Speaker Verification Based on G.729 and G.723.1 Coder Parameters and Handset Mismatch Compensation

Eric W. M. Yu, Man-Wai Mak, Chin-Hung Sit

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA

Abstract

A novel technique for speaker verification over a communication network is proposed. The technique employs cepstral coefficients (LPCCs) derived from G.729 and G.723.1 coder parameters as feature vectors. Based on the LP coefficients derived from the coder parameters, LP residuals are reconstructed, and the verification performance is improved by taking account of the additional speaker-dependent information contained in the reconstructed residuals. This is achieved by adding the LPCCs of the LP residuals to the LPCCs derived from the coder parameters. To reduce the acoustic mismatch between different handsets, a technique combining a handset selector with stochastic feature transformation is employed. Experimental results based on 150 speakers show that the proposed technique outperforms the approaches that only utilize the coder-derived LPCCs.

1. Introduction

As a result of the popularity of digital communication systems, there has been increasing interest in the automatic recognition of resynthesized coded speech [1], [2], [3]. For instance, 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 [2] and [3], 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.

In addition to the acoustic distortion caused by the transcoding process, transducer variability can also result in acoustic mismatches between the speech data gathered from different handsets. The sensitivity to handset variations means that handset compensation techniques are essential for practical speaker verification systems. Feature transformation is a possible approach to minimizing the mismatch caused by the transcoding process and handset variability. We have recently proposed a technique that combines maximum-likelihood based feature transformation and handset identification for speaker verification [4, 5]. In [6], we further demonstrated that the technique is also applicable to the resynthesized speech of six coders. This paper extends this technique to speaker verification over a digital communication network in which speaker-dependent features are extracted directly from coders' parameters.

In this work, the ITU-T G.729 [7] and G.723.1 [8] speech coders were employed to encode and decode the HTIMIT cor-

pus [9]. LP-derived cepstral coefficients (LPCCs) extracted from the coder parameters were used as the basic features for speaker verification. In order to improve the verification performance, additional feature vectors that take speaker characteristics in the LP residuals into account were added to the LPCCs. Results using cepstral mean normalization as channel compensation are also shown for comparison.

2. Coder-Derived Feature Vectors

G.729 is an 8 kbit/s toll-quality speech coding standard for personal communication and satellite systems. In G.729, forward adaptation is used to determine the synthesis filter parameters every 10 ms. These filter coefficients are then converted to line spectral frequencies (LSFs) and quantized using predictive two-stage vector quantization. Each of the 10 ms frames is split into two 5 ms subframes and the excitation for the synthesis filter is determined for each subframe. The long-term correlations in the speech signal are modelled using an adaptive codebook with fractional delay. An algebraic codebook with an efficient search procedure is used as the fixed codebook. The adaptive and fixed-codebook gains are vector quantized using a two-stage conjugate structure codebook. The entries from the fixed, adaptive, and gain codebooks are chosen every subframe using an analysis-by-synthesis search.

G.723.1 is a 5.3 and 6.3 kbit/s speech coding standard for multimedia services such as video conferencing. G.723.1 specifies multi-rate coders operated on 30 ms speech frames. After high pass filtering, each speech frame is divided into 4 subframes to obtain the LP coefficients. Quantized LSFs are then determined in the last subframe using predictive split vector quantization. The unquantized LP coefficients are used to construct a short-term perceptual weighting filter. Pitch is estimated every two subframes. Together with the computed impulse response, a pitch predictor, which contributes as a conventional adaptive codebook, is created. In high bit-rate operation, multipulse maximum-likelihood quantization excitation is used, and for low bit-rate operation, algebraic code excitation is used. The pitch predictor gains and fixed codebook gains are vector quantized, and their entries are chosen using an analysis-by-synthesis search.

In G.729 and G.723.1 decoders, the received bit-stream is decoded to obtain synthesis filter coefficients. Entries from the fixed, adaptive, and gain codebooks are also determined to form an excitation signal for the synthesis filter.

