Discrimination Power of Vocal Source and Vocal Tract Related Features for Speaker Segmentation

Wai Nang Chan, Nengheng Zheng, *Member, IEEE*, and Tan Lee, *Member, IEEE*

***Abstract*—This paper presents an analysis of the speaker discrimination power of vocal source related features, in comparison to the conventional vocal tract related features. The vocal source features, named wavelet octave coefficients of residues (WOCOR), are extracted by pitch-synchronous wavelet transform of the linear predictive (LP) residual signals. Using a series of controlled experiments, it is shown that WOCOR is less sensitive to spoken content than the conventional MFCC features and thus more discriminative when the amount of training data is limited. These advantages of WOCOR are exploited in the task of speaker segmentation for telephone conversation, in which statistical speaker models need to be built upon short speech segments. Experimental results show that the proposed use of WOCOR leads to noticeable reduction of segmentation errors.**

***Index Terms*—Speaker discrimination power, speaker segmentation, vocal source features, vocal tract features.**

I. Introduction

STATE-OF-THE-ART speaker recognition technology is predominantly based on statistical modeling of short-time features extracted from acoustic speech signals [1]–[3]. The recognition performance is determined by 1) discrimination power of the acoustic features and 2) effectiveness of statistical modeling techniques. In many applications, the amount of training data for speaker modeling are limited and so are the test data for recognition. It is desirable that the acoustic features can maintain good discrimination power regardless of the amount of speech data. This is particularly important for speaker segmentation, which is a special application of speaker recognition. Especially, in speaker segmentation for telephone conversations [4], [5], short speech segments are often encountered, making statistical speaker modeling less reliable. This paper describes a set of newly proposed acoustic features and analyzes their speaker discrimination power under different training and test conditions, which can be exploited in speaker segmentation.

Manuscript received September 14, 2006; revised April 23, 2007. This work was supported in part by the Hong Kong Research Grants Council under Earmarked Research Grant Ref. CUHK 4236/04E and a Central Allocation Grant Ref. CUHK1/02C. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

W.N. Chan was with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China. He is now with the Hang Seng Bank, Hong Kong, China (e-mail: dexterchan@gmail.com).

N. Zheng and T. Lee are with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China (e-mail: nhzheng@ee.cuhk.edu.hk; tanlee@ee.cuhk.edu.hk)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.900103

For automatic speech recognition (ASR), Mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) are the most commonly used acoustic features [6], [7]. With the primary goal of identifying different speech sounds, these features are believed to provide pertinent cues for phonetic classification. They characterize mainly the spectral envelope of a quasi-stationary speech segment. In the source-filter theory of speech production, the spectral envelope corresponds to the vocal tract system, which determines the articulation of sounds [6]. Therefore, the cepstral features extracted from a spoken utterance are closely related to its linguistic content.

Speaker recognition has a different objective, i.e., differentiating one speaker from the others. However, as a matter of fact, most existing speaker recognition systems use vocal tract features like MFCC [1], [3]. This indicates that the MFCC features do contain important speaker-specific information, in addition to the intended phonetic information. Ideally, if a large amount of phonetically balanced speech data are available for speaker modeling, the phonetic variability tends to be smoothed out so that speaker-specific aspects can be captured.

In the source-filter model, the vocal tract system is excited by the vocal source excitation signal, which is acoustically equivalent to the glottal airflow originated from the lungs and modulated at the larynx [8]. Voiced speech is generated with quasi-periodic vocal folds vibration, resulting in a quasi-periodic glottal waveform. The vibration frequency determines the pitch of voice. In [9], it was shown that temporal pitch variation is useful for text-dependent speaker recognition. Imperl *et al.* demonstrated the effectiveness of the amplitudes of pitch harmonics for speaker identification [10]. 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 [11], a separately recorded laryngograph signal was used to detect the GCI. In [12], 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 [12].

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 [8]. Thevenaz and Hugli [13] showed that the cepstrum of LP residual signal could be used to improve the performance of a text-independent speaker

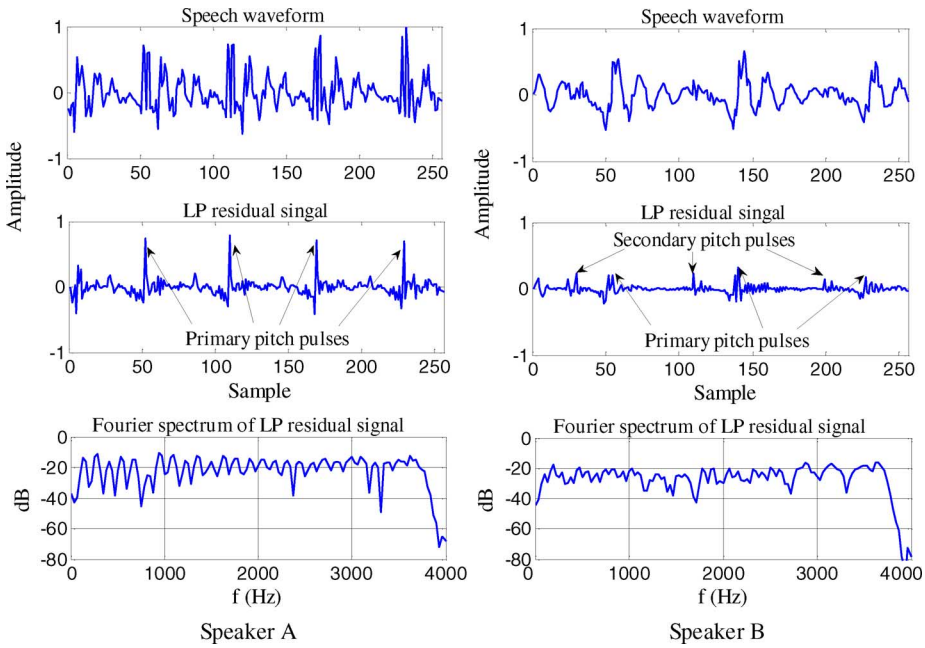


Fig. 1. Examples of speech waveforms and LP residual signals of two male speakers. Left: Speaker A. Right: Speaker B. Top to bottom: speech waveforms, LP residual signals, and Fourier spectra of LP residual signals.

