Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 273-290

Integrating Complementary Features from Vocal Source and Vocal Tract for Speaker Identification

Nengheng Zheng*, Tan Lee*, Ning Wang* and P. C. Ching*

Abstract

This paper describes a speaker identification system that uses complementary acoustic features derived from the vocal source excitation and the vocal tract system. Conventional speaker recognition systems typically adopt the cepstral coefficients, *e.g.*, Mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC), as the representative features. The cepstral features aim at characterizing the formant structure of the vocal tract system. This study proposes a new feature set, named the wavelet octave coefficients of residues (WOCOR), to characterize the vocal source excitation signal. WOCOR is derived by wavelet transformation of the linear predictive (LP) residual signal and is capable of capturing the spectro-temporal properties of vocal source excitation. WOCOR and MFCC contain complementary information for speaker recognition since they characterize two physiologically distinct components of speech production. The complementary contributions of MFCC and WOCOR in speaker identification are investigated. A confidence measure based score-level fusion technique is proposed to take full advantage of these two complementary features for speaker identification. Experiments show that an identification system using both MFCC and WOCOR significantly outperforms one using MFCC only. In comparison with the identification error rate of 6.8% obtained with MFCC-based system, an error rate of 4.1% is obtained with the proposed confidence measure based integrating system.

Keywords: Speaker Identification, Vocal Source Feature, Vocal Tract Feature, Information Fusion, Confidence Measure

1. Introduction

Speaker recognition is the process of determining a person's identity based on the intrinsic characteristics of his/her voice. In the source-filter model of human speech production, the

* Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong.
E-mail: nhzheng@ee.cuhk.edu.hk

[Received May 10, 2007; Revised August 9, 2007; Accepted August 12, 2007]

speech signal is modeled as the convolutional output of a vocal source excitation signal and the impulse response of a vocal tract filter system [Rabiner and Schafer 1978]. The most representative vocal tract related acoustic features are the cepstral coefficients, *e.g.*, Mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein 1980] and linear predictive cepstral coefficients (LPCC) [Furui 1981], which aim at modeling the spectral envelope, or the formant structure of the vocal tract. With the primary goal being identifying different speech sounds, these features are believed to provide pertinent cues for phonetic classification and have been successfully applied to automatic speech recognition [Rabiner and Juang 1993]. At the same time, these features are also implemented in most existing speaker recognition systems [Campbell 1997; Reynolds 2002]. This indicates that MFCC and LPCC features do contain important speaker-specific information, in addition to the intended phonetic information. Ideally, if a large amount of phonetically balanced speech data is available for speaker modeling, the phonetic variability tends to be smoothed out so that speaker-specific aspects can be captured.

The vocal source related features, *e.g.*, pitch and harmonics, on the other hand, characterize the vocal folds' vibration style in speech production and are closely related to the speaker-specific laryngeal system. The spoken contents have less effect on the variation of the vocal source excitation than on that of the vocal tract system [Miller 1963; Childers 1991]. This makes the vocal source derived acoustic features useful for speaker recognition, especially for text-independent cases. However, the usefulness of vocal source information for speaker recognition, although having been investigated in some literature, has not been thoroughly studied, let alone the efficient information retrieving techniques. In this paper, a novel vocal source feature is presented and implemented to supplement the vocal tract features in speaker recognition.

For voiced speech, the source excitation signal is a quasi-periodic glottal waveform, which is generated with quasi-periodic vocal fold vibration. The vibration frequency determines the pitch of voice. It has been shown that temporal pitch variation is useful for speaker recognition [Atal 1972; Sonmez 1998]. The amplitude of pitch harmonics has also been demonstrated to be an effective feature for speaker identification [Imperl *et al.* 1997]. To exploit detailed vocal source information, we need a method of automatically estimating the glottal waveform from the speech signal. This can be done by inverse filtering the speech signal with the vocal tract filter parameters estimated during the glottal closing phase (GCI). In Brookes and Chan [1994], a separately recorded laryngograph signal was used to detect the GCI. In Plumpe *et al.* [1999], a method of automatic GCI detection was proposed and the estimated glottal waveform was represented using the Liljencrants-Fant (LF) model. The model parameters were shown to be useful in speaker identification. However, this method worked well only for the typical voices in which the GCI clearly exists and the estimated

glottal waveform can be well explained by the LF model [Plumpe *et al.* 1999].

In linear predictive (LP) modeling of speech signals, the vocal tract system is represented by an all-pole filter. The prediction error, which is named the LP residual signal, contains useful information about the source excitation [Rabiner and Schafer 1978]. In Thevenaz and Hugli [1995], it is shown that the cepstrum of LP residual signal could be used to improve the performance of a text-independent speaker verification system. In He *et al.* [1995] and Chen and Wang [2004], the standard procedures for extracting MFCC and LPCC features were applied to LP residual signals, resulting in a set of residual features for speaker recognition. In Yegnanarayana *et al.* [2005], the speaker information present in LP residual signals was captured using an auto-associative neural network model. Murty and Yegnanarayana [2006] proposed to extract residual phase information by applying Hilbert transform on LP residual signals. The phase features were used to supplement MFCC in speaker recognition.

