Abstract

This paper investigates the importance of spectro-temporal characteristics of the source excitation signal for speaker recognition. We propose an effective feature extraction technique for obtaining essential time-frequency information from the linear prediction (LP) residual signal, which are closely related to the glottal excitation of individual speaker. With pitch synchronous analysis, wavelet transform is applied to every two pitch cycles of the LP residual signal to generate a new feature vector, called Wavelet Octave Coefficients of Residues (WOCOR), which provides additional speaker discriminative power to the commonly used linear predictive Cepstral coefficients (LPCC). Experimental evaluation over a Cantonese speaker recognition corpus demonstrates the effectiveness of WOCOR for speaker recognition. Recognition tests with WOCOR and LPCC outperform the conventional methods of using Mel Frequency Cepstral Coefficients (MFCC).

1. Introduction

Exploiting speaker specific information embedded within the speech signal for speaker characterization and recognition is one of the important techniques for the human-computer dialog system. According to the speech production model, human speech is the convolution output of the source excitation signal $u(n)$ and the impulse response of the vocal tract system $h(n)$ [1],

$$s(n) = u(n) * h(n)$$

State-of-the-art speaker recognition systems typically employ acoustic features mainly carrying vocal tract information, such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC). The importance and applicability of such vocal tract characteristics have been extensively acknowledged [2]. However, the usefulness of the vocal source excitation, as well as its effective retrieving technique, has not been thoroughly studied.

Many of the studies on vocal source information for speaker recognition have been focused on pitch and cepstral coefficients of the excitation signal (i.e., the LP residual signal) [4]. More recently, explicit temporal modeling of the glottal flow derivative waveform has also been studied, and the importance of glottal flow waveshape of the voiced sound for speaker recognition was demonstrated [6]. However, none of these studies revealed the time-frequency properties of the source excitation signal, which may be more appropriate to characterize the relatively fast time-varying excitation signal and could be helpful for speaker recognition.

In our previous work, we applied Haar transform to the LP residual signal with fixed duration of 32 ms. By dividing the Haar spectrum into different octave groups, the spectral details of the corresponding bandwidth can be achieved. Furthermore, dividing each octave group into subgroups, the temporal information within the analysis time span could also be obtained. The feature vector, HOCOR$, is believed to retain the spectro-temporal characteristics of the source excitation of the analysis frame, and had been demonstrated to be able to improve the speaker recognition performance when serving as a supplementary feature to LPCC [7].

In this article, instead of doing time-frequency analysis of the LP residual signal for each frame of a fixed duration, we adaptively select the analysis window length to be exactly equal to two pitch cycles by pitch synchronous analysis. Wavelet transform with the Dau-bechies wavelet and 6 dyadic (or octave) dilated baby wavelets is applied to the pitch synchronized LP residual signal. Similar to our previous feature extraction method, the wavelet transform coefficients corresponding to each baby wavelet are grouped together to form an octave group. And the energy of each octave group (or the time-indexed subgroups of each octave) is computed to form the so called Wavelet Octave Coefficients of Residues (WOCOR). While it does not provide the true glottal flow, information related to the frequency composition as well as the dynamic evolution within each pitch cycle can be characterized.

The remainder of this paper will first give a detailed outline of the proposed method and then experimental results will be presented to demonstrate the effectiveness of WOCOR for speaker recognition.

2. Time-frequency feature extraction

Most existing speech/speaker recognition systems adapt a frame-based feature extraction approach where vocal tract transfer function is assumed to be stationary within
the analysis frame. However, study of the glottal waveform showed that the dynamic evolution of the glottal source within a short period of time is much more non-stationary than that of the vocal tract system, i.e., the burst changing of the amplitude caused by the rapid vocal cord closing, and the time-varying characteristics over successive glottal cycles even during the sustained voicing [8]. The dynamic evolution of glottal waveform is highly related to the speaker specific larynx characteristics, and thus should be useful for speaker recognition.

To exploit the time-frequency information of the source excitation, we present a wavelet transform based feature extraction technique, as described in Figure 1.

**Voiced Speech Detection.** Only the voiced segment is concerned in our method since the unvoiced speech segment does not contain too much information of the vocal source production mechanism. A two step algorithm is used to determine the voiced sound. First, an energy detector is applied to remove the silence and most of the unvoiced portions that have considerably lower energy than the voiced segments. Then, the ‘filtered’ speech signal is fed through a zero-crossing detector to eliminate the remaining unvoiced sounds.

**Pitch Estimation.** Pitch is estimated using cepstrum analysis for every 32 ms of voiced speech [9]. The pitch period is used for window length selection and pitch pulses detection in the following steps.

