coding and transmission. The result suggests that speaker verification over a cell phone connection remains feasible even though the signal has been encoded to 8 Kb/s.

1 Introduction

The objective of a Speaker Verification (SV) system is to correctly accept legitimate registered users and reject impostors, who falsely claim to be legitimate users, therefore protecting restricted information or privileges. The task of recognizing or verifying a person's identity has gained relevance and interest as the technology allows us to perform critical operations or receive services remotely, such as on-line or telephone banking, shopping, trading, etc. [2]. Among all biometrics, "voice" has the advantage [3] that it doesn't require any sophisticated apparatus; individuals can provide speech samples in a very natural way and most people are accustomed to speaking to a handset. Furthermore, the availability of cell phone services and Internet access (wired or wireless) makes this kind of operations simple and low-cost.

Speaker verification is a subject that has been rather well studied [2-3]. Many new advances have also been reported. For example, Ref. [4] proposed the use of general Gaussian mixture models which offer improved speech modeling resulting in better verification accuracy; Ref. [5] reported significant performance improvement using

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

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 366–373, 2004.

minimum verification error training; and Li et al. in [6] proposed the method of utterance verification embedded in a human-machine dialog which can be used for both automatic registration and speaker verification. In this work, we focus on the issue of speech coding and its impact on the performance of a speaker verification system due to the fact that nearly all telecommunication networks today are digital and thus speech signals that are being transmitted through the networks are all encoded into bit-streams at various bit rates. Since speech coding is in general of a lossy type, it thus will inevitably introduce distortion to the decoded signal. An assessment of the impact of signal distortion due to coding upon automatic speech recognition was provided by [7] for the purpose of evaluating the potential detriment that speech coding may bring upon voice-enabled services. Here we turn our attention to the application of speaker verification with a similar motivation. However, unlike the earlier report on speech recognition based on coded speech [8], our work here includes a novel use of additional information already existent in the output of the speech coder; our system thus can be considered a new design.

SV can be classified into a 'text-dependent' mode in which the SV system knows the transcription of the utterance pronounced by the claimant; or a 'text-independent' mode in which the transcription is unknown and the utterance may be arbitrary. In this work, a ‘pseudo text-independent’ SV system was built, where the system doesn't know the transcription of the input utterance, but it does know it is within a closed set (see Section 2.3).

This paper is organized as follows. First, background information is presented in Section 2, including details of interest about the encoder, the speaker verification database, and how they were used to build the SV system. Section 2.3 presents the basic configuration of the experiments, and Section 3 describes the proposed use of additional information which is derived from the encoded bit-stream; for brevity, we call the system a bit-stream level system. Finally, Section 4 presents a comparison of results obtained with our scheme against those with conventional SV systems.

2 Background

2.1 Database

Our evaluation uses the YOHO database [1], which consists of a series of lock-combination sentences pronounced in American English by 138 subjects (106 male and 32 female), having a wide range of ages, jobs and education, including at least 4 speakers with foreign mother tongue.

This database is originally divided into two main sets: the ENRollment set and the VERification set; furthermore, ENR has 4 sessions with 24 utterances each, and VER has 10 sessions with 4 utterances each; resulting in a total of 13,248 enrollment utterances and 5,520 verification utterances. Although the length of each wave file is around 3 to 4 seconds, only about 2.5 seconds is active speech, which yields to roughly 240 seconds of active speech for ENR per speaker.

2.2 ITU-T G.729

ITU-T G.729 [9] is a set of speech coding standards recommended for digital cellular phones, operating at the rate of 8 kb/s. The recommendation ITU-T G.729 describes a "toll quality" 8 kb/s Conjugate-Structure Algebraic-Code-Excited Linear-Prediction encoder (CS-ACELP), with a frame rate of 10ms at 80 bits/frame. The input speech waveform is sampled at 8 kHz with each sample represented in a 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) parameters for quantization. The residual or excitation information consists of two components: periodic and random. Table 1 illustrates how the 80 bits are allocated to the complete set of encoder parameters.

Table 1. Bit allocation for various parameters in G.729

Parameter		Codeword	Subframe 1	Subframe 2	Total per frame
Line Spectrum pairs		**L0, L1, L2, L3**	–	–	18
Periodic component	Pitch Delay index for Adaptive Codebook	**P1, P2**	8	5	13
	Pitch-Delay Parity	**P0**	1	–	1
	Gains (pitch) for Adaptive Codebook	**Gp1, Gp2**	3	3	6
Random component	Fixed Codebook Index	**Ic1, Ic2**	13	13	26
	Fixed Codebook Sign	**S1**	4	4	8
	Algebraic Codebook Gains	**Ga1, Ga2**	4	4	8
Total					80

As shown in the table, a total of 18 bits per frame are spent for the short-term predictor in the form of line spectral pair parameters, while 62 bits are used for the residual (20 for the periodic part and 42 for the random component).