verification system. In [14] and [15], the standard procedures of extracting MFCC and LPCC features for ASR were applied to LP residual signals, resulting in a set of residual features for speaker recognition. In [16], the speaker information present in LP residual signals was captured using an autoassociative neural network model. Murty and Yegnanarayana [17] 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.

Fig. 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. In our previous work [18], a novel feature extraction technique based on time–frequency analysis of LP residual signals was described. The resulted feature parameters, named wavelet octave coefficients of residues (WOCOR), were obtained by applying pitch-synchronous wavelet transform to the residual signals. It was shown that the WOCOR features provide complementary information to MFCC and help improving speaker recognition performance [18], [19].

This paper presents an analysis of the speaker discrimination power of the vocal source-related WOCOR features, in comparison with the vocal tract-related MFCC features. Section II describes the feature extraction procedures for WOCOR and briefly reviews the MFCC feature extraction procedures. Section III describes the speech database to be used in the subsequent sections. Section IV compares the discrimination power of WOCOR and MFCC using a series of controlled experiments. Specifically, we investigate on the effect of linguistic content matching or mismatching between training and test data, and the effect of the amount of training data. Section V demonstrates the usefulness of WOCOR in speaker segmentation, in which the issues of content mismatching and data sparseness are rather critical. Section VI gives conclusions of this study.

II. Vocal Source and Vocal Tract Features

A. Vocal Source Features: WOCOR

As illustrated in Fig. 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 a good way 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 [18].

1) *Preprocessing.* An energy-based voice activity detection (VAD) procedure is applied to the input utterance to remove the silence and long pauses that contain no speech. Subsequently, the speech signal is preemphasized with a first-order filter, i.e., $E(z) = 1 - 0.97z^{-1}$.
2) *Voicing decision and pitch extraction.* Voicing status decision and pitch extraction are done with Talkin's Robust Algorithm for Pitch Tracking [20]. 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 [8]. We believe that such

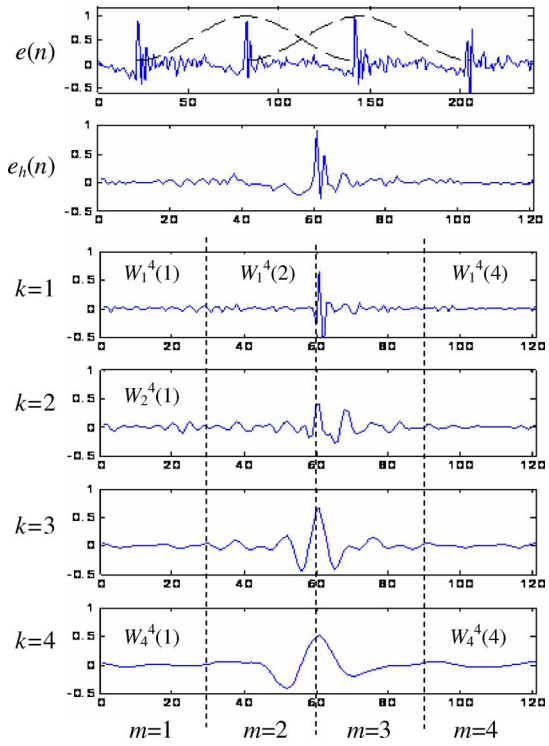


Fig. 2. Extraction of WOCOR features from a pitch-synchronous segment of residual signal. Here, $K = 4$ and $M = 4$.

noise-like signals carry relatively little speaker-specific information.

3) *LP inverse filtering.* The voiced speech is divided into nonoverlapping frames of 30-ms length. 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 [8]. To reduce intraspeaker variation, the amplitude of the residual signal within each voiced segment is normalized to the range $[-1, 1]$.

4) *Pitch-synchronous windowing.* Based on the pitch periods estimated in step 2), 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 Fig. 2. The windowed residual signal is denoted as $e_h(n)$.

5) *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 $a = \{2^k | k = 1, 2, \ldots, K\}$ and $b = 1, 2, \ldots, N$, and N is the window length. $\Psi^*(n)$ is the conjugate of the fourth-order Daubechies wavelet basis function $\Psi(n)$. a, and b are the scaling parameter and the translation parameter, respectively [21]. In this case, the LP residual signal is analyzed in K octave subbands. For a specific subband, the time-varying characteristics within the analysis window are measured as b changes.

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

$$W_k = \{ w(2^k, b) \, | b = 1, 2, \ldots, N \}, k = 1, \ldots, K. \tag{3}$$

To retain the temporal information, each octave group of coefficients is divided evenly into M subgroups, i.e.,

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

where M is the number of subgroups. The 2-norm of each subgroup 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:

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

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

Fig. 2 illustrates the extraction of WOCOR features from a pitch-synchronous segment of residual signal. It can been 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 subband are characterized by the wavelet coefficients. In this research, we are interested in telephone speech with the frequency band of 300 to 3400 Hz. To cover this range, we set $K = 4$ and four frequency subbands 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 subband 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. Our previous work showed that speaker recognition performance would not be significantly improved as M increases beyond 4 [18]. Thus, we fix $M = 4$ throughout this paper and the feature vector has 16 components.