In this work, a G.729 coder and a high-rate G.723.1 coder were used to code telephone speech in the HTIMIT corpus [9]. Feature vectors were derived from the coder parameters (LSFs) at the decoder side. The feature vectors are the

12-th order LP-derived cepstral coefficients (LPCCs), $\mathbf{c}_v = [c_v(1), \ldots, c_v(12)]^T$, which can be computed recursively from the quantized LP coefficients using [10]

$$c_v(n) = a(n) + \sum_{m=1}^{n-1} \left(\frac{m}{n}\right) c_v(m) a(n-m) \quad n = 1, 2, \ldots, P \tag{1}$$

where $\{a(n)\}_{n=0}^{P}$ are the quantized LP coefficients of the synthesis filter and P is the order of the filter.

In general, speaker verification systems use features representing the vocal tract only. However, the human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information (e.g. phonation, respiratory, and mixed-voiced and unvoiced, etc.). Therefore, verification performance can be improved if features derived from the excitation signals are also employed. The most obvious information that can be extracted from the excitation signals is the pitch period. It has been shown that pitch information is speaker dependent [11] and that by using a Bayesian network, density functions of pitch period can be combined with those of spectral envelopes for speaker verification [12]. Instead of extracting the pitch period from the excitation signals, we may consider the excitation signals as the additional information that is not encapsulated in the spectral parameters such as MFCCs and LPCCs. Research in speech coding has shown that proper encoding of LP-residues is detrimental to the quality of synthetic speech. It has also been shown that LP-residues of speech contains speaker-dependent information [13]. More interestingly, humans can recognize individuals by listening to LP-residues alone [14]. All of these evidences suggest the promise of LP-residues in speaker recognition.

In order to extract the additional speaker-dependent information from the coder parameters, we first reconstructed the LP residuals from the optimum vectors of the fixed, adaptive, and gain codebooks. Then, for each segment (10 ms for G.729 and 30 ms for G.723.1) of the residual signals, LP analysis was performed using the autocorrelation method [10] with a 30 ms asymmetric window for G.729 and a 22.5 ms Hamming window for G.723.1. The autocorrelation coefficients of windowed residual signals were computed and converted to LP coefficients using the Levinson-Durbin algorithm. A set of LPCCs, $\mathbf{c}_r = [c_r(1), \ldots, c_r(12)]^T$, can then be computed from the LP coefficients of the residual signals using a recursive formula similar to (1). As the synthetic speech is the convolution of the residual signals and the vocal-tract impulse response, we added the cepstrum of residuals to that of the vocal-tract transfer function. More specifically, the proposed feature vectors are defined as:

$$\mathbf{c} = \mathbf{c}_v + \mathbf{c}_r \tag{2}$$

where $\mathbf{c} = [c(1) \ldots c(12)]^T$.

3. Handset Mismatch Compensation

We extended our recently proposed feature transformation technique [4] to handset- and coder-mismatch compensation. The key idea is to transform the distorted features to fit the clean speech models. Specifically, for each speech frame, the recovered cepstral coefficients are determined by:

$$\hat{c}(n) = \sum_{k=1}^{K} g_k(\mathbf{c}) \left[\alpha_{k,n} \left(c(n)\right)^2 + \beta_{k,n} c(n) + \gamma_{k,n}\right] \tag{3}$$

where $\alpha_{k,n}, \beta_{k,n}$ and $\gamma_{k,n}$, $n = 1, \ldots, 12$, are the transformation parameters, $c(n)$ is the n-th component of the distorted

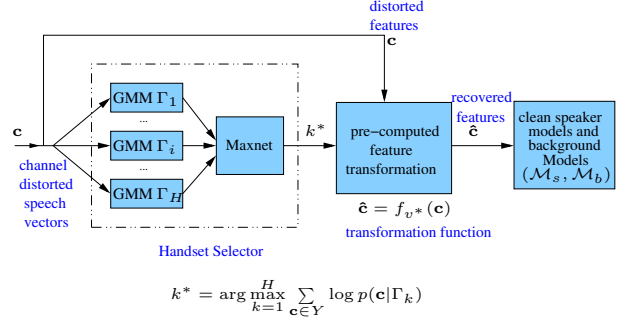$$k^* = \arg \max_{k=1}^{H} \sum_{\mathbf{c} \in Y} \log p(\mathbf{c}|\Gamma_k)$$

Figure 1: Combining handset identification and feature transformation for speaker verification.

cepstral vector $\mathbf{c}$, $\hat{c}(n)$ is the recovered cepstral component, and $g_k(\mathbf{c})$ is the posterior probability of selecting the k-th transformation given the distorted vector $\mathbf{c}$. These transformation parameters can be determined by maximizing the likelihood of the distorted data given a clean speech model [4].

