IMPROVEMENT OF SPEAKER RECOGNITION BY COMBINING RESIDUAL AND PROSODIC FEATURES WITH ACOUSTIC FEATURES

Shi-Han Chen and Hsiao-Chuan Wang

Dept. Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

When a speech signal is encoded in some low bit-rate coding formats, it becomes more difficult in distinguishing the speaker identities. This paper investigates the codec effect to acoustic and prosodic features. A new representation of prosodic features based on the piecewise fitting of the pitch contour is introduced. A method for including residual features based on LDA algorithm is suggested. By combining prosodic features with the acoustic features, we can improve the performance of speaker recognition system. A series of experiments is performed with coded speech affected by G.729A and GSM codec processes to demonstrate the effectiveness of our proposed method.

1. INTRODUCTION

The low bit-rate speech coding is achieved by lossy representation and deconstructed quantization of the original speech. This introduces the distortion to the speech signal even the distortion is unnoticed by human ears. The performance of speech or speaker recognition on the coded speech will be degraded because of the effect of codec process. Many works had focused on the issues of how to recover the performance of speech or speaker recognition in low bit-rate coded speech [1]~[6]. Some works has shown that the residual information is very useful for speaker recognition especially for coded speech. [7][8]

In this paper, we aim at the problems related to speaker identification and verification. At first, we examine the spectral distortion and the pitch variation due to codec process. Then we suggest that residual features as well as the prosodic features must be included in the speaker recognition process. A new representation of prosodic features is introduced.

Combining prosodic features with acoustic features is a way to improve the performance of speaker recognition. A series of experiments is performed with coded speech affected by G.729A and GSM codec process to demonstrate the effectiveness of our proposed method.

2. EFFECT OF CODEC PROCESS

In this study, two codec processes are examined. They are full-rate GSM speech coder and G.729A speech coder. The input speech signal is sampled in 8 kHz and represented in linear PCM format. The coding scheme of GSM is the regular pulse excitation—long term prediction (RPE-LTP). Reflection coefficients are calculated, transformed, and quantized into 8 LARs for transmission. The G.729A coder is based on the scheme of conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). LP coefficients are converted into line spectrum pairs (LSPs) and quantized using predictive two-stage vector quantization.

In order to understand the effect of codec process, we measure the spectral distortion by the following equation;

$$SD(f) = \frac{H_{original}(f)}{H_{decoded}(f)}, \qquad (1)$$

where $H_{original}(f)$ and $H_{decoded}(f)$ are the LP-based spectra obtained from original speech signal and decoded speech signal, respectively. An experiment is done on 2700 speech frames with GSM and G.729A coding schemes. The resulted spectral distortions for the case of GSM codec is plotted in Figure 1.

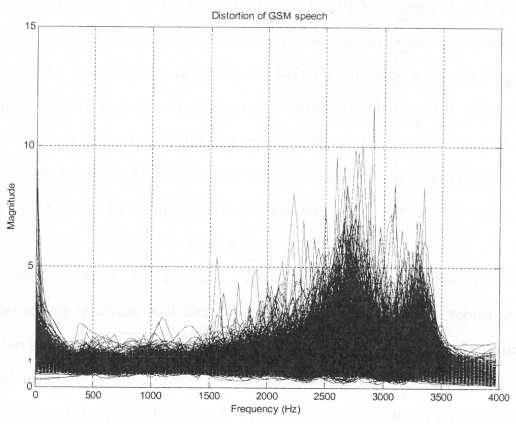
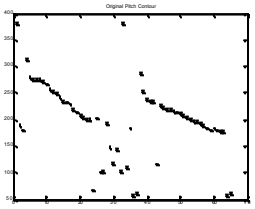
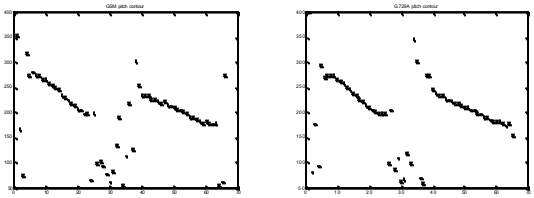


Figure 1. Spectral distortion due to GSM coding

The figure shows that the serious distortion appears at frequencies below 100 Hz and beyond 2 kHz. The similar situation happens in case of G.729A codec process. The spectral distortion at high frequencies will degrade the performance of speaker recognition since high frequency components carry more information for distinguishing speaker identities. However, pitch contours are much more robust to the codec process as shown in Figure 2.



(a) Original speech



(b) GSM coded speech (c) G.729A coded speech

Figure 2. Robustness of pitch contours

3. RESIDUAL FEATURES

The LP analysis of speech signal results in an all-pole model for representing the speech production. The LP coefficients are the acoustic features which are used for describing the speech spectrum. Once the LP coefficients are available, the residual signal can be obtained by the following equation,

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k), \qquad (2)$$

where $\{a_k\}$ are LP coefficients and p is the order of the LP model. The residual signal carries the information of excitation which is useful for distinguishing speaker identities.