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.

B. Vocal Tract Features: MFCC

The MFCC features have been widely used for speech recognition and speaker recognition. In this study, we use the standard procedures of extracting MFCC on a short-time frame basis as described below [7].

TABLE I
PHONETIC TRANSCRIPTIONS (IN IPA) OF THE TEN CANTONESE DIGITS

Digit	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
Transcription	*lɪŋ*	*jɐt*	*ji*	*sam*	*sei*	*ŋ*	*lʊk*	*tsʰɐt*	*pat*	*gɐu*

1) Short-time Fourier transform is applied every 10 ms with 20-ms Hamming window.
2) The magnitude spectrum is warped with a set of nonlinearly spaced triangular filters that are centered on equally spaced frequencies in Mel-scale.
3) The log-energy of each filter output is computed.
4) Discrete cosine transform (DCT) is applied to the filterbank output to produce the cepstral coefficients.

Aiming at characterizing two physiologically distinct components in speech production, WOCOR and MFCC contain complementary information for speaker discrimination. It has been shown that although WOCOR is not as effective as MFCC in speaker recognition, the integration of these two features gives a noticeable performance improvements over the MFCC-only system [18], [19].

III. SPEECH DATABASES

Two speech databases are used for the analysis of discrimination power in Section IV. They are CU2C, a continuous speech database of Cantonese, and the NIST 2004 Speaker Recognition Evaluation (SRE) database, which contains English speech. The NIST database is also used for the speaker segmentation experiment in Section V.

A. CU2C

CU2C is a large-scale Cantonese speech database developed at the Chinese University of Hong Kong [22]. Cantonese is one of the most popular Chinese dialects 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. There are 50 male and 34 female speakers in CU2C. Each of them has 18 sessions of recordings, which were collected over four to nine months with the minimum intersession interval of one week.

Our experiment is done on the male telephone speech in CU2C. The speech data were sampled at 8 KHz and encoded by 8-bit μ-law encoding. Only the digit-string utterances are used. For each speaker, we use the 72 utterances recorded in the last 12 sessions (six utterances per session), which were designated as the test sessions in the design of CU2C [22]. Each utterance contains 14 randomly generated Cantonese digits. The speakers were asked to read the digit sequence fluently. Table I gives the phonetic transcriptions of the ten Cantonese digits "0" to "9." Each digit is pronounced as a monosyllable that consists of no more than three phonetic units.

B. NIST 2004 SRE

NIST 2004 SRE database is for the task of speaker detection in summed-channel telephony conversations [23]. In this study, we use only part of the database. The data contains 36 conversations randomly selected from the whole set of 1200 conversations. The duration of each conversation is approximately 5 min. Each conversation involves two unknown speakers talking over telephones. There are ten male–male, ten female–female, and 16 male–female conversations. The speech was sampled at 8 KHz and encoded by 8-bit μ-law. Each of the selected conversations was manually divided into speaker-homogeneous segments. As a result, a total of 1857 speaker segments were marked in the 36 conversations. Excluding the silence and nonspeech periods, the segment duration is between 1 to 3 s in most cases.

IV. ANALYSIS OF DISCRIMINATION POWER

In speaker segmentation, especially for conversations, the hypothesized segments are usually very short and the spoken contents are not restricted. To investigate the effectiveness of MFCC and WOCOR in these kinds of applications, we analyze and compare their speaker discrimination power on the effect of 1) linguistic content mismatching and 2) the amount of training data. In these experiments, the MFCC feature vector is formed by the first 12 cepstral coefficients and the WOCOR feature vector has 16 components. MFCC- and WOCOR-based speaker models are established in parallel using the same training data, and their error rates under different training and test conditions are compared.

A. Effect of Content Mismatch

1) Method: Our analysis starts with the two-speaker verification problem, i.e., only two speakers need to be recognized. Without loss of generality, one of the speakers is designated as the target speaker, denoted by λ, and the other as the alternative speaker, denoted by μ. The experiment consists of a large number of tests. In each test, a content-specific speech segment from the target speaker is used. By "content-specific," we mean that the utterance contains a specific word (or a set of specific words).

Suppose that there exist L distinct spoken contents, which are denoted by $w_1, w_2, \ldots, w_L$, respectively. Let $\mathbf{X}_{\lambda,i}$ be a speech segment with content w_i spoken by λ. $\mathbf{X}_{\lambda,i}$ is tested against a target speaker model $\Theta_{\lambda,l}$ and an alternative speaker model $\Theta_{\mu,k}$, which are trained using speech segments with contents w_l and w_k, respectively. The likelihoods $P(\mathbf{X}_{\lambda,i} \mid \Theta_{\lambda,l})$ and $P(\mathbf{X}_{\lambda,i} \mid \Theta_{\mu,k})$ are computed. If $P(\mathbf{X}_{\lambda,i} \mid \Theta_{\lambda,l}) > P(\mathbf{X}_{\lambda,i} \mid \Theta_{\mu,k})$, $\mathbf{X}_{\lambda,i}$ is considered to be correctly recognized as being spoken by λ; otherwise, a recognition error is recorded. With a large number of test segments, a statistical error rate can be obtained. There are four possible cases of content matching and mismatching between test data and training data, which are listed and explained as in Table II.

