Integration of Complementary Acoustic Features for Speaker Recognition

Nengheng Zheng, *Student Member, IEEE*, Tan Lee, *Member, IEEE*, and P. C. Ching, *Senior Member, IEEE*

***Abstract*—This letter describes a speaker verification system that uses complementary acoustic features derived from the vocal source excitation and the vocal tract system. A new feature set, named the wavelet octave coefficients of residues (WOCOR), is proposed to capture the spectro-temporal source excitation characteristics embedded in the linear predictive residual signal. WOCOR is used to supplement the conventional vocal tract-related features, in this case, the Mel-frequency cepstral coefficients (MFCC), for speaker verification. A novel confidence measure-based score fusion technique is applied to integrate WOCOR and MFCC. Speaker verification experiments are carried out on the NIST 2001 database. The equal error rate (EER) attained with the proposed method is 7.67%, in comparison to 9.30% of the conventional MFCC-based system.**

***Index Terms*—Confidence measure, information fusion, LP residual signal, speaker verification, wavelet transform.**

I. Introduction

SPEAKER recognition is a 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 speech signal is modeled as the convolutional output of a vocal source excitation signal and the impulse response of a vocal tract filter system [1]. The most representative vocal tract-related acoustic features are the linear predictive cepstal coefficients (LPCC) and the Mel-frequency cepstral coefficients (MFCC), which aim at modeling the spectral envelope, or the formant structure of the vocal tract [2], [3]. MFCC and LPCC have been successfully applied to speaker recognition [4]. The usefulness of vocal source-related features, on the other hand, has also been investigated, though to a lesser extent. Such features include mainly pitch [5], harmonic structure [6], and phase information [7]. Brookes and Chan [8] and Plumpe *et al.* [9] attempted to estimate and model the glottal flow derivative waveform and used these parameters to identify individual speaker. These studies have demonstrated that the vocal source-related features provide complementary information to MFCC and LPCC.

Given an acoustic speech signal, it is very difficult, if not impossible, to recover the exact excitation signal. Based on the theory of linear prediction, the LP residual signal generated by inverse filtering contains useful information about the source excitation [1], [7]. Fig. 1 shows the speech waveforms of the vowel /a/ uttered by two male speakers and their corresponding LP residual signals. There are noticeable differences between the residual signals of the two speakers. In addition to the significant 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 higher. For both speakers, the short-time Fourier transforms of their residual signals give nearly flat spectra. Although the harmonic structures of the spectra reflect the periodicity, the pitch pulse-related time-frequency properties cannot be easily extracted from the Fourier spectra. To characterize the time-frequency characteristics of the pitch pulses, wavelet transform is more appropriate than the short-time Fourier transform.

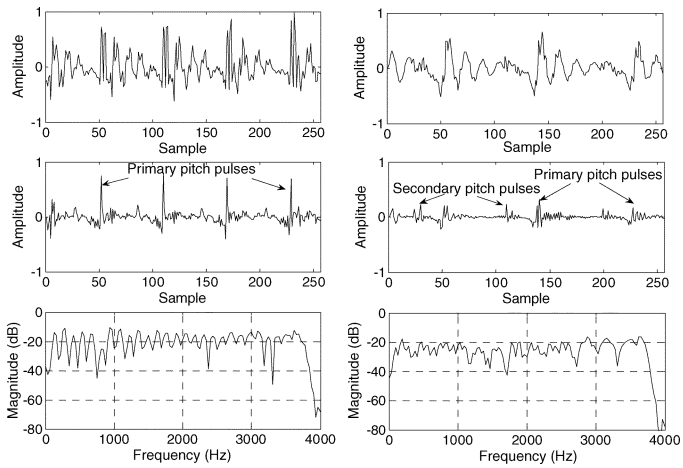

Fig. 1. Examples of speech waveforms and LP residual signals of two male speakers. (Left) Speaker A. (Right) Speaker B. (Top to bottom) Speech waveforms, LP residual signals, and Fourier spectra of LP residual signals.