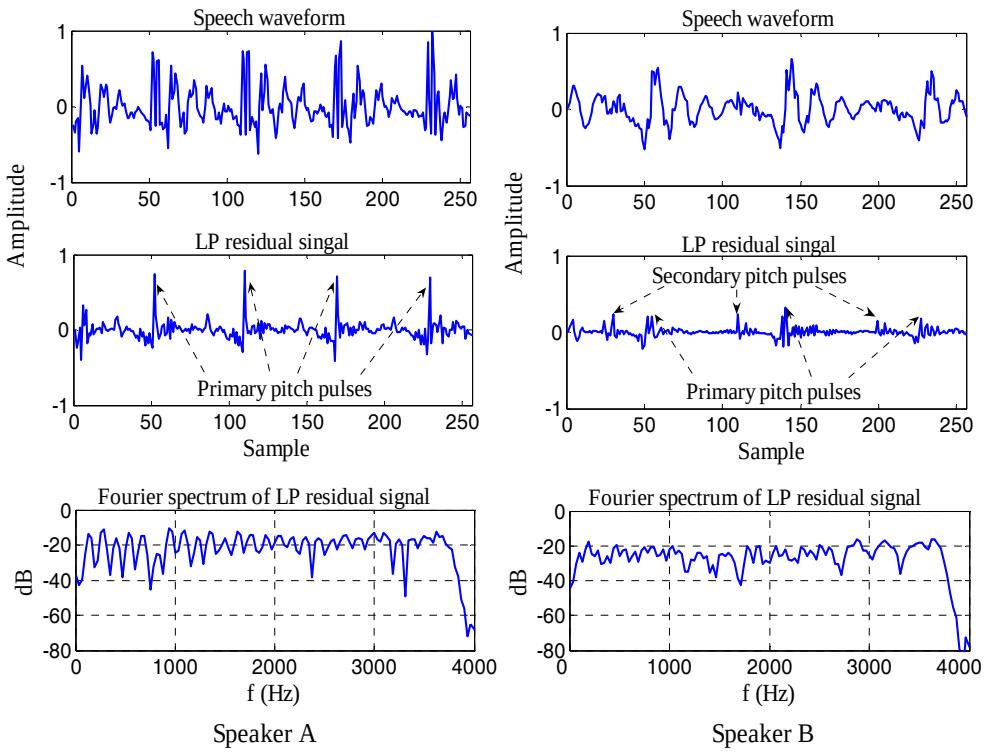


Figure 1. Examples of speech waveforms and LP residual signals of two male speakers. Left: Speaker A; Right: Speaker B; Top to bottom: speech waveforms, LP residual signals and Fourier spectra of LP residual signals

Figure 1 shows the speech waveforms of the vowel /a/ uttered by two different male speakers and the corresponding LP residual signals. There are noticeable differences between the two segments of residual signals. In addition to the difference between their pitch periods, the residual signal of speaker A shows much stronger periodicity than that of speaker B. For speaker B, the magnitudes of the secondary pulses are relatively high. In frequency domain, the Fourier spectra of the two residual signal segments look similar in that they have nearly flat envelopes. Although the harmonic peaks carry speaker-related periodicity information, the useful temporal information, *i.e.*, the amplitudes and the time locations of pitch pulses, are not represented in the Fourier spectra. To characterize the time-frequency characteristics of the pitch pulses, wavelet transform is more appropriate than the short-time Fourier transform.

This paper describes a novel feature extraction technique based on time-frequency analysis of the LP residual signal. The new feature parameters, called wavelet octave coefficients of residues (WOCOR), are generated by applying pitch-synchronous wavelet transform to the residual signal [Zheng *et al.* 2004]. The WOCOR features contain useful information for speaker characterization and recognition. More importantly, WOCOR and MFCC carry different speaker-specific information since they characterize two physiologically distinct components in speech production. As a result, combining these two complementary features will result in higher recognition performance than using only one set of features.

The performance of the information fusion system, however, is highly dependant on the effectiveness of the fusion technique implemented. In multi-modal biometric authentication systems, the reliability of authentication decisions from different classifiers may vary significantly in different tests. Therefore, it is very important to apply an efficient fusion technique to maximize the benefit through the information fusion. A number of information fusion techniques have been proposed for biometrics systems [Garcia-Romero *et al.* 2004; Ross *et al.* 2001; Toh and Tau 2005]. Generally, the information fusion can be done at: (i) feature level, (ii) score level, or (iii) decision level. This paper proposes a score level fusion technique for combining MFCC and WOCOR for speaker identification. Score level fusion is preferred because the matching scores are easily obtained and contain sufficient information for distinguishing different speakers. A confidence measure, which measures the confidence of MFCC in identification decision in comparison to that of WOCOR, is adopted as the fusion weight in each individual identification trial. The confidence measure provides an optimized fusion score by giving more weight to the feature of higher confidence in correct identification. The effectiveness of the proposed information fusion system is demonstrated by a set or speaker identification experiments.

The rest of this paper is organized as follows. Section 2 describes the feature extraction procedures for WOCOR and briefly reviews the MFCC feature extraction procedures. Section

3 demonstrates the usefulness of WOCOR in speaker identification and the complementary contributions of WOCOR and MFCC in speaker identification. Section 4 presents the confidence measure based score-level fusion technique for integrating MFCC and WOCOR for speaker identification. Some analysis of the identification results is presented in Section 5, which further elaborates the complementarity of MFCC and WOCOR in speaker recognition and the superiority of the proposed confidence measure based fusion technique over the fixed-weight fusion. Conclusions are given in Section 6.