2) Test Procedures: As stated in Section III-A, our analysis is carried out with the male telephone speech in CU2C. Each speaker has 72 utterances recorded in 12 different sessions. Each of the ten digits is treated as a distinct content. Thus, we have $L = 10$ different contents. All utterances are divided into digit segments by the forced-alignment technique using a set of pretrained hidden Markov models. Therefore, for each speaker, there are $72 \times 14 = 1008$ digit segments, i.e., about 100 segments for each digit.

Under each of the test conditions as described in Table II, a large number of speaker verification tests are performed.

TABLE II
FOUR DIFFERENT CONDITIONS OF CONTENT MATCHING AND MISMATCHING BETWEEN TEST DATA AND SPEAKER MODELS. THE TEST SEGMENT IS ASSUMED TO HAVE CONTENT w_i. THE TARGET MODEL AND THE ALTERNATIVE MODEL ARE TRAINED ON CONTENTS w_l AND w_k, RESPECTIVELY

(T, A)	$l = i$ $k = i$	Test content matched with both target and alternative speaker models
$(\overline{T}, \overline{A})$	$l \neq i$ $k \neq i$	Test content mismatched with both target and alternative speaker models
$(T, \overline{A})$	$l = i$ $k \neq i$	Test content matched with target model and mismatched with alternative model
$(\overline{T}, A)$	$l \neq i$ $k = i$	Test content mismatched with target model and matched with alternative model

Each test involves one digit segment from a designated target speaker. Both the target speaker model and the alternative speaker model are represented by a single Gaussian distribution. They are trained to be content-specific, i.e., each model corresponds to a specific Cantonese digit. The training data are N randomly selected digit segments that satisfy the respective test condition. For example, under the condition (T, A), if the test segment carries the digit "0," N segments of "0" from the target speaker are used to train the target model and N segments of "0" from the alternative speaker are used to train the alternative model. The test segment itself must not be included in the training data. Under the condition $(T, \overline{A})$, the alternative model is trained on one of the nine mismatched contents, resulting in nine independent tests. Similarly, the number of tests required for the conditions $(\overline{T}, \overline{A})$ and $(\overline{T}, A)$ are 81 and 9, respectively.

With the 50 speakers in the database, there are $50 \times 49 = 2450$ possible combinations of target speaker and alternative speaker. For each of the combinations, the same test procedures are applied to all digit segments of the target speaker.

3) Results and Discussion: Fig. 3 shows the test results for MFCC- and WOCOR-based speaker models with N varying from 1 to 64. When the test segment has the same content as both the target and the alternative models, i.e., the (T, A) condition, MFCC is more discriminative than WOCOR [Fig. 3(a)]. MFCC-based models perform well even with very little training data, while the performance of WOCOR-based models depend greatly on the amount of training data. With sufficient training data, the discrimination power of WOCOR becomes comparable to MFCC. Moving from (T, A) to $(T, \overline{A})$, both MFCC- and WOCOR-based models show performance improvement [Fig. 3(b)]. This is because the alternative model likelihood is lowered due to content mismatching while the target model likelihood remains the same.

When the target model is trained on a different content from the test segment, both MFCC and WOCOR become less effective. Fig. 3(c) and (d) shows that MFCC suffers more from content mismatching. Under the $(\overline{T}, \overline{A})$ condition, the best error rates attained by MFCC and WOCOR are 29.8% and 17.7%, respectively, with $N = 64$. $(\overline{T}, A)$ is considered the most unfavorable condition for speaker recognition. In this case, MFCC-based models perform badly regardless of the amount of training data, and WOCOR-based models keep improving as the amount of training data increases. With $N = 64$, the error rates for MFCC and WOCOR are 70.6% and 33.3%, respectively.

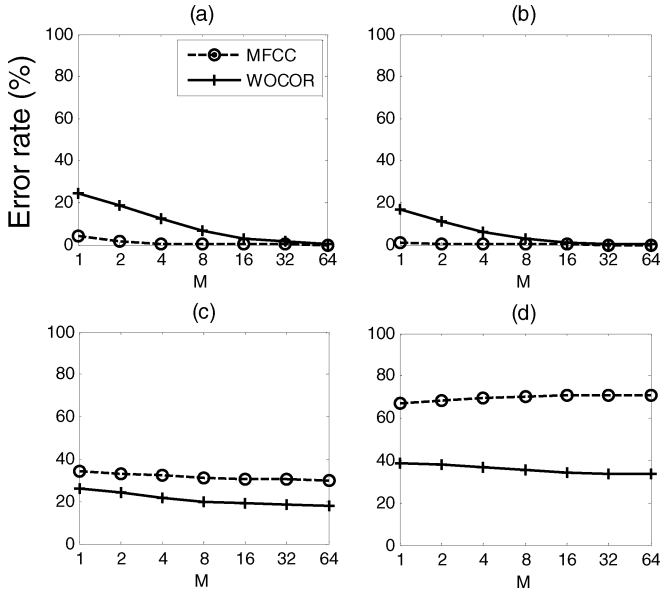


Fig. 3. Two-speaker test results for MFCC- and WOCOR-based speaker models under different test conditions. (a) (T, A). (b) $(T, \overline{A})$. (c) $(\overline{T}, \overline{A})$. (d) $(\overline{T}, A)$.

From the above analysis, WOCOR is found to be more discriminative for speaker recognition than MFCC when there is a content mismatch between the test utterance and the training data. In other words, MFCC is more sensitive to spoken content than WOCOR.

B. Effect of the Amount of Training Data

In this section, the discrimination power of MFCC and WOCOR are compared under text-independent training conditions with varying amount of training data. Similar to Section IV-A, we perform two-speaker verification experiments based on statistical speaker models.