If we perform LP analysis on the residual signal, we can get another set of LP coefficients and result in another residual signal.

$$d(n) = e(n) - \hat{e}(n) = e(n) - \sum_{k=1}^{q} b_k e(n-k), \qquad (3)$$

where $\{b_k\}$ are LP coefficients of residual signal $e(n)$, and q is the order of this LP model. Hence, residual features are represented by a set of LP coefficients $\{b_k\}$.

LP coefficients can be converted into cepstral coefficients which are called the LP-based cepstral coefficients (LPCCs). We can also use LP coefficients to construct an all-pole model and find its spectrum. This spectrum is exactly the spectrum of the model output signal. By applying a 20-band mel-scale filter-bank to the spectrum, we can compute mel-frequency cepstral coefficients (MFCCs). All the cepstral coefficients are mean subtracted to eliminate the channel effect.

4. PROSODIC FEATURES

The residual signal obtained by Eq(3) is low-passed and then the autocorrelation is calculated. For each frame, locations of top 3 maxima of the autocorrelation function indicate the pitch candidates. A pitch tracking algorithm based on Viterbi search is applied to find the pitch contour. The pitch contour exists only in the segments of voiced speech. The region of no pitch contour is the segment of speech-pause or unvoiced speech. For each voiced segment, the MSE-based piecewise linear stylization [9] is performed to convert the pitch contour into sections of straight lines. Figure 3 illustrates the linear fitting of pitch contours.

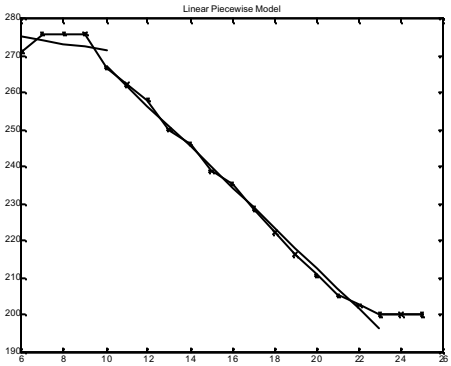


Figure 3. Piecewise linear fitting of pitch contour

Based on the segmented pitch contour, we define the following parameters;

(1) Pitch related features

log-p (pitch of a frame),
log-P (mean of pitch in a segment),
log-max (maximum pitch in a segment),
log-min (minimum pitch in a segment),
log-range (pitch range in a segment),
slope (slope of pitch in a segment)

(2) Duration related features

log-D (duration of a segment),
log-dvoice (voiced duration),
log-dpause (pause duration)

(3) Energy related feature

log-E (mean energy of a segment)

This set of parameters is called the prosodic features.

5. SPEAKER MODELS AND SPEAKER RECOGNITION

The acoustic features represented by cepstral coefficients are used to form a feature vector. The Gaussian mixture model (GMM) method is applied on these feature vectors for speaker recognition [10]. A GMM is generated for each speaker. For speaker identification, we find the speaker model which has the maximum posteriori probability for a given test utterance. For speaker verification, a hypothesis test is performed for a given test utterance. The probability of a test utterance belonging to a speaker model is represented by a likelihood score. A threshold must be preset for the hypothesis test.

When the residual features are included in speaker recognition, $\{b_k\}$ are combined with $\{a_k\}$ to form a new acoustic feature vector. The LDA algorithm is used to reduce the feature dimension.

A prosodic parameter is modeled by a single-dimension Gaussian distribution. In the recognition phase, a test model of corresponding parameter is generated from the test utterance. The KL distance is applied to measure the similarity between test model and target model so that a similarity score is calculated. A weighted-and-sum algorithm is used to combine the scores calculated from different prosodic parameters. When the prosodic features are merged with acoustic features in recognition process, a linear combination formula is defined to integrate the scores.

$$score_{total} = w_1 \times score_{acoustic} + w_2 \times score_{prosodic} \quad (4)$$

6. EXPERIMENTS

6.1. Databases

The telephone speech database MAT160 developed in Taiwan and the database NIST99 for one-speaker detection task are used in our experiments. MAT160 is used for speaker identification experiments. It is a database of 81 male and 79 female speakers. 16 utterances of about one second length of each speaker in MATDB_4 subset are the testing data. The remained 50 utterance of each speaker are for training the speaker model.

NIST99 is used for speaker verification experiments. It is a database of 230 male and 309 female target speakers. There are 3420 one-speaker detection segments, and the length of all these segments is longer than 8 seconds and less than one minute.

6.2. Speaker identification

(1) Baseline experiment

The purpose of this experiment is to examine the codec effect to speech signal as well as to the acoustic features. The result is summarized in Table 1.

Table 1. Error rate -- Baseline experiment (%)

Features	Original speech	G.729A coded speech	GSM coded speech
LPCC13	4.61	11.52	8.48
LPCC13+*Pre	6.13	12.30	9.84
MFCC13	1.21	5.23	2.89
MFcc13+*Pre	1.64	5.51	3.20

*Pre – a pre-emphasis filter is applied

It shows that MFCC is much better than LPCC. The other fact is that the pre-emphasis makes the performance even worse.

