SPEAKER VERIFICATION FROM CODED TELEPHONE SPEECH USING STOCHASTIC FEATURE TRANSFORMATION AND HANDSET IDENTIFICATION

Eric W. M. Yu, Man-Wai Mak, and Sun-Yuan Kung *

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

Abstract. A handset compensation technique for speaker verification from coded telephone speech is proposed. The proposed technique combines handset selectors with stochastic feature transformation to reduce the acoustic mismatch between different handsets and different speech coders. Coder-dependent GMM-based handset selectors are trained to identify the most likely handset used by the claimants. Stochastic feature transformations are then applied to remove the acoustic distortion introduced by the coder and the handset. Experimental results show that the proposed technique outperforms the CMS approach and significantly reduces the error rates under six different coders with bit rates ranging from 2.4 kb/s to 64 kb/s. Strong correlation between speech quality and verification performance is also observed.

1 Introduction

Due to the proliferation of electronic banking and electronic commerce, recent research has focused on verifying speakers' identity over the telephone. A challenge of telephone-based speaker verification is that transducer variability could 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 resolving the mismatch problem. This approach includes cepstral mean subtraction (CMS) [1] and signal bias removal [2], which approximate a linear channel by the long-term average of distorted cepstral vectors. However, they do not consider the effect of background noise. The codeword-dependent cepstral normalization (CDCN) [3] is a

* This work was supported by The Hong Kong Polytechnic University, Grant No. A442 and RGC Project No. PolyU 5129/01E. S. Y. Kung is on sabbatical from Princeton University, USA. He is currently a Distinguished Chair Professor of The Hong Kong Polytechnic University.

more general approach that accounts for the effect of background noise. However, it works well only when the noise level is low.

A technique that combines stochastic feature transformation and handset identification was proposed in [4] for the compensation of channel mismatch in telephone-based speaker verification. It was demonstrated that the technique can significantly reduce verification error rate.

As a result of the popularity of digital communication systems, there has been increasing interest in the automatic recognition of resynthesized coded speech [5], [6], [7]. For example, speaker verification based on GSM, G.729, and G.723.1 resynthesized speech was studied in [6]. It was shown that the verification performance generally degrades with coders' bit rate. As the perceptual quality of coded speech generally decreases with coders' bit rate, the verification performance decreases with decreasing perceptual quality of speech. To improve the verification performance of G.729 coded speech, techniques that require knowledge of the coder parameters and coder internal structure were proposed in [6] and [7]. However, the performance of these improved techniques is still poorer than that achieved by using resynthesized speech.

As [6] and [7] focus on using coder parameters and pitch information for speaker verification, channel compensation was limited to CMS and RASTA processing. This paper, on the other hand, applies a more advanced channel compensation technique [4] for speaker verification over a digital communication network. As the technique operates directly on the coded telephone speech, no access to the coder parameters and structure will be required. In order to study the performance on coded speech with a wide range of compression ratios, six coders (G.711, G.726, GSM, G.729, G.723.1, and LPC) were employed to generate the coded speech.

Unlike [4], where LP-derived cepstral coefficients (LPCC) were used as features, we employed mel-frequency cepstrum coefficients (MFCC) [8] as the feature vectors in this work. Speaker verification results based on uncoded and coded corpora are presented. Results using CMS as channel compensation are also shown for comparison.

2 Stochastic Feature Transformation

Stochastic matching [9] is a popular approach to speaker adaptation and channel compensation. Its main idea is to transform distorted data to fit the clean speech models or to transform the clean speech models to better fit the distorted data. In the case of feature transformation, the channel is represented by either a single cepstral bias ($\mathbf{b}$) or a bias together with an affine transformation matrix (A). In the latter case, the component-wise form of the transformed vectors is given by

$$\hat{x}_{t,i} = f_\nu(\mathbf{y}_t)_i = a_i y_{t,i} + b_i \tag{1}$$

where $\mathbf{y}_t$ is a D-dimensional distorted vector, $\nu = \{a_i, b_i\}_{i=1}^{D}$ is the set of transformation parameters, and f_ν denotes the transformation function. Intuitively, the bias $\{b_i\}$ compensates the convolutive distortion and the parameters $\{a_i\}$ compensate the effects of noise.