The analysis is conducted on the 36 conversations in the NIST 2004 SRE database, as described in Section III-B. Given a conversation, one of the speakers is designated as the target speaker λ and the other as the alternative speaker μ. There are many segments from each speaker in the conversation. Each segment is represented by a sequence of feature vectors (MFCC or WOCOR). The target model and the alternative model are trained with a subset of N feature vectors from one selected segment of the respective speaker. In case that the segment contains less than N feature vectors, it would not be used for the test. By varying N, we can control the amount of training data.

Let $S_{\lambda,j}$ denote the jth segment of λ, and $F_{\lambda,j}$ be the collection of N feature vectors randomly selected from $S_{\lambda,j}$. The target speaker model trained with $F_{\lambda,j}$ is denoted by Θ_λ. The alternative speaker model, denoted by Θ_μ, is trained with $F_{\mu,k}$, which consists of N randomly selected feature vectors from the kth segment of speaker μ. Let $S_{\lambda,p}$ denote a test segment from the target speaker, where $p \neq j$, and $F_{\lambda,p}$, which is composed of N randomly selected feature vectors from $S_{\lambda,p}$, are used for the computation of the likelihoods $P(F_{\lambda,p} \mid \Theta_\lambda)$ and $P(F_{\lambda,p} \mid \Theta_\mu)$. If $P(F_{\lambda,p} \mid \Theta_\lambda) > P(F_{\lambda,p} \mid \Theta_\mu)$, $S_{\lambda,p}$ is considered to be correctly recognized as being spoken by λ; otherwise, a recognition error is recorded. The same test procedures

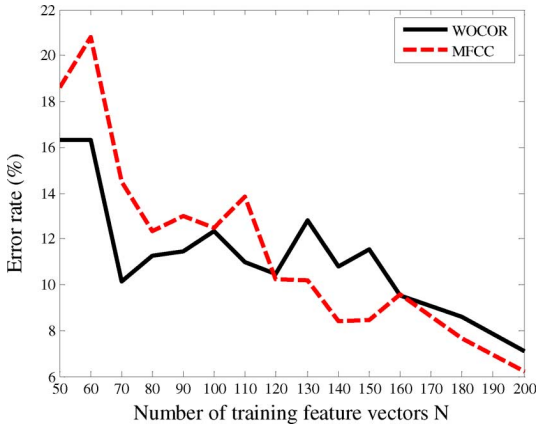


Fig. 4. Test results for MFCC- and WOCOR-based speaker models (all 36 conversations).

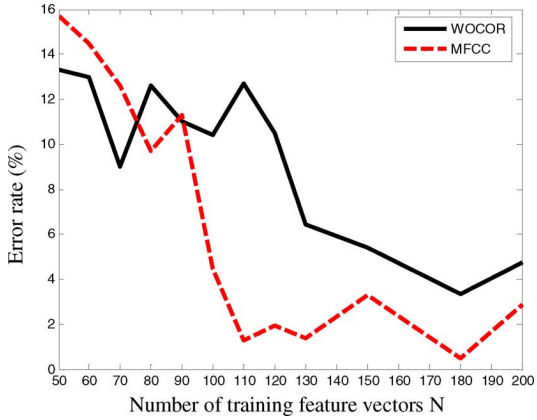


Fig. 5. Test results for MFCC- and WOCOR-based speaker models (ten male–male conversations).

are repeated for all segments in the conversation and for all of the 36 conversations. There are totally 56 000 tests from which a statistical error rate is obtained.

Fig. 4 compares the percentage error rates attained by MFCC- and WOCOR-based speaker models. N is varied from 50 to 200, which are equivalent to speech duration of 0.5 to 2 s. The figure shows that WOCOR performs better than MFCC when relatively few feature vectors are used to represent a speaker. MFCC-based models keep improving as training data increases and catch up with WOCOR at $N = 120$. With $N < 120$, both the test and the training speech are shorter than 1.2 s, which can accommodate only a small number of phonemes. It is highly likely that the test speech and the training speech have different contents. As shown in Section IV-A, WOCOR is less susceptible to content mismatching and thus can provide better discrimination than MFCC in this case. When more speech data are used, the phonetic content becomes rich and balanced so that MFCC-based models can provide an effective representation of speaker characteristics and perform well in speaker recognition. In Section IV-A, there was also a content mismatching problem between test data and training data. However, increasing the amount of training data in this case does not enrich the phonetic coverage since all training segments were restricted to contain a specific Cantonese digit. Thus, the performance of MFCC was found to be always worse than WOCOR regardless of how much training data was used.

Figs. 5–7 give the test results on male–male, female–female, and male–female conversations, respectively. For male–female cases, WOCOR works better than MFCC up to $N = 200$. For those conversations where the speakers are of the same gender, the advantage of WOCOR is not obvious unless N is very small. This confirms that WOCOR is more speaker discriminative than MFCC in this task of text-independent speaker verification, when the amount of training and test data is limited.

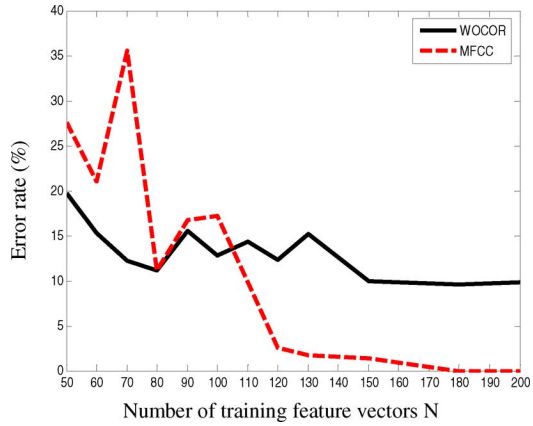


Fig. 6. Test results for MFCC- and WOCOR-based speaker models (ten female–female conversations).