**Windowing.** The voiced speech is segmented with a rectangular window with variable length equals to 2.5 times of the estimated pitch period.

**LP Inverse Filtering.** The windowed speech frame is inverse filtered to generate the LP residual signal. The 12th order LP coefficients are computed by covariance method [1].

**Pitch Synchronous Analysis.** The residual signal of the voiced speech essentially represents bursts at the vocal cord closing. If we define the period between two successive bursts to be a pitch cycle, then pitch synchronous analysis can be achieved by detecting these bursts.

**Wavelet Transform of** $u(n)$. With the above process, the analysis window for wavelet transform is now strictly constrained to be exactly 2 pitch cycles and 1 pitch cycle overlap with every window starting at the excitation epoch. The wavelet transform of $u(n)$ can be expressed as

$$ W_u(a,b) = \frac{1}{\sqrt{|a|}} \sum u(n) \Psi^\ast \left( \frac{n}{a} - b \right) $$

where $\Psi(n)$ is the mother wavelet function, $a$ is the dilation (or scaling) parameter, and $b$ is the translation parameter. The resolution in time can be trade-off for resolution in frequency by selecting various scaling parameters. For a specific resolution, the time-varying characteristics can be measured as the translation parameter $b$ changes. Figure 2 shows a segment of two pitch cycles residual signal (top panel) and its 3 wavelet transforms in different scales (the bottom 3 panels). As shown, as $a$ increases, time resolution decreases, whilst the frequency resolution improves. Also, the time varying characteristics of $u(n)$ can be measured from $W_u(a,b)$ at different translation parameters.

**Time-Frequency Feature Generation.** To generate the time-frequency feature vector from $u(n)$, we first do wavelet transform using Daubechies wavelet with dyadic (or octave) scaling parameter,

$$ a = \{2^k | k = 1, 2, \ldots, 6 \} $$

and translation parameter,

$$ b = 1, 2, \ldots, N $$

where $N$ equals to the length of $u(n)$. All the $W_u(a,b)$’s with a specific scaling parameter can be considered as the frequency analysis of the signal with a

---

**Figure 1.** Block diagram for Time-Frequency feature extraction of the source excitation.

**Figure 2.** Analyzed LP residual signal (top panel) and its wavelet transforms (bottom three panels) with scaling parameter $a=2, 4$ and $8$, respectively.
particular time-frequency resolution, and can be grouped together as
\[ W_k = \left\{ W_s (2^j, b) | b = 1, 2, \cdots N, k = 1, 2, \cdots 6 \right\} \] (5)
Each \( W_k \) is called an octave group. Finally, we derive the new feature vector as
\[ \text{WOCOR}_\alpha = \left\{ \| W_s \|_k = 1, 2, \cdots 6 \right\} \] (6)
where \( \| \cdot \| \) denotes the 2-norm operator. In this case, the feature vector has 6 elements containing only frequency information, but not the temporal characteristics. To retain the temporal details, each octave group can be equally divided into 2 subgroups and then the energy of each subgroup is computed to generate a double sized feature vector noted as WOCOR\(_\alpha\). There are now 12 elements in the WOCOR\(_\alpha\) and provides a certain degree of temporal information of the constituent frequency components. To extend further so as to obtain more detailed temporal characteristics, each octave group can be further divided into \( \alpha (\alpha \geq 2) \) subgroups,
\[ W_s ^\alpha (j) = \left\{ W_s ^\alpha (2^j, b) | b \in \{ j-1 : j \} \times \text{Round}(N/\alpha) \right\} \] (7)
Note that the final subgroup of each octave may have more or less components than \( \text{Round}(N/\alpha) \). Finally, a \( 6\alpha \)-dimensional feature vector can be generated
\[ \text{WOCOR}_\alpha = \left\{ \| W_s ^\alpha (j) \|_k = 1, 2, \cdots 6 \right\} \] (8)

3. Experiments

3.1. Speech corpus

The speech database used in this experiment is a subset of the Cantonese database collected at the Chinese University of Hong Kong for speaker recognition purpose. In this subset, there are totally 42 male speakers, each of whom completed 10 recording sessions within 4 months with at least one week interval between two successive sessions. The first 4 parts were enrollment sessions with 30 utterances each, while the last 6 parts were testing ones with 6 utterances per session. In each session, the speakers were prompted to read at their normal loudness and speaking rates a series of Hong Kong ID card number (e.g. “T984760”). The speech data were collected through a microphone in a reasonably quiet recording room. The sampling frequency is 8000 Hz.