In this work, we will consider the bias term only (i.e. $a_i = 1$ for all i) because our previous results [4] have shown that the zero- and 1st-order transformations achieve a comparable error reduction.

Given a clean GMM speech model

$$\Lambda_X = \{\omega_j^X, \mu_j^X, \Sigma_j^X\}_{j=1}^M \tag{2}$$

derived from the clean speech of several speakers (ten speakers in this work) and distorted speech $\mathbf{y}_t$, $t = 1, \ldots, T$, the maximum likelihood estimates of ν can be obtained by maximizing an auxiliary function

$$\begin{aligned} Q(\nu'|\nu) &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \cdot \log\left\{\omega_j^X p(\mathbf{y}_t|\mu_j^X, \Sigma_j^X, \nu')\right\} \\ &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \cdot \log\left\{\omega_j^X p(f_{\nu'}(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X) \cdot |J_{\nu'}(\mathbf{y}_t)|\right\} \end{aligned} \tag{3}$$

with respect to ν'. In (3), ν' and ν represent respectively the new and current estimates of the transformation parameters, T is the number of distorted vectors, $\nu' = \{b_i'\}_{i=1}^D$ denotes the transformation, $|J_{\nu'}(\mathbf{y}_t)|$ is the determinant of the Jacobian matrix whose (r, s)-th entry is given by $J_{\nu'}(\mathbf{y}_t)_{rs} = \partial f_{\nu'}(\mathbf{y}_t)_s / \partial y_{t,r}$, and $h_j(f_\nu(\mathbf{y}_t))$ is the posterior probability given by

$$h_j(f_\nu(\mathbf{y}_t)) = P(j|\Lambda_X, \mathbf{y}_t, \nu) = \frac{\omega_j^X p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X)}{\sum_{l=1}^M \omega_l^X p(f_\nu(\mathbf{y}_t)|\mu_l^X, \Sigma_l^X)} \tag{4}$$

where $\{\omega_j^X\}_{j=1}^M$ are the mixing coefficients in Λ_X and

$$\begin{aligned} p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X) &= (2\pi)^{-\frac{D}{2}} |\Sigma_j^X|^{-\frac{1}{2}} \\ &\cdot \exp\{-\tfrac{1}{2}(f_\nu(\mathbf{y}_t) - \mu_j^X)(\Sigma_j^X)^{-1}(f_\nu(\mathbf{y}_t) - \mu_j^X)\}. \end{aligned} \tag{5}$$

Ignoring the terms independent of ν' and assuming diagonal covariance (i.e. $\Sigma_j^X = \text{diag}\left\{(\sigma_{j1}^X)^2, \ldots, (\sigma_{jD}^X)^2\right\}$), (3) can be written as

$$Q(\nu'|\nu) = \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \left\{-\frac{1}{2} \sum_{i=1}^D \frac{(y_{t,i} + b_i' - \mu_{ji}^X)^2}{\left(\sigma_{ji}^X\right)^2}\right\} \tag{6}$$

In the M-step of each EM iteration, we maximize $Q(\nu'|\nu)$ to obtain

$$\mathbf{b}' = \frac{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t))(\Sigma_j^X)^{-1}(\mu_j^X - \mathbf{y}_t)}{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t))(\Sigma_j^X)^{-1}} \tag{7}$$

where $f_\nu(\mathbf{y}_t) = \mathbf{y}_t + \mathbf{b}$, μ_j^X and Σ_j^X, $j = 1, \ldots, M$, are the mean vectors and covariance matrices of an M-center Gaussian mixture model (Λ_X) representing the clean speech.

3 Handset Selector

Unlike speaker adaptation where the transformation parameters can be estimated during recognition, in speaker verification we need to estimate the transformation parameters before verification takes place. This is because we do not know the claimant's identity in advance. If the transformation parameters are estimated based on claimant's speech obtained in a single verification session only, all the transformed vectors, regardless of the claimant's genuineness, will be mapped to a region very close to the claimed model in the clean feature space. As a result, the claimant will likely be accepted regardless of whether he/she is a genuine speaker or an impostor.