2. Vocal Source and Vocal Tract Features

2.1 Vocal Source Features: WOCOR

As illustrated in Figure 1, Fourier spectrum is not good at characterizing the time-frequency properties of the pitch pulses in the residual signal. Wavelet transform has been well known to be an effective method for transient signal representation. Therefore, the proposed WOCOR feature extraction is based on wavelet transform, rather than Fourier transform, of the residual signal. The process of extracting the WOCOR features is formulated in the following steps:

1) *Voicing decision and pitch extraction.* Voicing status decision and pitch extraction are done with Talkin's Robust Algorithm for Pitch Tracking [Talkin 1995]. Only voiced speech is retained for subsequent processing. In the source-filter model, the excitation signal for unvoiced speech can be approximated as random noise [Rabiner and Schafer 1978]. We believe that such noise-like signals carry relatively little speaker-specific information.

2) *LP inverse filtering.* The voiced speech is divided into non-overlapping frames of 30 ms long. The LP residual signal $e(n)$ is obtained from each frame by inverse filtering the speech signal $s(n)$, *i.e.*,

$$e(n) = s(n) - \sum_{k=1}^{12} a_k s(n-k) \tag{1}$$

where the LP filter coefficients a_k are computed using the autocorrelation method [Rabiner and Schafer 1978]. To reduce intra-speaker variation, the amplitude of the residual signal within each voiced segment is normalized to the range [-1, 1].

3) *Pitch-synchronous windowing.* Based on the pitch periods estimated in Step 1, pitch pulses in the residual signal are located by detecting the maximum amplitude within each pitch period. For each pitch pulse, pitch-synchronous wavelet analysis is applied with a Hamming window of two pitch periods long. Let t_{i-1}, t_i and t_{i+1} denote the locations of three successive pitch pulses. The analysis window for the pitch pulse at t_i spans from t_{i-1} to t_{i+1}, as illustrated in Figure 2. The windowed residual signal is denoted as $e_h(n)$.

4) *Wavelet transform of residual signal.* The wavelet transform of $e_h(n)$ is computed as:

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_n e_h(n) \Psi^* \left(\frac{n-b}{a} \right) \tag{2}$$

where $a = \{2^k \,|\, k = 1, 2, \cdots, K\}$ and $b = 1, 2, \cdots, N$, and N is the window length. $\Psi^*(n)$ is the conjugate of the 4th-order Daubechies wavelet basis function $\Psi(n)$. a and b are the scaling parameter and the translation parameter, respectively [Daubechies 1992]. In this case, the LP residual signal is analyzed in K octave sub-bands. For a specific sub-band, the time-varying characteristics within the analysis window are measured as b changes.

5) *Generation of WOCOR feature parameters.* We have K octave groups of wavelet coefficients, *i.e.*,

$$W_k = \left\{ w\left(2^k, b\right) \;\middle|\; b = 1, 2, \cdots, N \right\}, \quad k = 1, 2, \cdots, K \tag{3}$$

To retain the temporal information, each octave group of coefficients is divided evenly into M sub-groups, *i.e.*,

$$W_k^M(m) = \left\{ w\left(2^k, b\right) \;\middle|\; b \in \left(\frac{(m-1)N}{M}, \frac{mN}{M} \right] \right\}, \quad m = 1, 2, \cdots, M \tag{4}$$

where M is the number of sub-groups. The 2-norm of each sub-group of coefficients is computed to be one of the feature parameters. As a result, the complete feature vector is composed of $K \cdot M$ parameters as follows,

$$\text{WOCOR} = \left\{ \left\| W_k^M(m) \right\| \;\middle|\; \begin{matrix} m = 1, 2, \cdots, M \\ k = 1, 2, \cdots, K \end{matrix} \right\} \tag{5}$$

where $\|\bullet\|$ denotes the 2-norm operation.

Figure 2 illustrates the extraction of WOCOR features from a pitch-synchronous segment of residual signal. It can be seen that, with different values of k, the signal is analyzed with different time-frequency resolutions. The time-frequency properties of the signal in each sub-band are characterized by the wavelet coefficients. In this research, we are interested in telephone speech with the frequency band of 300 - 3400 Hz. To cover this range, we set K = 4 and the four frequency sub-bands at different octave levels are defined accordingly: 2000 - 4000 Hz (W_1), 1000 – 2000 Hz (W_2), 500 - 1000 Hz (W_3), and 250 - 500 Hz (W_4). The parameter M determines the temporal resolution attained by the WOCOR parameters. If M = 1, all the coefficients of a sub-band are combined into a single feature parameter, and no temporal information is retained. On the other hand, if a large M is used, such that each coefficient acts as an individual feature parameter, a lot of unnecessary temporal details are included and the feature vector tends to be noisy and less discriminative. A low feature

dimension is also desirable for effective statistical modeling. In Section 3.3, the effect of *M* on recognition performance will be investigated experimentally.

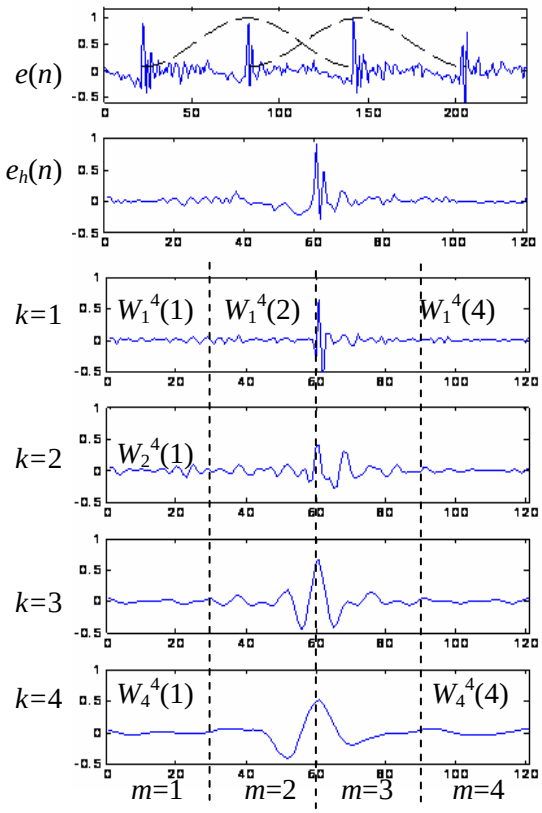


Figure 2. Extraction of WOCOR features from a pitch-synchronous segment of LP residual signal. Here K = 4 and M = 4