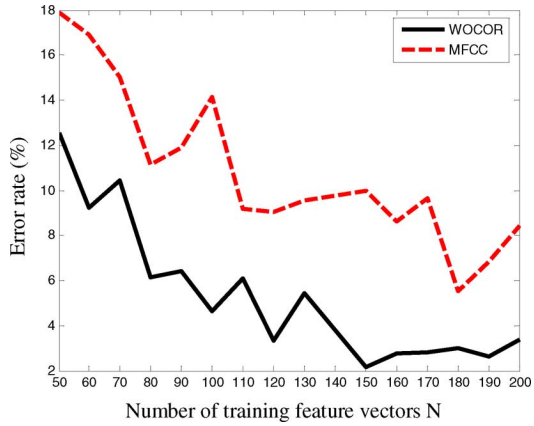


Fig. 7. Test results for MFCC- and WOCOR-based speaker models (16 male–female conversations).

V. Use of Vocal Source Features in Speaker Segmentation

WOCOR and MFCC are regarded as the representatives of vocal source and vocal tract related features, respectively. From the analysis in Section IV, we have the following observations.

- WOCOR is less sensitive to spoken content than MFCC.
- In text-independent applications, WOCOR is more discriminative than MFCC when the amount of training data is small.

In the following, we describe how these properties of WOCOR can be exploited in speaker segmentation.

A. Problem

Speaker segmentation, also known as speaker diarization, is a task of dividing an input speech signal into homogeneous segments, each of which contains the speech of exactly one speaker [24]. The applications include audio indexing and searching, automatic labeling, and transcription of audio archives that involve multiple speakers. In this study, we focus on the problem of speaker segmentation in telephone conversation between two unknown speakers. The same 36 conversations as described in Section III-B are used.

There are two basic problems to be addressed in speaker diarization. First, speaker turning points, i.e., the time instants when there are changes of speakers, need to be determined. Second, the speech segments separated by the turning points are associated with particular speakers. Speaker segmentation algorithms have been mostly based on statistical modeling of the conventional MFCC features [25]. Usually there is no or very little prior knowledge about the speakers such that no speaker model can be established beforehand. The speaker models need to be built from preliminarily hypothesized segments in the speech signal being processed. For telephone conversations, these hypothesized segments are very short and the spoken contents are not restricted. Such an application scenario calls for acoustic features like WOCOR, which are relatively less content-sensitive.

B. Basic Algorithm

Speaker turning point detection and segment clustering can be done sequentially in one pass [26], [27] or iteratively in multiple passes [28]–[30]. Turning points are hypothesized based on local change of acoustic properties. The hypothesized turning points divide the speech signal into many segments. These segments are clustered into a certain number of speaker-homogenous groups. Statistical modeling techniques are used for both turning point detection and segment clustering. Our basic algorithm of speaker segmentation is based on the previous work reported in [5], [30], [25]. It is a noniterative process that consists of three steps.

1) Preliminary Segmentation: Preliminary segmentation is to find a set of hypothesized speaker turning points. This is done with the DISTBIC technique proposed by Delacourt [5], which involves sequential use of spectral distance measurement and the Bayesian information criterion (BIC). When applying spectral distance measurement, we consider a 2-s window of speech in each measurement. The window is divided into two equal parts and each of them is represented by a multivariate Gaussian distribution. In [5], it was shown that the generalized likelihood ratios (GLR) and Kullback–Leibler distance (KLD) perform better than other distance measures. In our experiment, we use the KLD. Let f and g denote the two multivariate Gaussian distributions, respectively. The KLD is computed as

$$\mathrm{KLD}(f,g) = \frac{1}{2} \cdot \mathrm{tr}\left\{(\Sigma_f^{-1}+\Sigma_g^{-1})(\mu_f-\mu_g)(\mu_f-\mu_g)^T+\Sigma_f\Sigma_g^{-1}+\Sigma_g\Sigma_f^{-1}-2\mathbf{I}\right\} \quad (6)$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, Σ and μ are, respectively, the covariance matrix and the mean vector of the Gaussian distributions, and $\mathbf{I}$ is the identity matrix.

The window slides over the conversation at the step size of 100 ms. As a result, we obtain a time-varying KLD curve. The potential speaker turning points are detected at the peaks of the KLD curve. Subsequently, the turning points are refined using the ΔBIC value [5]. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a sequence of feature vectors from N successive speech frames. In this study, N is set to 200, and the frame shift is 10 ms. Consider two subsequences of $\mathbf{X}$: $\mathbf{X}_l = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i\}$ and $\mathbf{X}_r = \{\mathbf{x}_{i+1}, \mathbf{x}_{i+2}, \ldots, \mathbf{x}_N\}$, where $1 < i < N$. The following two hypotheses are defined [26].

- H_0 —- $\mathbf{X}$ is generated by a single Gaussian distribution denoted by $\mathcal{N}(\mu, \Sigma)$.
- H_1 —- $\mathbf{X}_l$ and $\mathbf{X}_r$ are generated by two distinct Gaussian distributions $\mathcal{N}(\mu_l, \Sigma_l)$ and $\mathcal{N}(\mu_r, \Sigma_r)$, respectively.