This letter 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 a pitch-synchronous wavelet transform to the residual signal [10]. Experimental results show that the WOCOR parameters provide complementary information to the conventional MFCC features for speaker recognition. We propose to use a confidence measure to combine the likelihood scores obtained based on the WOCOR and the MFCC features. The confidence measure is derived to reflect the different discrimination power of WOCOR and MFCC in each recognition trial.

II. Feature Extraction From LP Residual Signal

The process of extracting the proposed WOCOR features is formulated in the following steps.

Manuscript received June 5, 2006; revised August 10, 2006. This work was supported in part by a research grant awarded by the Research Grant Council of the Hong Kong SAR Government. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alfred Hanssen.

The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (e-mail: nhzheng@ee.cuhk.edu.hk; tanlee@ee.cuhk.edu.hk; pcching@ee.cuhk.edu.hk).

Digital Object Identifier 10.1109/LSP.2006.884031

1) *Voicing decision and pitch extraction*. Voicing decision and pitch extraction are done by the robust algorithm for pitch tracking [11]. Only voiced speech is kept for subsequent processing. In the source-filter model, the excitation signal for unvoiced speech is approximated as a random noise [1]. We believe that such a noise-like signal carries little speaker-specific information in the time-frequency domain.
2) *LP inverse filtering*. For each voiced speech portion, a sequence of LP residual signals of 30 ms long is obtained by inverse filtering the speech signal, i.e.,

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

where the filter coefficients a_k are computed on Hamming windowed speech frames using the autocorrelation method [1]. The $e(n)$'s of neighboring frames are concatenated to get the residual signal, and their amplitude is normalized within [−1,1] to reduce intra-speaker variation.
3) *Pitch-synchronous windowing*. With the pitch periods estimated in step 1, pitch pulses in the residual signal are located. 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}. The windowed residual signal is denoted as $e_h(n)$.
4) *Wavelet transform of the 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 fourth-order Daubechies wavelet basis function $\Psi(n)$. a and b are the scaling parameter and the translation parameter, respectively [12]. In this letter, the speaker verification experiments are carried out on telephone speech, which has the frequency band of 300–3400 Hz. $K = 4$ is selected such that the signal is decomposed into four sub-bands at different octave levels: 2000–4000 Hz $(k = 1)$, 1000–2000 Hz, $\cdots$, 250–500 Hz $(k = 4)$. At a specific sub-band, the time-varying characteristics within the analysis window are measured as b changes.
5) *Generating the feature parameters*. We now have four octave groups of wavelet coefficients, i.e.,

$$W_k = \left\{ w(2^k, b) | b = 1, 2, \cdots, N \right\}, \quad k = 1, \cdots, 4. \tag{3}$$

Each octave group of coefficients is divided evenly into M sub-groups, i.e.,

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

The two-norm of each sub-group of coefficients is computed to be a feature parameter. As a result, the complete feature vector is composed as

$$\mathrm{WOCOR}_M = \left\{ \left\| W_k^M(m) \right\| \,\middle|\, \begin{matrix} m = 1, \cdots, M \\ k = 1, \cdots, 4 \end{matrix} \right\} \tag{5}$$

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

In the case of $M = 1$, all the coefficients of a sub-band are combined to form a single feature parameter, and therefore, all temporal information is lost. On the other hand, if $M = N$, each coefficient is included as an individual component in the feature vector. This may introduce too much unnecessary detail so that the features become less discriminative. From a statistical modeling point of view, a relatively low feature dimension is also desirable. In Section III-B, the effect of M on recognition performance will be investigated experimentally.

To summarize, given a speech utterance, a sequence of WOCOR_M feature vectors is obtained by pitch-synchronous analysis of the LP residual signal. Each feature vector consists of $4M$ components, which are expected to capture useful spectro-temporal characteristics of the residual signal.

III. Speaker Verification Experiments

A. Experimental Setup

Speaker recognition problems are categorized into speaker verification and speaker identification. In this letter, we focus on the task of speaker verification, which aims to validate the identity claimed by a person.