To summarize, given a speech utterance, a sequence of WOCOR feature vectors is obtained by pitch-synchronous wavelet transform of the LP residual signal. The WOCOR features are expected to capture spectro-temporal characteristics of the residual signal, which is useful for speaker characterization and recognition.

2.2 Vocal Tract Features: MFCC

The MFCC features have been widely used for speech and speaker recognition. In this study, we use the standard procedures of extracting MFCC on a short-time frame basis as described below [Davis and Mermelstein 1980]:

1) Short-time Fourier transform is applied every 10 ms with 30-ms Hamming window.

2) The magnitude spectrum is warped with a Mel-scale filter bank that consists of 26 filters,

which emulates the frequency resolution of human auditory system.

3) The log-energy of each filter output is computed.

4) Discrete cosine transform (DCT) is applied to the filter-bank output to produce the cepstral coefficients.

The MFCC feature vector has 39 components, including the first 12 cepstral coefficients, the log energy, as well as their first and second order time derivatives.

Aiming at characterizing two physiologically distinct components in speech production, WOCOR and MFCC contain complementary information for speaker discrimination. The effectiveness of WOCOR and its complementarity to MFCC for speaker recognition will be investigated in the following sections.

3. Experiments

3.1 Speaker Identification System

Figure 3 gives the block diagram of the speaker identification system using MFCC and WOCOR. In the pre-processing stage, the speech signal is first pre-emphasized with a first order filter $H(z) = 1 - 0.97z^{-1}$. Then energy-based voice activity detection (VAD) technique is applied to remove the silent portion. The speech signal is passed through for MFCC and WOCOR generation, respectively. For each feature set, speaker models are trained with the UBM-GMM technique [Reynolds *et al.* 2000] in the training stage. A universal background model (UBM) is first trained using the training data from all speakers. Then a Gaussian mixture model (GMM) is adapted from the UBM for each speaker using the respective training data. In the test stage, for each identification trial, likelihoods scores of the two feature sets are first computed and then a score-level fusion is implemented, *i.e.*,

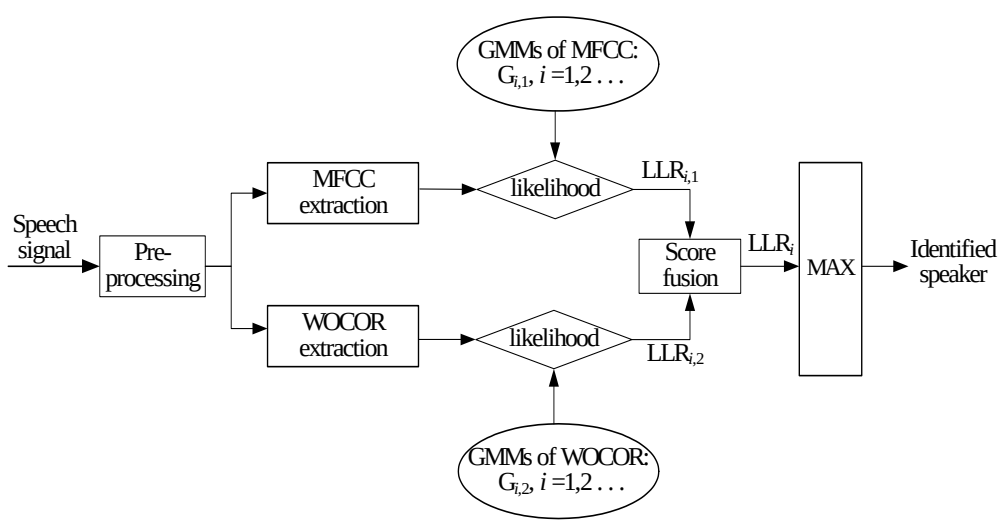


Figure 3. Block diagram of the speaker identification system using MFCC and WOCOR

$$\mathrm{LLR}_i = \mathbf{f}(\mathrm{LLR}_{i,1}, \mathrm{LLR}_{i,2}),\ i = 1, 2, \cdots, N \tag{6}$$

where $\mathrm{LLR}_{i,1}$ and $\mathrm{LLR}_{i,2}$ are likelihood scores obtained from MFCC and WOCOR, respectively, $\mathbf{f}$ is the combination function and N is the number of speakers. Although in real application, the test utterances could come from the unregistered impostors. In this study, we only deal with the closed-set speaker identification. That is, all the test utterances must come from one of the 50 male speakers. The one whose models give the highest matching score is marked as the identified speaker.