The periodic part of the residual consists of pitch estimates **P**s, which provides an index pointer for a position in the adaptive codebook to facilitate "long-term" prediction spanning over a pitch period; and pitch gains **Gp**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 codebook. Note 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 periodicity. It can thus be considered a crude measure of vocality.

The random part consists of the algebraic codebook indices and signs (**Ic**s and **S**s) and the fixed (algebraic) codebook gain (**Ga**s). This component is related to the excitation function that cannot be properly represented with both the long and short-term predictors.

Figure 1 depicts how the decoded bit-stream is used in various components of the decoder/receiver for the reconstruction of speech waveform. Here we assume that the inner layer of information is available to the speaker verification system. We thus refer to such an SV system "the bit-stream level system" without ambiguity.

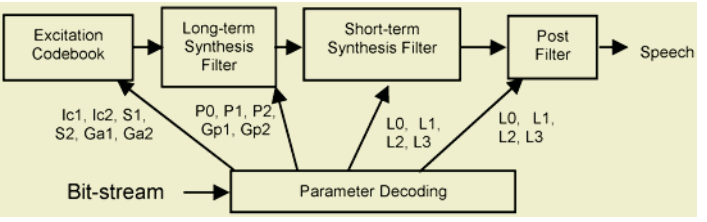

Fig. 1. Decoded parameters at the decoder for speech synthesis

Typical applications of G.729 speech coder include Voice over IP, satellite communications, and digital cellular phone service.

2.3 Experimental Setup

The application scenario considered consists of an individual requesting access to restricted information or privilege from a remote location (via a cellular phone or a Voice over IP connection) to a server, which has full access to the bit-stream transmitted (for example, an SV service provided by a cellular phone carrier, or an SV server connected to the Internet).

The SV system follows the Gaussian mixture universal background model (GMM-UBM) paradigm, also known as general or world background model [3]. There are 138 different models (one for each speaker) plus one speaker independent universal background model (UBM), a silence and a short pause. Although the transcriptions are unknown, it is known that the utterances have three words (lock combinations).

In order to build a UBM background model, the database was repartitioned into two sets: I and II, keeping half of the individuals on each set (53 male and 16 female per set). Each set has an ENR and VER subsets. Two separate runs of SV experiments were performed, where in one case, the entire set II was used to estimate (train) the background model, while in the other, only set I was used for registered users and verification attempts. By doing this, it is ensured that none of the individuals used to estimate the background model would be present during verification.

We use the tool kits of HTK [10] to train and test the models. Each individual's model and the UBM consist of Gaussian mixtures (single-state HMMs) with 40 mixture components, which attempt to model the different vocal tract configurations [4] as represented in the feature vectors.

We test the hypothesis {H_0: the claimant is indeed the registered user} against {H_1: the claimant is an impostor}, using the log likelihood ratio (LLR) computed as follows:

$$\textbf{Accept } H_0 \textbf{ if } \theta(\mathbf{O}) = \log \frac{P(\mathbf{O} \mid \lambda_c)}{P(\mathbf{O} \mid \lambda_{UBM})} \geq \tau \tag{1}$$

where λc is the claimed identity's model, λ_{UBM} is the universal background model, and $\mathbf{O}$ is the sequence of observed feature vectors. The decision whether to accept or reject the claimant depends on the threshold τ. For analysis purposes, results are presented using DET (Detection Error Tradeoff) plots, therefore leaving the choice of τ

open to suit a desired application. False-Alarm corresponds to False Acceptance (FA) and Miss corresponds to False Rejection (FR).

3 Bit-Stream Level Speaker Verification

As mentioned in Section 2.2, a quantized version of the spectral envelope information is available in the bit-stream from the 10 LSP parameters (Line Spectrum Pair frequencies). These parameters have a one-to-one correspondence to the LPC coefficients (Linear Prediction Coefficients), which can be further transformed to cepstral domain using the following recursion [11]:

$$c[n] = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c[k]\, a_{n-k} \tag{2}$$

where we have used the convention of 1-*A(z)* for the inverse filter and a_0=1 and a_n=0 for $n>p$. It is clear that when p is known, $c[1]$ … $c[p]$ are sufficient to recover back the LPC coefficients. The effect of truncating the LPC-Cepstral coefficients (also called rec-cepstrum [12]), or multiplying the coefficients by a rectangular window, is a convolution of the log power spectrum with the Fourier transform of a rectangle window (i.e., a *sinc* function), causing smoothing of the power spectrum estimate from LPC coefficients, and reducing the sharpness of the formant peaks [13]. This smoothing effect is up to a point desirable, since sharp formant peaks are often artifact themselves.