We adopt the state-of-the-art GMM-UBM approach of statistical speaker modeling and speaker verification [13]. The universal background model (UBM) is a Gaussian mixture model (GMM). It is built on a large amount of training data and serves as the impostor model in verification. For each target speaker, a speaker model is adapted from the UBM using the respective training data. Given a test utterance, the log-likelihood ratio (LLR) is obtained by

$$\mathrm{LLR} = \log P(s|\lambda_c) - \log P(s|\lambda_U) \tag{6}$$

where $P(s|\lambda_c)$ and $P(s|\lambda_U)$ are the likelihoods given by the claimed speaker model and the UBM, respectively. If LLR is higher than a preset threshold θ, the claimant is accepted. Otherwise, it is rejected. There are two types of errors: false acceptance (FA) of an impostor and false rejection (FR) of the genuine speaker. With different values of θ, different FA and FR rates can be attained. This leads to a detection error trade-off (DET) curve [14], on which each point corresponds to a specific value of θ, with the horizontal and vertical coordinates being the FA rate and the FR rate, respectively. The performance of a speaker verification system can also be measured in terms of the equal error rate (EER), which is attained by choosing θ such that FA rate equals FR rate.

Speaker verification experiments are conducted on the NIST 2001 one-speaker detection database [15]. Only verification results from male speakers are reported in that paper. The database consists of a development set and an evaluation set. The development set is used to determine the parameter M for WOCOR_M (see Section III-B) and to train the fusion weights for WOCOR_M and MFCC (see Section III-C and IV). The evaluation set includes 74 male speakers. Each speaker has a training utterance of about 2 min long, which is used to adapt the respective speaker model from the UBM. There are totally 850 test utterances with duration of 15∼ 45 s. Each test utterance is evaluated against 11 hypothesized speakers, including the genuine speaker and ten attacked speakers. All utterances are cellular telephone speeches, and the training and test speeches are from different numbers.

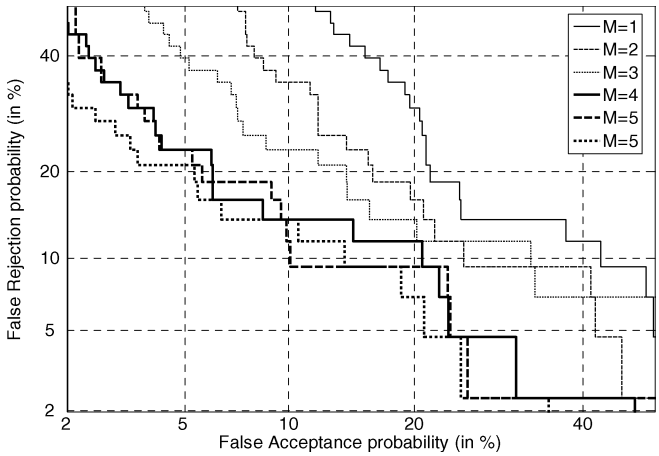

Fig. 2. Performance of WOCOR_M-based speaker recognition for different values of M.

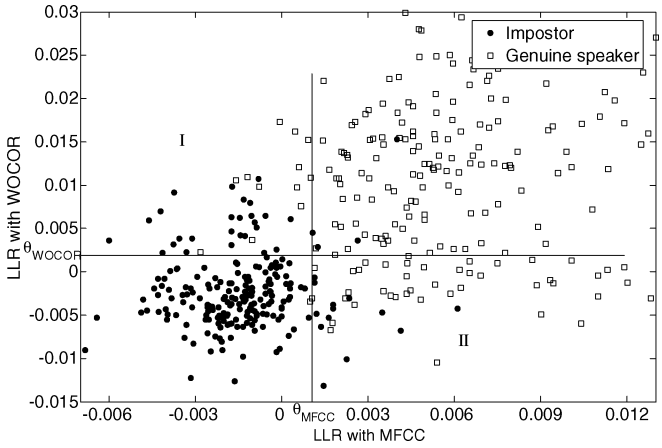

Fig. 3. Distribution of 2-D LLR scores for genuine and impostor trials. θ_{WOCOR} and θ_{MFCC} are chosen such that FR=FA.