Therefore, to apply stochastic transformation to telephone-based speaker verification, we need to derive one set of transformation parameters for each type of handsets. During verification, the transformation parameters corresponding to the most likely handset are used to transform the distorted features. This can be achieved by applying our recently proposed handset selector [10]. Specifically, each handset is associated with one set of transformation parameters; during verification, an utterance of claimant's speech is fed to H GMMs (denoted as $\{\Gamma_k\}_{k=1}^{H}$). The most likely handset is selected according to

$$k^* = \arg\max_{k=1}^{H} \sum_{t=1}^{T} \log p(\mathbf{y}_t|\Gamma_k) \tag{8}$$

where $p(\mathbf{y}_t|\Gamma_k)$ is the likelihood of the k-th handset. Then, the transformation parameters corresponding to the k^*-th handset are used to transform the distorted vectors.

4 Experiments and Results

Uncoded and Coded Corpora: In this work, the HTIMIT corpus [11] and six coded HTIMIT corpora containing resynthesized coded speech were used to evaluate the feature transformation technique. The HTIMIT corpus 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). Each speaker was assigned a personalized 32-center GMM that models the characteristics of his/her own voice. For each GMM, the feature vectors derived from the SA and SX sentence sets of the corresponding speaker were used for training. A collection of all speakers in the speaker set was used to train a 64-center GMM background model ($\mathcal{M}_b$). The handset "senh" was used as the enrollment handset.

To evaluate the performance of the feature transformation technique on the coded HTIMIT corpora, six different codecs were employed in this work: G.711 at 64 kb/s, G.726 at 32 kb/s, GSM at 13 kb/s, G.729 at 8 kb/s, G.723.1 at 6.3 kb/s, and LPC at 2.4 kb/s. Six sets of coded corpora were obtained by coding the speech in HTIMIT using these coders. The encoded utterances were then decoded to produce resynthesized speech. Feature vectors were extracted from

each of the utterances in the uncoded and coded corpora. The feature vectors were 12-th order mel-frequency cepstrum coefficients (MFCC) [8]. These vectors were computed at a frame rate of 14 ms using a Hamming window of 28 ms.

Feature Transformation: The uncoded clean utterances of 10 speakers were used to create a 2-center GMM (Λ_X) clean model (i.e. $M = 2$ in (2)). Using this model and the estimation algorithms described in Section 2, a set of coder-dependent feature transformation parameters ν were computed for each handset in each coded corpus. In particular, the utterances from handset "senh" were considered as clean and were used to create Λ_X, while those from other 8 handsets (cb1-cb4, el1-el3, and pt1) were used as distorted speech. As the experimental results in [4] show that the difference in error rates is not significant among stochastic transformations with zero-th, 1-st and 2-nd order, we used zero-th order transformations for all handsets and coders in this work.

Coder-Dependent Handset Selectors: Six handset selectors, each of them consisting of ten GMMs $\{\Gamma_k^{(i)}; i = 1, \ldots, 6 \text{ and } k = 1, \ldots, 10\}$, were constructed from the SA and SX sentence sets of the coded corpora. For example, GMM $\Gamma_k^{(i)}$ represents the characteristics of speech derived from the k-th handset of the i-th coded corpus. As we assume that in most practical situations the receiver will know the type of coders being used (otherwise it will not be able to decode the speech), there will not be any error in choosing the handset selector. The only error that will be introduced is the incorrect decisions made by the chosen handset selector. This error, however, is very small, as demonstrated in the latter part of this paper.

Verification Procedures: During verification, a vector sequence $\mathbf{Y}$ derived from a claimant's utterance (SI sentence) was fed to a coder-dependent handset selector corresponding to the coder being used by the claimant. According to the outputs of the handset selector (8), a set of coder-dependent transformation parameters was selected. The features were transformed and then fed to a 32-center GMM speaker model ($\mathcal{M}_s$) to obtain a score ($\log p(\mathbf{Y}|\mathcal{M}_s)$), which was then normalized according to

$$S(\mathbf{Y}) = \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b) \tag{9}$$

where $\mathcal{M}_b$ represents the background model. The normalized score $S(\mathbf{Y})$ 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). Similar to [12], the vector sequence was divided into overlapping segments to increase the resolution of the error rates.