As the transformation parameters are handset-dependent, it is necessary to determine one set of transformation parameters for each type of handsets that will be used by the claimants. A handset selector [5] will also be needed to identify the handset types used by the claimants during verification so that an appropriate set of transformation parameters can be selected. The combination of handset identification and feature transformation is illustrated in Fig. 1.

4. Experiments and Results

4.1. Speech Corpus and Features

To evaluate the proposed technique, the uncoded HTIMIT corpus [9] and its transcoded variants were employed. HTIMIT was obtained by playing a subset of the TIMIT corpus through a set of telephone handsets (cb1-cb4, el1-el4, and pt1) and a Sennheizer head-mounted microphone (senh). Speakers in the corpus were divided into a speaker set (50 male and 50 female) and an impostor set (25 male and 25 female). The SA and SX sentence sets of Handset senh were used for enrollment, while the SI sentence sets from all handsets were used for verification. As a result, we were able to compare the coder-derived features under both handset matched and handset mismatched conditions.

There are three different feature representations in the experiments:

- Feature A: LPCCs obtained from the front-end of the LP coefficients in the encoder (Section 3.2.1-2 of [7] and Section 2.4 of [8]).

- Feature B: LPCCs computed from the quantized LSFs in the decoders (i.e. $\mathbf{c}_v$ as defined in (1)).

- Feature C: Sum of Feature B and the LPCCs derived from the residual $\mathbf{c}_r$ (i.e. $\mathbf{c}$ as defined in (2)).

To obtain Feature A, the speech in HTIMIT was first filtered with a 2nd-order pole/zero highpass filter with a cutoff frequency at 140 Hz. LP analysis was performed once per 10 ms for G.729-transcoded speech using the autocorrelation method with a 30 ms asymmetric window, while it was performed once

per 30 ms for G.723.1-transcoded speech with a 22.5 ms Hamming window. The 10th-order LPCCs were then computed from the LP coefficients.

4.2. Speaker Enrollment and Verification

For each of the feature sets (A, B and C) mentioned in Section 4.1, we used the SA and SX sentence sets of Handset senh to train 100 personalized 32-center GMMs ($\mathcal{M}_{s,\omega}$ where $\omega \in \{A, B, C\}$) that model the characteristics of the speakers in the speaker set. All of the 100 speakers in the speaker set were also used to train a 64-center GMM background models for each feature set, i.e. $\mathcal{M}_{b,\omega}$ where $\omega \in \{A, B, C\}$. The background models were shared by all speaker models during verification.

The clean utterances of 10 speakers from the speaker set were used to create a 2-center GMM (Λ_X) clean model. Using this model and the estimation algorithms described in [4], a set of feature transformation parameters $\nu = \{\alpha_{k,n}, \beta_{k,n}, \gamma_{k,n}\}$ were computed for each handset. In particular, the SA and SX utterances from handset "senh" were considered as clean and were used to create Λ_X, while the SA and SX utterances spoken by the same 10 speakers but using other 9 handsets (cb1-cb4, el1-el4, and pt1) were used as the distorted speech. Note that the objective is to model the statistical difference between the clean and distorted speech, not the statistical characteristics of these 10 speakers. Therefore, it is not absolutely necessary to choose these speakers from a held-out set. Note also that only the training utterances from the enrollment sessions were used for estimating the transformation parameters.