Although spectral envelope conveys information that characterizes a person's identity, the residual carries another component (it is well known that it is still possible to guess the identity of the talker by simply listening to the LPC residual signal). Several techniques have been proposed to extract these characteristics from the residual, including LPC analysis over the residual [15] and appending the residual parameters [12] to the feature vector.

Our proposed feature vector appends residual information to the LPC-Cepstral coefficients, in the form of a fundamental frequency measure ($\log f_0$) and a measure of vocality (mv_{k}) estimated by combining the pitch gain (**Gp**) and codebook gain (**Ga**) as follows:

$$\begin{aligned} \Gamma p_k = \text{median}\{&Gp2_{k-2}, Gp1_{k-1}, Gp2_{k-1}, \\ &Gp1_k, Gp2_k, Gp1_{k+1}, Gp2_{k+1}, Gp1_{k+2}\} \end{aligned} \tag{3}$$

and similarly for Γa_{k}, to find:

$$mv_k = \ln\left(\frac{\Gamma p_k}{\Gamma a_k}\right) \tag{4}$$

where index k denotes the frame number, and Γs are results of a 40ms moving median, therefore removing spurious glitches.

The structure of our 54-dimensional feature vector is:

$$FV = \left[c_1, c_2, \ldots, c_{16}, \log(f_0), mv, [\Delta], [\Delta^2]\right]. \tag{5}$$

A baseline experimental setup 'A' consists of a conventional SV system applied to the original set of waveforms in YOHO. The SV system uses 12 MFCC (Mel-frequency cepstral coefficients) plus an energy term.

Additionally, Δ and Δ^2 are appended, resulting in a 39-dimensional feature vector.

Similarly, in order to illustrate the impact on SV performance using G.729 coded waveform, experimental setup 'B' applies the same scheme as in setup 'A' to a transcoded version of the database.

The performance of the proposed feature vector eq.(5) is tested in experimental setup 'C'.

For every experimental setup described above, robustness was tested in noisy conditions with SNR values of 20dB and 15dB. Noise level was adjusted using ITU-T P.56 recommendation software [14].

4 Experimental Results

As shown in Figure 2, a conventional SV system is capable of achieving an equal error rate (i.e., when %FA=%FR) of slightly more than 1.6% without obvious noise interference. When additive noise is present at 20dB SNR, the performance deteriorates to about 3.5% and to about 5% when SNR is 15dB.

When the speech signal undergoes G.729 coding and decoding, the SV system that takes the reconstructed waveform as input can only achieve roughly 2%, 4% and 5% equal error rate under clean, 20dB SNR and 15 dB SNR conditions, respectively, as shown in the second plot (Set B) of Figure 3. Note that the coder G.729 is generally considered toll quality at 8 kb/s. The result shows that while the distortion introduced by the coder may not be obvious to our perceptual apparatus, it is causing deterioration in the performance of a speaker verification system. The result is consistent with that of [7] and the recommendation is to avoid transcoding if possible.

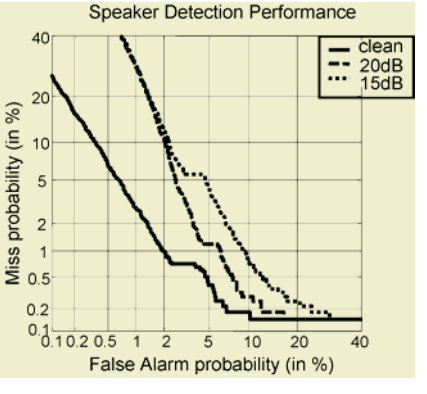

Fig. 2. Speaker Verification results from experimental setup A. Clean and noisy data with SNR values of 20 dB and 15 dB, using conventional MFCC features extracted from the waveform

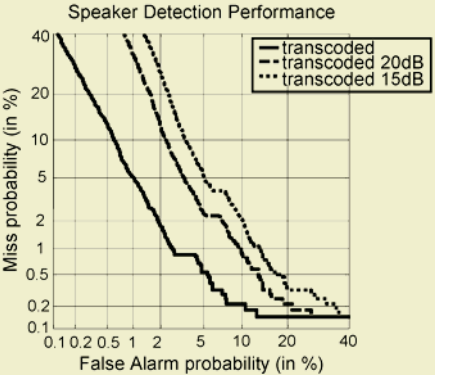

Fig. 3. Speaker Verification results from experimental setup B. Transcoded waveforms and noise added before the codification for SNR values of 20 dB and 15 dB