Verification Results: The experimental results are summarized in Tables 1, 2, and 3. 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

Codec	Equal Error Rate (%)								
	cb1	cb2	cb3	cb4	el1	el2	el3	pt1	senh
Uncoded (128 kb/s)	4.85	5.67	21.19	16.49	3.60	11.11	5.14	11.74	1.26
G.711 (64 kb/s)	4.88	5.86	21.20	16.73	3.67	11.08	5.21	12.04	1.34
G.726 (32 kb/s)	6.36	8.71	22.67	19.61	6.83	14.98	6.68	16.42	2.66
GSM (13 kb/s)	6.37	6.10	19.90	15.93	6.21	17.93	9.86	16.42	2.35
G.729 (8 kb/s)	6.65	4.59	20.15	15.08	6.18	14.28	6.71	11.93	2.67
G.723.1 (6.3 kb/s)	7.33	5.49	20.83	15.59	6.56	14.71	6.58	14.03	3.30
LPC (2.4 kb/s)	10.81	10.30	29.68	24.21	8.56	19.29	10.56	14.97	3.43

Table 1. Equal error rates (in %) achieved by the baseline approach (without handset selectors and feature transformation) on speech corpora coded by different coders. The enrollment handset is "senh".

Codec	Equal Error Rate (%)								
	cb1	cb2	cb3	cb4	el1	el2	el3	pt1	senh
Uncoded (128 kb/s)	4.00	3.02	10.69	6.62	3.36	5.16	5.67	5.67	3.67
G.711 (64 kb/s)	4.06	3.07	10.73	6.70	3.43	5.26	5.74	5.84	3.75
G.726 (32 kb/s)	5.65	4.42	11.78	8.00	5.61	7.95	6.97	9.07	5.12
GSM (13 kb/s)	5.25	4.10	11.32	8.00	4.95	7.04	7.47	7.58	4.73
G.729 (8 kb/s)	5.43	4.37	11.81	7.98	5.16	7.38	7.32	7.21	4.69
G.723.1 (6.3 kb/s)	6.40	4.60	12.36	8.53	6.11	8.50	7.31	8.28	5.62
LPC (2.4 kb/s)	6.34	5.51	14.10	9.22	6.35	8.95	8.95	9.55	4.57

Table 2. Equal error rates (in %) achieved by the cepstral mean subtraction (CMS) approach on speech corpora coded by different coders. The enrollment handset is "senh".

genuine speakers. Average EERs of the uncoded and coded corpora are plotted in Figure 1. The average EER of a corpus is computed by taking the average of all the EERs corresponding to different handsets of the corpus.

The results show that the transformation technique achieves significant error reduction for both uncoded and coded corpora. In general, the transformation approach outperforms the CMS approach except for the LPC coded corpus. From the results in Table 1, we observe that the error rates of LPC coded corpus are relatively high before channel compensation is applied. An informal listening test reveals that the perceptual quality of LPC coded speech is very poor, which means that most of the speaker's characteristics have been removed by the coding process. This may degrade the performance of the transformation technique.

Nowadays, G.711 and GSM coders are widely used in fixed-line and mobile communication networks respectively, and G.729 and G.723.1 have become standard coders in teleconferencing systems. These are the areas where speaker verification is useful. LPC coders, on the other hand, are mainly employed in applications where speaker verification is not very important (e.g., toys). As the feature transformation technique outperforms CMS in areas where speaker verification is more important, it is a better candidate for compensating coder- and channel-distortion in speaker verification systems.

Codec	Equal Error Rate (%)									Handset Selector Accuracy (%)
	cb1	cb2	cb3	cb4	el1	el2	el3	pt1	senh	
Uncoded (128 kb/s)	1.63	1.27	9.65	4.47	1.41	3.58	3.37	3.08	1.09	97.92
G.711 (64 kb/s)	1.52	1.26	9.57	4.53	1.41	3.53	3.33	3.21	1.17	98.02
G.726 (32 kb/s)	2.55	2.55	11.66	6.05	2.74	6.19	4.17	5.82	2.29	97.73
GSM (13 kb/s)	3.13	2.44	11.13	7.10	3.10	6.34	6.29	5.58	2.67	96.91
G.729 (8 kb/s)	3.94	3.27	9.99	6.63	4.18	6.17	6.20	4.70	2.89	96.39
G.723.1 (6.3 kb/s)	3.94	3.42	10.74	6.83	4.49	6.70	5.80	5.71	3.41	96.27
LPC (2.4 kb/s)	5.68	5.93	17.33	11.05	7.14	10.50	9.34	8.89	3.95	94.39

Table 3. Equal error rates (in %) achieved by combining 0th-order stochastic transformation with coder-dependent handset selectors on speech corpora coded by different coders. The accuracy achieved by the handset selectors is also shown. The enrollment handset is "senh".