During verification, a vector sequence Y derived from a claimant's utterance (SI sentence) was fed to the handset selector [5]. According to the outputs of the handset selector, a set of transformation parameters was selected. The features were transformed and then fed to the speaker model ($\mathcal{M}_{s,\omega}$) corresponding to the claimed identity to obtain a score ($\log p(Y|\mathcal{M}_{s,\omega})$), which was then normalized according to

$$S(Y) = \log p(Y|\mathcal{M}_{s,\omega}) - \log p(Y|\mathcal{M}_{b,\omega}) \qquad (4)$$

where $\mathcal{M}_{b,\omega}$ represents the background model and $\omega \in \{A, B, C\}$. The normalized score $S(Y)$ was compared with a threshold to make a verification decision. In this work, the threshold for each speaker was adjusted to determine the equal error rate (EER).

4.3. Results and Discussions

The experimental results for G.729- and G.723.1-transcoded speech are summarized in Table 1. A baseline experiment (without using the handset selectors and feature transformations) and an experiment using CMS as channel compensation were also conducted for comparison. All error rates are based on the average of 100 genuine speakers and 50 impostors. Average EERs for the three different features are plotted in Fig. 2. The average EER is computed by taking the average of all the EERs corresponding to different handsets.

It is obvious from Table 1 that CMS degrades the performance of the system when the enrollment and verification sessions use the same handset (senh). In the matched handset condition, we found that the transformation technique is slightly inferior to the baseline in all the three feature representations. This is because a few utterances were considered as recorded from handsets other than senh. On the other hand, when feature transformation is employed under handset mismatched conditions, the handset selectors are able to detect the
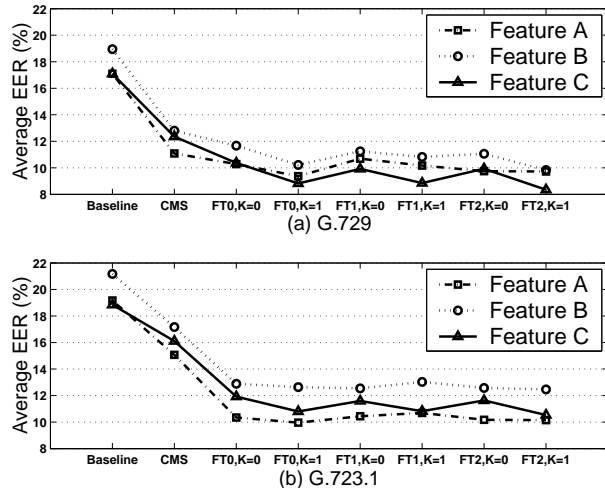

Figure 2: Average EERs (over both matched and mismatched conditions) achieved by the three different feature representations using the baseline, CMS and feature transformation. FTx, where $x = 0, 1, 2$, denotes the order of feature transformation in (3), and K denotes the number of transformations.

most likely handset and facilitate the subsequent transformation of the distorted features. As a result, the transformation technique achieves significant error reduction as compared to the baseline and CMS for all the three feature representations under mismatched conditions.

As shown in Fig. 2, the verification performance becomes poorer when the feature vectors are obtained from the quantized LSFs (Feature B). On the other hand, Feature C is superior to Feature B and in some cases even better than Feature A. Since additional speaker characteristics can be found in the excitation signals, the proposed feature vectors (Feature C), which take account of the LP residuals, can further improve the verification performance. We can also notice from Fig. 2 that transforming the features derived from the G.729 codecs can improve the performance to a level beyond the one attainable by extracting LPCCs from the encoders' front-end.

5. Conclusions

A technique that improves the performance of speaker verification from G.729 and G.723.1 coded telephone speech is proposed. In this technique, additional speaker-dependent information that takes account of the LP-residual spectra is incorporated into the proposed feature representation. A new channel compensation approach for verifying speakers from coded telephone speech has also been presented. Results show that adding the LPCCs of LP-residuals to the coder-derived LPCCs achieves the best result. It was also found that the transformation technique outperforms the CMS approach and significantly reduces the error rates of a baseline system.

6. References

[1] J. M. Huerta and R. M. Stern, "Speech recognition from GSM coder parameters," in *Proc. 5th Int. Conf. on Spoken Language Processing*, 1998, vol. 4, pp. 1463–1466.