The performance of WOCOR_M is evaluated in comparison with the conventional MFCC parameters. To each input utterance, an energy-based voice activity detection (VAD) is applied to remove the silence portions and the speech signal is pre-emphasized. The generation of WOCOR is described as in Section II. The extraction of MFCC follows the standard procedures as described in [3].

1) Short-time Fourier transform is applied every 10 ms with a 30-ms Hamming window.
2) The magnitude spectrum is warped with a Mel-scale filter bank that consists of 26 filters. Log-magnitude of each filter output is calculated.
3) Discrete cosine transform (DCT) is applied to the filter bank output.

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. Since the speech data used in our experiments were recorded via telephone networks, the method of cepstral mean normalization (CMN) [16] is applied to eliminate the convolutional channel distortion.

B. Determining the Parameter M

As discussed earlier, the value of M controls the size of the WOCOR_M feature vector and how much temporal detail can be captured. We use the development data set to compare the performance of WOCOR_M with different values of M. The test results are given by the DET curves in Fig. 2. It is clear that WOCOR_M 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 four components, an EER of 21.5% is achieved. The DET curves show that, with M increasing from 1 to 4, the verification performance is significantly improved. For $M > 4$, the improvement becomes less noticeable. Therefore, in the following experiments, we will use WOCOR_4, which consists of 16 feature components.

C. WOCOR as Complementary Feature to MFCC

We first evaluate the performances of WOCOR_4 and MFCC individually using the evaluation set. As shown in Fig. 4, the MFCC-based speaker verification system significantly outperforms the WOCOR one. It is noted that, despite the performance difference, the two approaches make complementary decisions in many cases. Fig. 3 shows the distribution of the LLR scores attained by WOCOR and MFCC for the genuine and impostor trials. Let θ_{MFCC} and θ_{WOCOR} denote the LLR thresholds of the MFCC and WOCOR-based systems. They partition the score distribution into four regions. We are interested in the cases that the two systems give different decisions, which are labeled Region I and II, respectively. For Region I, the MFCC system proposes rejection, and the WOCOR system suggests acceptance. It can be seen that the MFCC system falsely rejects some genuine utterances, which can actually be accepted by the WOCOR system. Some of the impostor utterances that are falsely accepted by the WOCOR system can be rejected by the MFCC system. Similar observation can be made on Region II.

To take advantages of the complementarity, a simple method of score-level linear fusion is formulated as follows:

$$\mathrm{LLR} = w_t \mathrm{LLR}_1 + (1 - w_t)\mathrm{LLR}_2 \tag{7}$$

where LLR_1 and LLR_2 are calculated from the MFCC and the WOCOR-based systems, respectively. 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 EER is selected for the evaluation trials. As a result, we have $w_t = 0.84$. The evaluation results are given in Fig. 4. The system that combines the contributions of MFCC and WOCOR has superior performance over that of using MFCC only. The EER is reduced from 9.30% to 8.32%.

IV. Information Fusion With Confidence Measure

While information fusion with predefined weighting as given in (7) can improve verification performance, it does not necessarily provide the best result. Fixed weighting is unable to cover explicitly the different performance levels of WOCOR and MFCC for individual verification trials. It is found in some cases that, although one of the features gives the correct decision with a higher confidence, the fused score, however, results in a wrong decision. To avoid this undesirable consequence, we propose to apply a confidence measure (CM) for the score fusion.

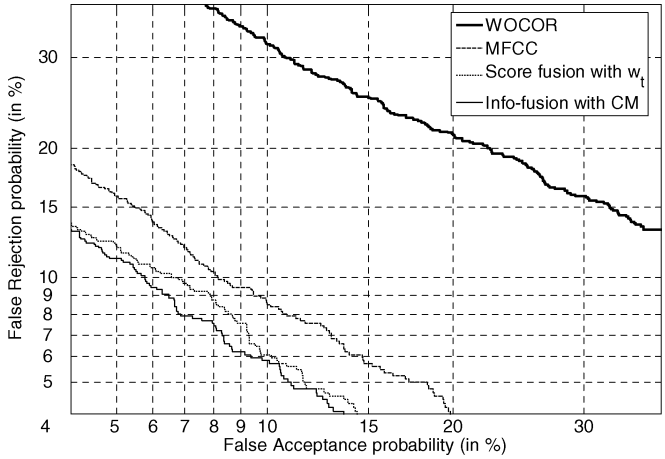

Fig. 4. Performances of speaker verification systems with WOCOR, MFCC, and the two methods of score fusion.