3.2 Speech Databases: CU2C

CU2C is a continuous speech database of Cantonese developed at the Chinese University of Hong Kong [Zheng *et al.* 2005]. Cantonese is one of the most popular Chinese dialects and is spoken by tens of millions of people in southern China. CU2C was designed to facilitate general speaker recognition research. It contains parallel utterances collected over fixed-line telephone channel and desktop computer microphones. The spoken contents include Hong Kong personal identity numbers, randomly generated digit strings, and phonetically balanced sentences. In this study, the speaker identification experiments are conducted on the sentence subset of the male speakers. There are 50 male speakers, each having 18 sessions of speech data with 10 utterances in each session. The first 4 sessions are used for training the speaker models. Sessions 5 to 8 are used as development data for training the weighting parameters for the score level fusion of MFCC and WOCOR. The last 10 sessions are used as the evaluation data, and there are totally 5000 identification trials (50 speakers, 100 trials per speaker). All the utterances are text-independent telephone speech with matched training and testing conditions (the same handset and fixed line telephone network). The speech data were sampled at 8 KHz and encoded by 8-bit μ-law encoding. The speech data of each speaker are collected over 4 to 9 months with the minimum inter-session interval of 1 week. Therefore, the challenge of the long-term intra-speaker variation for speaker recognition can be addressed by the database.

3.3 Determining the Parameter *M* for WOCOR

As discussed earlier, the value of M controls the size of the WOCOR feature vector and how much temporal detail can be captured. First, we compare the performance of WOCOR with different values of M. Figure 4 shows the identification error rate (IDER) of WOCOR in which M varies from 1 to 6. The identification error rate is defined as:

$$\mathrm{IDER} = \frac{\text{Number of incorrect identification trials}}{\text{Number of identification trials}} \times 100\% \tag{7}$$

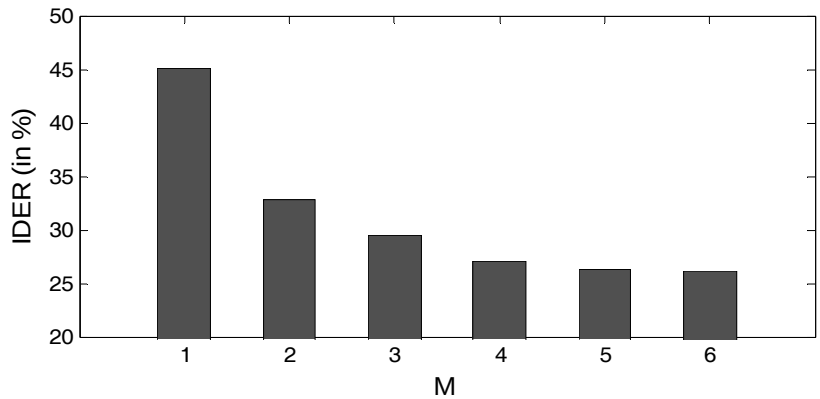


Figure 4. The speaker identification results of WOCOR for different values of M

It is clear that WOCOR in general provide a certain degree of speaker discrimination power. For $M = 1$, *i.e.*, no temporal detail is captured and the feature vector has only 4 components, an IDER of 45.1% is achieved. With *M* increasing from 1 to 4, the IDER is significantly reduced to only 27.0%. For $M > 4$, the improvement becomes less noticeable. Therefore, in the following experiments, we will use WOCOR with $M = 4$, which consists of 16 feature components.

3.4 Wavelet vs. Fourier Transform of LP Residual Signal

To demonstrate the superiority of wavelet transform over Fourier transform for feature extraction from the LP residual signal, we compare the speaker identification performances of WOCOR and the Fourier spectrum-based vocal source features. To do so, we apply the MFCC feature extraction process on the LP residual signal to generate another set of vocal source features, noted as MFCC_{res}. Speaker identification experiments with WOCOR and MFCC_{res} result in IDERs of 27.0% and 52.0%, respectively. That is, WOCOR significantly outperforms MFCC_{res}. This is reasonable because MFCC focuses on extracting the spectral envelope-related features, and, as given in Fig. 1, spectral envelopes of LP residual signals are almost the same for different speakers. On the other hand, WOCOR tries to capture the spectro-temporal information in the residual signals, which is quite different between speakers.

3.5 Speaker Identification Results

We evaluate the speaker identification performances of MFCC and WOCOR individually. In addition, we evaluate the system with both MFCC and WOCOR, using the same evaluation data described in Section 3.2 for all three performance evaluations. In this case, information

fusion is performed as a score-level linear fusion, *i.e.*,

$$\mathrm{LLR}_i = w_t \mathrm{LLR}_{i,1} + (1 - w_t)\mathrm{LLR}_{i,2} \tag{8}$$

The fusion weight w_t is experimentally determined using the development data set. That is, w_t is varied from 0 to 1, and the value giving the smallest IDER is selected for the evaluation trials. Figure 5 shows IDER vs. w_t curve with the development data. As illustrated, the best performance is achieved at around $w_t = 0.80$. Actually, the identification performance is not very sensitive to w_t at around $w_t = 0.80$. The performances of MFCC- and WOCOR-based systems and the information fusion system with $w_t = 0.80$ are evaluated over the evaluation data and the results are as given in Table 1. As shown, the MFCC-based speaker identification system significantly outperforms the WOCOR system. It is noted that, despite the performance difference, the two approaches make complementary decisions in many cases, which will be further elaborated in Section 5, and the combining system has superior performance over that using MFCC only. The IDER is reduced from 6.8% to 4.7%, a relative improvement of about 30%.