ΔBIC value is given by the likelihood ratio of H_0 and H_1, minus a penalty [26], i.e.,

$$\Delta\mathrm{BIC}(i) = N\log|\Sigma| - i\log|\Sigma_l| - (N-i)\log|\Sigma_r| - \lambda P \quad (7)$$

where P is a penalty term used to compensate the model complexity difference, and λ is a parameter that controls the threshold for turning point detection. P is defined as [5]

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N \quad (8)$$

where d is the feature dimension.

It is assumed that there is no turning point in $\mathbf{X}$ if $\Delta\mathrm{BIC}(i) < 0$ for all i. If there exists a turning point, its location is given by

$$i^\star = \arg\max_i \Delta\mathrm{BIC}(i) \quad (9)$$

2) Segments Clustering: This is a bottom-up process in which similar segments are merged iteratively. The turning points detected above divide the conversation into many segments. Each of these segments is treated as an initial cluster. At each iteration of clusters merging, the two closest clusters are combined to form a new cluster. The ΔBIC defined as in (7) is used to measure the "closeness" between a pair of clusters. The entire clustering process ends when there are only two clusters left, which are assumed to represent the two speakers in the conversation. The speech data in each cluster are used to train a GMM speaker model with 32 mixture components.

3) Resegmentation: The two GMM speaker models are used to resegment the conversation using the Viterbi algorithm. It determines the most probable path that toggles between the two speaker states. There is a constraint of minimum segment length of 0.5 s. The resulted state sequence indicate speaker occurrences in the conversation.

C. Use of WOCOR

Given that WOCOR is more discriminative than MFCC for short speech segments, we propose the following way of incorporating WOCOR into the basic segmentation algorithm.

- In preliminary segmentation, WOCOR is used to replace MFCC in the computation of both KLD in (6) and ΔBIC in (7) for turning point detection. Since the size of analysis window is 2 s, assuming the turning point is in the middle of the window, the speech data available for estimating the Gaussian distribution is only 1-s long. In this case, WOCOR is expected to be more discriminative than MFCC.

- In segments clustering, WOCOR is made contributive to the ΔBIC score at the initial stage of segments clustering when the clusters contain very little data. At each iteration, the average amount of speech data over all existing clusters is observed. If it is less than 2 s of speech, ΔBIC is computed as

$$\Delta\text{BIC} = W_{\text{MFCC}} \cdot \Delta\text{BIC}_{\text{MFCC}} + W_{\text{WOCOR}} \cdot \Delta\text{BIC}_{\text{WOCOR}} \tag{10}$$

  where $\Delta\text{BIC}_{\text{MFCC}}$ and $\Delta\text{BIC}_{\text{WOCOR}}$ are the ΔBIC scores produced by the MFCC- and WOCOR-based speaker models, respectively. If the average amount of data is more than 2 s of speech, the clustering decision depends only on MFCC.
  In our experiments, the optimal weighting factors $W_{\text{MFCC}} = 0.3$ and $W_{\text{WOCOR}} = 0.7$ are determined empirically based on two conversations that are randomly selected from the 1200 conversations in NIST 2004 database (see Section III-B). These two conversations are different from the 36 test conversations.
- In resegmentation, WOCOR is not used at all. The resegmentation algorithm remains the same as the basic algorithm described in Section V-B. The duration of each conversation is about 5 min. Assuming that each speaker talks half of the time, the speech data used for building the speaker models are about 2.5 min long, for which MFCC is considered to be more appropriate than WOCOR.

D. Experimental Results

The performance of speaker segmentation is assessed in terms of 1) the accuracy of speaker turning point detection and 2) the speaker coverage. There are two types of errors in turning point detection, namely false alarm (FA) and missed detection (MD). The false alarm rate (FAR) and the missed detection rate (MDR) are defined as

$$\text{FAR} = \frac{\text{Number of false alarmed turning points}}{\text{Number of detected turning points}} \tag{11}$$

$$\text{MDR} = \frac{\text{Number of missed turning points}}{\text{Number of reference turning points}} \tag{12}$$

where the number of reference turning points is given by the manual reference segmentation.

For the measurement of speaker coverage, the false alarm coverage (FACov) and the missed detection coverage (MDCov) are defined as [31]

$$\text{FACov} = \frac{\sum_{\rho} \text{duration of false alarmed portion in segment } \rho}{\sum_{\rho} \text{duration of detected segment } \rho} \tag{13}$$

$$\text{MDCov} = \frac{\sum_{\theta} \text{duration of missed portion in segment } \theta}{\sum_{\theta} \text{duration of reference segment } \theta}. \tag{14}$$

The experimental results are illustrated as in Tables III and IV. The baseline system has an overall error rate (the average of FAR and MDR) of 28.8% in turning point detection. With the proposed use of WOCOR, the error rate is reduced to 24.0%.

TABLE III
RESULTS OF SPEAKER SEGMENTATION EXPERIMENTS: SPEAKER TURNING POINTS DETECTION

Segmentation Errors	Without use of WOCOR	With use of WOCOR
FAR	31.6%	27.3%
MDR	25.9%	20.7%
(FAR+MDR)/2	28.8%	24.0%

TABLE IV
RESULTS OF SPEAKER SEGMENTATION EXPERIMENTS: SPEAKER COVERAGE

Segmentation Errors	Without use of WOCOR	With use of WOCOR
FACov	12.3%	9.94%
MDCov	7.40%	5.30%
(FACov+MDCov)/2	9.85%	7.62%

In terms of the speaker coverage, the error rate improves from 9.85% to 7.62%.