[2] T. F. Quatieri, E. Singer, R. B. Dunn, D. A. Reynolds, and

	Matched handset (senh)			Mismatched handsets (cb1–pt1)		
Trans. Method	Feature A	Feature B	Feature C	Feature A	Feature B	Feature C
Baseline	3.54	4.02	3.87	17.10	18.95	17.10
CMS	5.50	5.77	5.89	11.08	12.80	12.35
FT0, K=1	3.64	4.00	3.93	10.28	11.67	10.37
FT0, K=2	3.94	4.33	4.13	9.37	10.22	8.82
FT1, K=1	3.79	4.17	4.06	10.69	11.25	9.92
FT1, K=2	4.10	4.55	4.31	10.17	10.82	8.85
FT2, K=1	3.82	4.15	4.17	9.75	11.06	9.93
FT2, K=2	4.02	4.43	4.14	9.72	9.83	8.35

(a) G.729

	Matched handset (senh)			Mismatched handsets (cb1–pt1)		
Trans. Method	Feature A	Feature B	Feature C	Feature A	Feature B	Feature C
Baseline	4.70	6.67	6.46	19.17	21.17	18.85
CMS	8.54	10.47	10.89	15.07	17.17	16.11
FT0, K=1	4.84	6.68	6.64	10.34	12.89	11.92
FT0, K=2	4.81	6.61	6.76	9.96	12.64	10.80
FT1, K=1	4.87	6.72	6.64	10.44	12.54	11.59
FT1, K=2	5.28	7.40	6.75	10.69	13.03	10.82
FT2, K=1	5.01	6.69	6.83	10.17	12.57	11.63
FT2, K=2	4.78	6.53	6.75	10.15	12.46	10.53

(b) G.723.1

Table 1: Equal error rates (in %) for matched and mismatched handsets achieved by using Feature A (LPCCs obtained from the encoders' front-end), Feature B (LPCCs computed from quantized LSFs) and Feature C (sum of Feature B and the LPCCs derived from the LP-residuals). The columns "Mismatch handsets (cb1–pt1)" list the average EERs of the mismatched handsets: cb1, cb2, cb3, cb4, el1, el2, el3, el4, and pt1. FTx, where $x = 0, 1, 2$, denotes the order of feature transformation in (3), and K denotes the number of transformations.

J. P. Campbell, "Speaker and language recognition using speech codec parameters," in *Proc. Eurospeech'99*, 1999, vol. 2, pp. 787–790.

[3] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, J. P. Campbell, and E. Singer, "Speaker recognition using G.729 codec parameters," in *Proc. ICASSP'2000*, 2000, pp. 89–92.

[4] M. W. Mak and S. Y. Kung, "Combining stochastic feautre transformation and handset identification for telephone-based speaker verification," in *Proc. ICASSP'2002*, 2002.

[5] C. L. Tsang, M. W. Mak, and S. Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," in *Proc. ICSLP'02*, 2002, pp. 2329–2332.

[6] W. M. Yu, M. W. Mak, and S. Y. Kung, "Speaker verification from coded telephone speech using stochastic feature transformation and handset identification," in *Pacific-Rim Conference on Multimedia 2002*, 2002, pp. 598–606.

[7] International Telecommunication Union, *ITU-T Recommendation G.729: Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*, 1996.

[8] International Telecommunication Union, *ITU-T Recommendation G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, 1996.

[9] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.

[10] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, New York: Springer-Verlag, 1976.

[11] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Int. Conf. on Spoken Language Processing*, 1996, vol. 3, pp. 1800–1803.

[12] M. Arcienega and A. Drygajlo, "A Bayesian network approach for combining pitch and spectral envelope for speaker verification," in *COST 275 Workshop - The Advent of Biometrics on the Internet*, Rome, Nov. 2002, pp. 99–102.

[13] P. Thevnaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communications*, vol. 17, pp. 145–157, 1995.

[14] T. C. Feustel, G. A. Velius, and R. J. Logan, "Human and machine performance on speaker identity verifiation," *Speech Technology*, vol. 89, pp. 169–170, 1989.