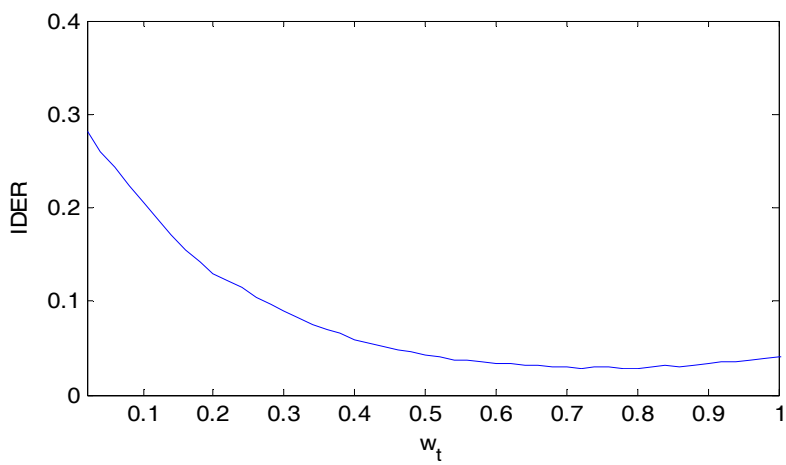


Figure 5. Speaker identification performance with various w_t

Table 1. Speaker identification performances

Systems	IDER (in %)
WOCOR	27.0
MFCC	6.8
MFCC+WOCOR	4.7

4. Information Fusion with Confidence Measure

While information fusion with a pre-defined fusion weight as given in (8) can improve identification performance, it does not necessarily provide the best result. Fixed weight is unable to cover explicitly the different performance levels of MFCC and WOCOR for individual identification trials. As a result, for some cases, although one of the features gives the correct decision, the fused score may not necessarily result in correct decision. For example, consider four types of identification trials as given in Table 2, in which MFCC and WOCOR give different contributions to speaker identification, and the info-fusion as (8) results in different decisions as well. In Type I and II trials, MFCC gives incorrect decisions while WOCOR gives correct decisions. The combined system makes correct decisions in Type I trials while making incorrect decisions in Type II trials. In Type III and IV trials, MFCC gives correct decisions while WOCOR gives incorrect decisions, and the combined system makes correct decisions in Type III trials while producing incorrect decisions in Type IV trials. To avoid the undesired outputs in Type II and IV trials, an ideal solution should be capable of distinguishing these four types of trials and give null weight to MFCC in Type I and II trials and null weight to WOCOR in Type III and IV trials. Although such an ideal solution is not available in real-world applications, we propose to apply a confidence measure based fusion method, which adopts varying weight in individual trials and avoids most of the identification errors introduced by information fusion.

Table 2. Different contributions of MFCC and WOCOR in four types of identification trials

	Type I	Type II	Type III	Type IV
MFCC	incorrect	incorrect	correct	correct
WOCOR	correct	correct	incorrect	incorrect
MFCC+WOCOR	correct	incorrect	correct	incorrect

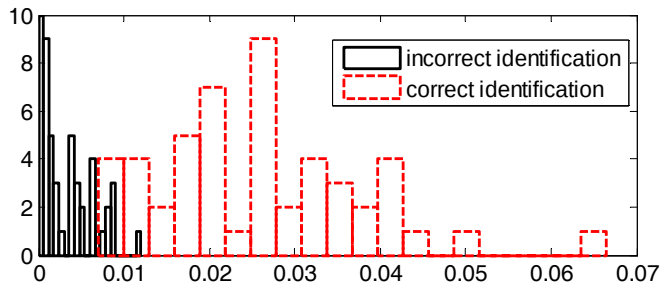
4.1 Speaker Discrimination Power

Analysis of the matching scores shows that, generally, in a correct identification, the difference of the scores between the identified speaker and the closest competitor is relatively larger than that in an incorrect identification. The score difference can therefore be adopted for measuring the speaker discrimination power, *i.e.*,

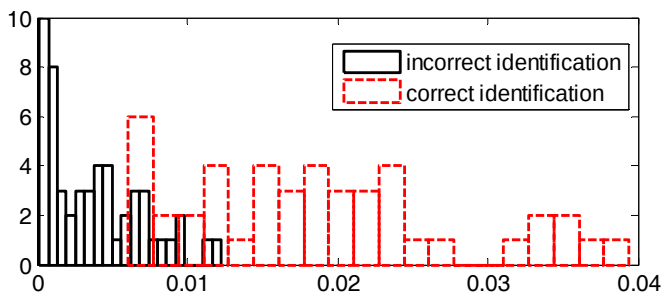
$$D = \frac{\max_i\{\mathrm{LLR}_i\} - \mathrm{sec}\,ond \max_i\{\mathrm{LLR}_i\}}{\left|\max_i\{\mathrm{LLR}_i\}\right|} \tag{9}$$

where LLR_i is the likelihood score of the *i*-th speaker. The normalization of the difference over $\left|\max_i\{\mathrm{LLR}_i\}\right|$ aims to equalize the dynamic ranges of *D* for different features.

Figure 6 shows the histograms of *D* for MFCC and WOCOR. It is clear that, for both features, a correct identification is generally associated with a larger *D* than an incorrect identification. Therefore, a larger *D* implies that the corresponding feature has higher confidence for speaker identification. Obviously, it is desirable to take into account *D* for score fusion in each identification trial instead of using the fixed weight.