VI. CONCLUSION

The MFCC features have been routinely utilized in most speaker recognition applications. Statistical models trained by MFCC describe the speakers' voice characteristics and, at the same time, model the variation of speech contents. Our study reveals that MFCC would become less discriminative in speaker recognition when there is a content mismatch between the test speech and the training speech. In this study, we use WOCOR as a representative of vocal source related features and demonstrate its effectiveness in speaker discrimination. WOCOR is found to be less sensitive to spoken content than MFCC, and thus for text-independent speech data, WOCOR is more discriminative than MFCC when the amount of training data is small. WOCOR is particularly useful for speaker segmentation, in which speaker models need to be built from limited amount of text-independent data. We have shown that WOCOR can be effectively utilized in some intermediate steps of speaker segmentation, to achieve better performance than using MFCC only.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.
[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 19–41, 2000.
[3] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 4072–4075.
[4] A. E. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, "Unsupervised speaker segmentation of telephone conversations," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 565–568.
[5] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Commun.*, vol. 32, pp. 111–126, 2000.
[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

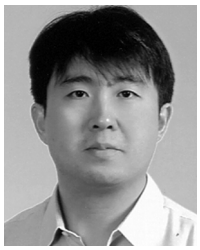
[9] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687–1697, 1972.
[10] B. Imperl, Z. Kacic, and B. Horvat, "A study of harmonic features for speaker recognition," *Speech Commun.*, vol. 22, no. 4, pp. 385–402, 1997.
[11] D. M. Brookes and D. S. F. Chan, "Speaker characteristics from a glottal airflow model using robust inverse filtering," *Proc. Inst. Acoust.*, vol. 16, no. 5, pp. 501–508, 1994.
[12] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–585, Sep. 1999.
[13] P. Thevenaz and H. Hugli, "Usefulness of the LPC residue in text-independent speaker verification," *Speech Commun.*, vol. 17, no. 1-2, pp. 145–157, 1995.
[14] J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification," in *Proc. Eurospeech*, 1995, pp. 313–316.
[15] S.-H. Chen and H.-C. Wang, "Improvement of speaker recognition by combining residual and prosodic features with acoustic features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 93–96.
[16] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 409–413.
[17] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.
[18] N. Zheng, P. C. Ching, and T. Lee, "Time frequency analysis of vocal source signal for speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2004, pp. 2333–2336.
[19] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 181–184, Mar. 2007.
[20] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995.
[21] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, PA: SIAM, 1992.
[22] N. Zheng, C. Qin, T. Lee, and P. C. Ching, "CU2C: A dual-condition Cantonese speech database for speaker recognition applications," in *Proc. Oriental-COCOSDA*, 2005, pp. 67–72.
[23] "The NIST year 2004 speaker recognition evaluation plan," [Online]. Available: http://www.nist.gov/speech/tests/spk/2004/index.htm
[24] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 5, pp. 953–956.
[25] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast news and telephone conversations," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, 2004 [Online]. Available: http://www.ll.mit.edu/IST/pubs/0501_torres.pdf.
[26] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, pp. 127–132.
[27] S. E. Tranter and D. A. Reynolds, "Speaker diarization for broadcast news," in *Proc. Odyssey Speaker Lang. Recognition Workshop*, 2004, pp. 337–344.
[28] J.-L. Gauvain, L. Lamel, and G. Adda, "Audio partitioning and transcription for broadcast data indexation," *Multimedia Tools Applicat.*, vol. 14, pp. 187–200, 2001.
[29] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE Automatic Speech Recognition Understanding*, 2003, pp. 411–416.
[30] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Improving speaker diarization," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04F)*, 2004 [Online]. Available: http://www.limsi.fr/Individu/barras/publis/rt04f_diarization.pdf.
[31] W. N. Chan, T. Lee, N. Zheng, and H. Ouyang, "Use of vocal source features in speaker segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 657–660.

Wai Nang Chan received the B.Eng. and M.Phil. degrees from the Chinese University of Hong Kong, Hong Kong, China, in 2004 and 2006, respectively.

He is currently with Hang Seng Bank, Hong Kong. His research interest is on speaker verification and segmentation.

Nengheng Zheng (M'06) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 1997 and 2002, respectively, and the Ph.D. degree from the Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2006, all in electrical and electronic engineering.

From 1997 to 1999, he was a Member of Technical Staff with the Fujian Start Computing Group Company, Ltd, Fuzhou, China. He was a Visiting Student at the Center for Information Science, Peking University, Beijing, China, in 2004. Currently, he is a Research Associate in the Department of Electronic Engineering, CUHK. His research interests include automatic speech and speaker recognition, nonlinear modeling of speech, and speech enhancement.



Tan Lee (M'03) received the B.Sc. and M.Phil. degrees in electronics in 1988 and 1990, respectively, and the Ph.D. degree in electronic engineering in 1996, all from the Chinese University of Hong Kong (CUHK), Hong Kong, China.

Since 1999, he has been a faculty member at the Department of Electronic Engineering, CUHK. Currently, he is in charge of the Digital Signal Processing and Speech Technology Laboratory, which is composed of over 20 research staff and postgraduate students. He was a Guest Researcher at the Department of Speech, Music, and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden, from 1997 to 1998. His research covers many different areas of speech processing, including automatic speech and speaker recognition, text-to-speech, tone modeling for Chinese, forensic speech processing, and speech enhancement for hearing prostheses. He initiated and coordinated a number of pioneering projects on the research and development of Chinese spoken language technologies in Hong Kong.

He is a member of the International Speech Communication Association (ISCA). He was the Chairman of IEEE Hong Kong Chapter of Signal Processing from 2005 to 2006. He is an Associate Editor of the *EURASIP Journal of Applied Signal Processing*.