It is obvious from the last column of Table 1 and Table 2 that CMS degrades the performance of the system when the enrollment and verification sessions use the same handset (senh). When the transformation technique is employed under this matched condition, the handset selectors are able to detect the most likely handset (i.e. senh) and faciliate the subsequent transformation of the distorted features. As a result, the error rates become very close to the baseline.

As observed from the experimental results, verifcation based on uncoded telephone speech performs better than that based on coded telephone speech. However, since the distortion introduced by G.711 is very small, the error rates of uncoded and G.711 coded corpora are similar.

In general, the verification performance of the coded corpora degrades when the bit rate of the corresponding codec decreases (Figure 1). However, the performance among the GSM, G.729, and G.723.1 coded speech does not obey this rule occasionally for some handsets. After CMS was employed for channel compensation, the error rates were reduced for all the uncoded and coded corpora while a stronger correlation between bit rates and verification performance can be observed among the GSM, G.729, and G.723.1 coded speech. Using the transformation technique, the error rates are reduced further while correlation between bit rates and verification performance becomes very obvious among the coded speech at various bit rates. As the perceptual quality of the coded speech is usually poorer for lower rate codecs, we conclude that a strong correlation between the coded speech quality and the verification performance exists.

Comparing with the results in [4], it is obvious that using MFCC as features is more desirable than using LPCC. For example, when MFCC are used the average error rate for the uncoded speech is 9.01%, whereas the error rate increases to 11.16% when LPCC are used [4].

5 Conclusions

A new channel compensation approach for verifying speakers from coded telephone speech has been presented. The proposed approach combines stochastic transformation with handset identification. Results show that the transforma-

tion technique outperforms the CMS approach and significantly reduces the error rates of a baseline system. The error rate achieved by the transformation technique correlates with the bit rate of the codec and hence reflects the perceptual quality of the coded speech. In this work, we also observed that MFCC outperform LPCC in representing speakers' characteristics.

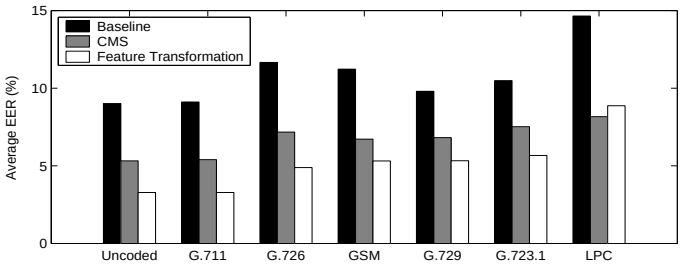

Fig. 1. Average EERs achieved by the baseline, CMS, and transformation approaches. Note that the bit rate of coders decreases from left to right, with "uncoded" being the highest (128 kb/s) and LPC the lowest (2.4 kb/s).

References

1. B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.
2. M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp. 19–30, Jan 1996.
3. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Pub., Dordrecht, 1992.
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. 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.
6. T. F. Quatieri, E. Singer, R. B. Dunn, D. A. Reynolds, and J. P. Campbell, "Speaker and language recognition using speech codec parameters," in *Proc. Eurospeech'99*, 1999, vol. 2, pp. 787–790.
7. 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.
8. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, August 1980.
9. A. Sankar and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.
10. K. K. Yiu, M. W. Mak, and S. Y. Kung, "A GMM-based handset selector for channel mismatch compensation with applications to speaker identification," in *2nd IEEE Pacific-Rim Conference on Multimedia*, 2001, pp. 1132–1137.
11. D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.
12. M. W. Mak and S. Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithms with application to speaker verification," *IEEE Trans. on Neural Networks*, vol. 11, no. 4, pp. 961–969, 2000.