(a) MFCC



(b) WOCOR

Figure 6. Histogram of speaker discrimination power D of MFCC and WOCOR

4.2 Confidence Measure Based Score Fusion

Although the optimal method of combining the scores from MFCC and WOCOR with the knowledge of the discrimination power is not known, the relative discrimination power of MFCC and WOCOR can be considered as a confidence measure, with which a better fusion weight can be derived to improve the identification performance. In each identification trial, the confidence measure is defined the discrimination ratio of the two features, *i.e.*,

$$\mathrm{CM} = \left|D_1 / D_2\right| \tag{10}$$

where D_1 and D_2 are the speaker discrimination power of MFCC and WOCOR, respectively. A larger CM implies that the MFCC-based system has a higher confidence in giving correct identification result than the WOCOR-based one. Then, the fusion weight for the specific identification trial is derived as:

$$w_{CM} = -\log \frac{1}{1+e^{-\alpha\cdot(\mathrm{CM}-\beta)}} \tag{11}$$

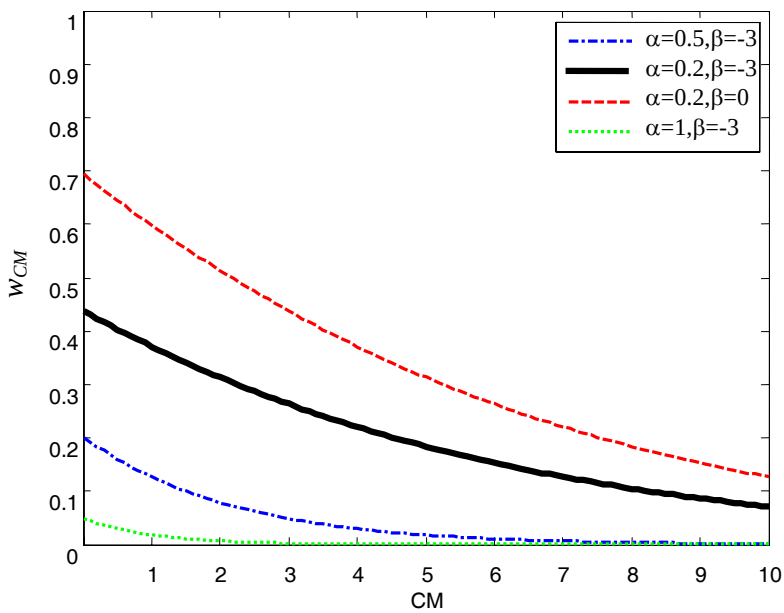


Figure 7. Mapping contours from CM to w_{CM}

where α and β control the slope of the mapping contour from CM to w_{CM}, as illustrated in Figure 7. The solid line curve in Figure 7 is used in this study. The corresponding parameters $\alpha = 0.2, \beta = -3$ are trained using the development data.

Score-level fusion based on CM is then carried out according to:

$$\mathrm{LLR}_i = \mathrm{LLR}_{i,1} + w_{CM}\,\mathrm{LLR}_{i,2} \tag{12}$$

With w_{CM}, the fused score combines better weighted likelihoods obtained from MFCC and WOCOR in each individual trial based on the contributions of the respective features in that trial.

As illustrated in Figure 7, when CM increases, w_{CM} becomes very small, and the decision will not be heavily affected by WOCOR. On the other hand, a small CM corresponds to a large w_{CM}, which means more impact from WOCOR.

As shown in Table 3, the CM-based score level fusion leads to a further performance improvement over the fixed-weight fusion. In summary, the IDERs attained with WOCOR and MFCC, in conjunction with the two methods of score fusion are 27.0%, 6.8%, 4.7%, and 4.1%, respectively.

Table 3. Speaker identification performances

Systems	IDER (in %)
WOCOR	27.0
MFCC	6.8
Fixed-weight fusion	4.7
Fusion with CM	4.1

5. Analysis of the Identification Results

Table 4 elaborates how the integration of the two complementary features affects the identification performances. The identification trials are divided into 4 subsets according to the performances of MFCC and WOCOR: (i) correct identification with both MFCC and WOCOR (McWc), (ii) incorrect identification with both MFCC and WOCOR (MiWi), (iii) incorrect identification with MFCC while correct identification with WOCOR (MiWc), and (iv) correct identification with MFCC while incorrect identification with WOCOR (McWi). Among the 5000 identification trials, there are 3328, 244, 95 and 1333 trials for these 4 subsets, respectively. The number of identification errors with MFCC, WOCOR and the integrated systems within each subset are given in the table.

Table 4. Number of errors of 4 identification subsets by different systems

Subsets	McWc	MiWi	MiWc	McWi
Number of trials	3328	244	95	1333
MFCC	0	244	95	0
WOCOR	0	244	0	1333
Fixed weight fusion	0	163	7	65
Fusion with CM	0	167	19	19

We are only interested in the last 3 subsets, which have errors with at least one kind of features. For the MiWi subset, although both MFCC and WOCOR give incorrect identification results, the combined system gives correct results for some trials. For example, the number of identification errors is reduced from 244 to 163 with the fixed weight fusion and to 167 with the CM-based fusion. That is, about one third of the errors have been corrected.

Table 5 gives an example demonstrating how the score fusion can give correct result even though both MFCC and WOCOR give error results. In this example, the true speaker is S5. It is shown that although S5 only ranks at the 6th and the 2nd with MFCC and WOCOR, respectively, in both integrating systems, it ranks at the first and therefore is correct identified.