We first define the speaker discrimination power of WOCOR and MFCC in a specific verification trial as

$$D_i = \frac{\log P(s_i|\lambda_{\mathrm{c},i}) - \log P(s_i|\lambda_{\mathrm{U},i})}{|\log P(s_i|\lambda_{\mathrm{U},i})|}, \quad i = 1, 2 \qquad (8)$$

where $\log P(s_i|\lambda_{\mathrm{c},i})$ and $\log P(s_i|\lambda_{\mathrm{U},i})$ denote the log-likelihoods given by the client model and the background model, respectively. The subscript i denotes the different features: $i = 1$ for MFCC and $i = 2$ for WOCOR. In this way, $\mathrm{LLR}_i = \log P(s_i|\lambda_{\mathrm{c},i}) - \log P(s_i|\lambda_{\mathrm{U},i})$ is exploited to measure the discrimination attained by the respective features. If $D_i \gg 0$ or $D_i \ll 0$, the features are considered to have high confidence in the decision of accepting a genuine speaker or rejecting an impostor. If D_i is close to 0, the confidence is low and the decision tends to be uncertain. The normalization by $|\log P(s_i|\lambda_{\mathrm{U},i})|$ aims to equalize the dynamic ranges of D_i for different features.

In each trial, we compute the discrimination ratio of MFCC and WOCOR as

$$\mathrm{DR} = |D_1/D_2|. \qquad (9)$$

A larger DR implies that the MFCC-based system has a higher confidence than the WOCOR-based one. Accordingly, a CM based on DR is defined

$$\mathrm{CM} = -\log \frac{1}{1 + e^{(-\alpha \cdot (\mathrm{DR} - \beta))}} \qquad (10)$$

where $\alpha = 0.75$ and $\beta = 2$ are determined from the development data.

Score-level fusion based on the confidence measure is then carried out according to

$$\mathrm{LLR} = \mathrm{LLR}_1 + \mathrm{LLR}_2 \cdot \mathrm{CM}. \qquad (11)$$

With CM, the fused score combines better weighted LLR_1 and LLR_2. From (10), when DR increases, CM becomes very small, and the decision will not be heavily affected by WOCOR. On the other hand, a small DR corresponds to a large CM, which means more impact from WOCOR.

As shown in Fig. 4, the score fusion based on CM leads to a further performance improvement over the fixed-weight fusion. In summary, the EERs attained with WOCOR and MFCC, in conjunction with the two methods of score fusion, are 21.8%, 9.30%, 8.32%, and 7.67%, respectively.

V. Conclusion

We have shown that the proposed WOCOR features, which are vocal source related, contain speaker-specific information for speaker recognition applications. The WOCOR features provide additional information to the conventional MFCC features in speaker verification. This complementarity is exploited by applying a novel confidence measure-based score fusion technique that gives a much improved overall verification accuracy. In comparison with the EER of 9.30% obtained with the MFCC-based system, an EER of 7.67% is obtained by the combined use of MFCC and WOCOR.

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[3] 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.

[4] J. P. Campbell, "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[5] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody based speaker recognition," in *Proc. Eurospeech*, 1997, pp. 1391–1394.

[6] 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.

[7] 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.

[8] D. M. Brooke 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.

[9] 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.

[10] N. H. Zheng, P. C. Ching, and T. Lee, "Time frequency analysis of vocal source signal for speaker recognition," in *Proc. Int. Conf. Spoken Language Processing*, 2004, pp. 2333–2336.

[11] 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.

[12] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[15] The NIST 2001 Speaker ID Evaluation Protocol. [Online]. Available: http://www.nist.gov/speech/tests/spk/200I/index.htm.

[16] B. S. Atal, "Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.