(2) Including residual features

In this experiment, MFCC converted from codec parameters or derived from decoded speech is used as the speech feature. When the residual features are included, the feature vector is a combination of LP coefficients $\{a_k\}$ and $\{b_k\}$. But the vector dimension is kept in 13. The LDA algorithm is used to reduce the dimension of acoustic features when the residual features are included. The experiment results are shown in Table 2.

Table 2. Error rate – Including residual features (%)

		Original	G729A	GSM
No residual features		1.21	5.23	2.89
MFCC derived from	codec parameters		3.09	2.39
	decoded speech	1.13	4.06	2.38

The results in Table 2 show that the residual features do help the speaker identification.

(3) Including prosodic features

In this experiment, acoustic features (MFCC) and prosodic features are combined for speaker identification. The result is shown in Table 3.

Table 3. Error rate – Including prosodic features (%)

	Original	G.729A	GSM
acoustic features only	1.13	4.06	2.38
Acoustic + prosodic features	0.82	3.48	2.07

It is obvious that the prosodic features give further improvement in the recognition accuracy.

6.3. Speaker verification

(1) Baseline experiment

The MFCC features are used in this experiment. The result is shown in Figure 3. The degradation due to codec process is small. The equal error rates for the original speech and the decoded speech are around 12%.

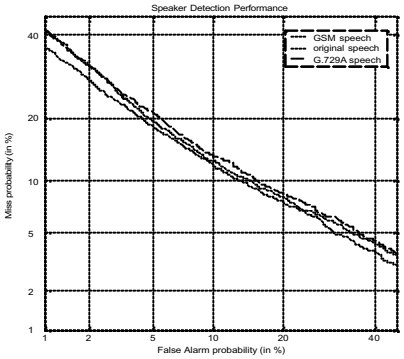


Figure 3. Detection error tradeoff (DET) curve -- Baseline

(2) Including prosodic features

The performance of including prosodic features is shown in Figure 4. The equal error rate has been reduced to around 10%.

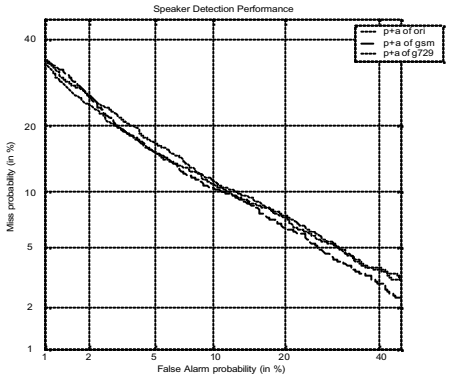


Figure 4. Detection error tradeoff (DET) curve -- Including prosodic features

The experimental result shows that the inclusion of prosodic features does improve the performance, even though the improvement is only about 2% for the NIST database.

7. CONCLUSION

We have investigated the spectral distortion due to the codec process. It motivates us to include prosodic features in the speaker recognition. A new representation of prosodic features based on piecewise linear fitting of pitch contour is introduced. This proposed method which combines acoustic and prosodic features for speaker recognition has been proved to be effective in improving the performance in our experiments.

ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

REFERENCES

[1] M.G. Kuitert & L. Boves, "Speaker verification with GSM coded telephone speech," *Proc. EUROSPEECH 1997*, Rhodes, Vol.2, pp. 975-978, 1997

[2] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, J.P. Campbell, "Speaker and Language Recognition Using Speech Codec Parameters," *Proc. EUROSPEECH 1999*, Vol.2, pp. 787-790, 1999

[3] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini,, "GSM speech coding and speaker recognition," *ICASSP-00,* Vol. 2, pp. 1085-1088, 2000.

[4] A. T. Yu and Hsiao-Chuan Wang, "A Study on the Recognition of Low Bit-Rate Encoded Speech," *Proc. ICSLP 1998*, pp. 38-41, 1998

[5] J. M. Huerta and R. M. Stern, "Speech Recognition from GSM Coder Parameters," *Proc. ICSLP-98*, Vol 4, pp. 1463-1466, 1998

[6] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, J. P. Campbell, and E. Singer, "Speaker Recognition using G.729 speech codec parameters", *Proc. ICASSP '00*, Vol. 2, pp. 1089-1092, 2000

[7] J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification," Proc. of *EUROSPEECH'95,* Vol. 1, pp. 313-316, Sept. 1995, Madrid, Spain

[8] P. Thevenaz and H. Hugli, "Usefulness of the LPC-Residue in text-independent speaker verification," *Speech Communication*, Vol. 17, pp. 145-157. 1995.

[9] K. Sonmez, E. Shriberg, L. Heck & M. Weintraub, "Modeling Dynamic ProsodicVariation for Speaker Verification", *ICSLP-98*, vol. 7, pp. 3189-3192, Sydney

[10] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, 3(1):72 - 83, 1995.