The results of the two one-error identification subsets McWi and MiWc in Table 4 demonstrate the superiority of the CM-based score fusion over the fixed-weight fusion. For the fixed-weight fusion system, although the number of errors in the MiWc subset is significantly reduced from 95 to 7, there are 65 errors introduced to the McWi subset, which have been correctly identified with MFCC only. For the CM-based system, the number of this kind of newly introduced errors is significantly reduced to 19, with only a slight increase in errors in MiWi and MiWc subsets. As a whole, the number of total identification errors is reduced from 339 with MFCC only to 235 with fixed-weight fusion, and further reduced to 205 with CM-based fusion.

Table 5. Ranking the speaker scores in an identification trial.

Rank	MFCC	WOCOR	Fixed weight fusion	Fusion with CM
1	S7: -1.7718	S34:1.5732	**S5:-0.4364**	**S5:-1.0903**
2	S27:-1.7718	**S5:1.5730**	S27:-0.4445	S27:-1.0977
3	S10:-1.7722	S48:1.5640	S34:-0.4446	S7: -1.0984
4	S42:-1.7743	S35:1.5620	S41:-0.4448	S10:-1.1000
5	S1: -1.7756	S39:1.5619	S46:-0.4448	S41:-1.1005
6	**S5:-1.7760**	S46:1.5510	S7: -0.4452	S46:-1.1015
7	S41:-1.7788	S41:1.5561	S10:-0.4465	S42:-1.1027

6. Conclusions

This paper presents a novel feature extraction technique to generate the vocal source related acoustic features from the LP residual signal. We have shown that the proposed WOCOR features contain speaker-specific information for speaker recognition applications. The WOCOR features provide additional information to the conventional MFCC features in speaker recognition. This complementarity is exploited by applying a novel confidence measure based score fusion technique which gives a much improved overall speaker identification accuracy. In comparison with the identification error rate of 6.8% obtained with MFCC only, an error rate of 4.1% is obtained with the proposed information fusion system. That is a relative improvement of 40%.

Acknowledgements

This research was supported partially by an Earmarked Research Grant (Ref. CUHK 4236/04E) and a Central Allocation Grant (Ref. CUHK1/02C) from the Hong Kong Research Grants Council.

References

Atal, B. S., "Automatic speaker recognition based on pitch contours," *Journal of the Acoustical Society of America*, 52(6), 1972, pp. 1687-1697.

Brookes, D. M. and D. S. F. Chan, "Speaker characteristics from a glottal airflow model using robust inverse filtering," *Proceedings of Institute of Acoustics*, 16(5), 1994, pp. 501-508.

Campbell, J. P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, 85(9), 1997, pp. 1437-1462.

Chen, S.-H. and H.-C. Wang, "Improvement of speaker recognition by combining residual and prosodic features with acoustic features," In *Proceedings of 29th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 93-96.

Childers, D. G. and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, 90(5), 1991, pp. 2394-2410.

Daubechies, I., *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

Davis, S. B. and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 1980, pp. 357-366.

Furui, S., "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 1981, pp. 254 - 272.

Garcia-Romero, D., J. Fierrez-Aguilar, J. Gonzalez-Rodriguez and J. Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," In *ESCA Workshop on Speaker and Language Recognition, Odyssey*, 2004, pp. 105-110.

He, J., L. Liu and G. Palm, "On the use of features from prediction residual signals in speaker identification," In *Proceedings of Eurospeech*, 1995, pp. 313-316.

Imperl, B., Z. Kacic and B. Horvat, "A study of harmonic features for speaker recognition," *Speech Communication*, 22(4), 1997, pp. 385-402.

Miller, J. E. and M. V. Mathews, "Investigation of the glottal waveshape by automatic inverse filtering," *Journal of the Acoustical Society of America*, 35, 1963 pp.1876.

Murty, K. S. and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, 13(1), 2006, pp. 52-55.

Plumpe, M. D., T. F. Quatieri and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, 7(5), 1999, pp. 569-585.

Rabiner, L. R. and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

Rabiner, L. R. and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

Reynolds, D. A., "An overview of automatic speaker recognition technology," In *Proceedings of 27th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 4072-4075.

Reynolds, D. A., T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10(1-3), 2000, pp. 19-41.

Ross, A., A. Jain and J.-Z. Qian, "Information fusion in biometrics," In *Proceedings of 3rd International Conference on Audio- and Video-Based Person Authentication*, 2001, pp. 354-359.

Sonmez, K., E. Shriberg, L. Heck and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," In *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 3189-3192.

Talkin, D., "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, ed. By W. B. Kleijn and K. K. Paliwal, Elsevier, 1995.

Thevenaz, P. and H. Hugli, "Usefulness of the LPC residue in text-independent speaker verification," *Speech Communication*, 17(1-2), 1995, pp. 145-157.

Toh, K.-A. and W.-Y. Yau, "Fingerprint and speaker verification decisions fusion using a functional link network," *IEEE Transactions on System, Man and Cybernetics B*, 35(3), 2005, pp. 357-370.

Yegnanarayana, B., K. S. Reddy and S. P. Kishore, "Source and system features for speaker recognition using AANN models," In *Proceedings of 26th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 409-413.

Zheng, N., P. C. Ching and T. Lee, "Time frequency analysis of vocal source signal for speaker recognition," In *Proceedings of International Conference on Spoken Language Processing*, 2004, pp. 2333-2336.

Zheng, N., C. Qin, T. Lee and P. C. Ching, "CU2C: A dual-condition Cantonese speech database for speaker recognition application," In *Proceedings of Oriental-COCOSDA*, 2005, pp. 67-72.