When additional feature parameters are used as described in eq.(5), the performance is slightly better than experimental setup B, without having to synthesize the

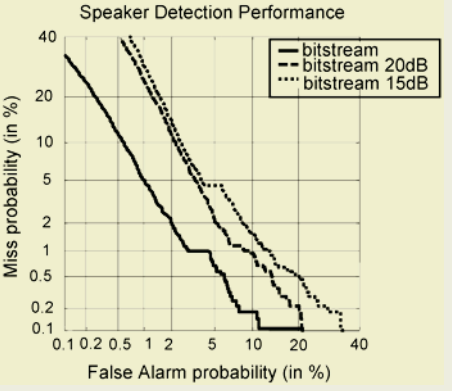

Fig. 4. Speaker Verification results from experimental setup C, using the proposed feature vector, extracted directly from the bit-stream (without synthesis)

quantized parameters into a waveform for preprocessing. Under a "clean" condition, use of the augmented feature vector is able to reduce the equal error rate to nearly 2%, as shown in Figure 4 (Set C). Even when noise is present, at 20dB and 15dB SNR, respectively, the equal error rates achieved by the new proposed system are approximately 4% and 4.8%. The augmented features show robustness comparable to MFCCs.

5 Conclusions and Future Work

Given the availability of mobile and telecommunication infrastructure that allows remote, ubiquitous access to critical service transactions, SV is an area that has gained attention recently. We have presented a bit-stream level SV system, that incorporates the residual in terms of an estimate of the log pitch frequency, and a measure of vocality, derived from G.729 pitch gains and codebook gains.

The experimental results show that this somewhat crude incorporation of the residual-derived (or excitation-derived) feature matches the performance of MFCCs extracted from transcoded speech; setting a baseline and leaving the opportunity to improve the performance by using the measure of vocality to distinguish the segments of speech that characterize the best the anatomic characteristics of the speakers.

References

1. Campbell, J.P., Jr.: Testing with the YOHO cd-rom voice verification corpus. Proc. ICASSP, (1995)
2. Furui A.: Recent Advances in Speaker Recognition. First Int. Conf. Audio- and Video-based Biometric Person Authentication. Switzerland, (1997) 237-252
3. Reynolds, D.A.: An Overview of Automatic Speaker Recognition Technology. Proc. ICASSP, (2002)
4. Reynolds, D.A., Rose R.: Robust Text-Independent Speaker Identification Using Gaussians Mixture Speaker Model. IEEE Transactions on Speech and Audio Processing, (1995)

5. Rosenberg, Aaron E., Siohan O., S. Parthasarathy: Speaker verification using minimum verification error training. Proc. ICASSP, (1998)
6. Li, Q., Juang, B.-H., Zhou, Q. and Lee, C.-H.: Automatic Verbal Information Verification for User Authentication. IEEE Transactions on Speech and Audio Processing, (2000) 585-596
7. Kim, H.K., and Cox, R.: Bitstream-based feature extraction for wireless speech recognition. Proc. ICASSP, (2000)
8. Zhong, X.: Speech coding and transmission for improved recognition in a communication network. PhD Dissertation, Georgia Institute of Technology, (2000)
9. ITU-T Recommendation G.729, Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP), (1996)
10. Young S., et al.: The HTK Book, Cambridge University, Version 3.2 ed., (2002)
11. Huang X., Acero A., Hon H.W.: Spoken language processing, Prentice Hall, (2001)
12. Quatieri, T. F., et al: Speaker Recognition Using G.729 Speech Codec Parameters, Proc. ICASSP, (2000)
13. Rabiner L. and Juang B.H.: Fundamentals of Speech Recognition, Prentice Hall, (1993)
14. ITU-T Recommendation G.191, Software tool library 2000 user's manual, (2000)
15. Yu Eric W.M., Mak Man-Wai, Sit Chin-Hung and Kung Sun-Yuan: Speaker verification based on G.729 and G.723.1 coder parameters and handset mismatch compensation. Proc. of the 8th European Conference on Speech Communication and Technology, (2003)
16. Besacier L., Grassi S., Dufaux A., Ansorge M. and Pellandini F.: GSM Speech coding and Speaker Recognition. Proc. ICASSP, (2000)