3.2. Baseline system

The feature vector used in the baseline system was the 36-dimensional LPCC\(_\text{D-A}\), which contains the 12-dimensional Linear Predictive Cepstral Coefficients and their first and second order time difference coefficients (ΔLPCC and ΔΔLPCC). The LPCC feature vector was generated by Hamming windowed speech signal (both voiced and unvoiced portions) with 32 ms window length and 22 ms overlaps. The 12th order LP coefficients were estimated by autocorrelation method. The speaker models were trained by 128 components Gaussian Mixture Model (GMM) [10] over the first 3 enrollment sessions. Each speaker has 36 identification tests with an utterance for every single test. In verification, 5 cohort speakers were selected for every one over the fourth enrollment session [11]. Each speaker has 36 tests against himself and 1296 tests from the impostors, excluding his cohort speakers.

3.3. Recognition results with WOCOR\(_\alpha\)

The model training technique for WOCOR\(_\alpha\) was the same as that for the LPCC\(_\text{D-A}\). Figure 3 shows the recognition results of WOCOR\(_\alpha\) at various \( \alpha \). As shown, as \( \alpha \) increases, there is a significant decline of the error rate when \( \alpha < 4 \), which illustrates the importance of the temporal information of the excitation signal for speaker characterization. Also it can be seen from the figure that no further considerable reduction in the recognition error rate can be achieved for \( \alpha > 4 \). This may be due to the redundancy between adjacent feature components of the subgroups with very small time span, and the insufficient training data for the relatively large feature size.

![Figure 3. Speaker recognition results with WOCOR\(_\alpha\)](image)

**IDER / EER (%)**

<table>
<thead>
<tr>
<th>Identification Error Rate</th>
<th>Verification Equal Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>15%</td>
</tr>
<tr>
<td>2%</td>
<td>20%</td>
</tr>
<tr>
<td>3%</td>
<td>25%</td>
</tr>
<tr>
<td>4%</td>
<td>30%</td>
</tr>
<tr>
<td>5%</td>
<td>35%</td>
</tr>
<tr>
<td>6%</td>
<td>40%</td>
</tr>
<tr>
<td>7%</td>
<td>45%</td>
</tr>
<tr>
<td>8%</td>
<td>50%</td>
</tr>
<tr>
<td>9%</td>
<td>55%</td>
</tr>
<tr>
<td>10%</td>
<td>60%</td>
</tr>
</tbody>
</table>

3.4. Information fusion for speaker recognition

For speaker recognition with fused vocal tract (LPCC\(_\text{D-A}\)) and vocal source information (WOCOR\(_\alpha\)), the final decision score is based on a weighted combination of the decision scores of each feature vector,
\[ S = w_t S_t + w_s S_s \]
with \( w_t + w_s = 1 \)

where the subscripts \( t \) and \( s \) correspond to vocal tract and vocal source features, respectively. For comparison, WOCOR\(_\alpha\), which has the same feature size as LPCC\(_\text{D-A}\), is used in the following tests. The weighting coefficients can be determined experimentally. As can
be seen from Figure 4, both identification and verification achieve the best performance when $w_i = 0.9$.

Table 1 gives the respective recognition results of the vocal tract information (LPCC_D_A), the vocal source information (WOCOR6), and their information fusion. Compared with the recognition results with LPCC_D_A, the discriminative power of WOCOR6 is not convincing. For example, at $w_i = 0.9$, the information fusion results in 1.59% for IDER and 2.26% for EER, which offers a relative improvement of the performance by 19.7% and 14.4%, respectively. The recognition results of MFCC_D_A, which has been employed in most of the speaker recognition systems, is 1.92% for IDER and 2.48% for EER, as also given in Table 1. The proposed information fusion technique outperforms that with only MFCC_D_A by 17.2% for IDER and 8.87% for EER relatively. An extensive speaker recognition test is undergoing and more detailed results will be reported in due course.

![Figure 4: Recognition results with information fusion](image-url)

**Figure 4:** Recognition results with information fusion

**Table 1:** Recognition results for different features

<table>
<thead>
<tr>
<th>Performance</th>
<th>IDER (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPCC_D_A</td>
<td>1.98</td>
<td>2.64</td>
</tr>
<tr>
<td>WOCOR6</td>
<td>29.17</td>
<td>14.45</td>
</tr>
<tr>
<td>Info-fusion(wi=0.9)</td>
<td>1.59</td>
<td>2.26</td>
</tr>
<tr>
<td>MFCC_D_A</td>
<td>1.92</td>
<td>2.48</td>
</tr>
</tbody>
</table>

### 4. Conclusions

In this paper, we present a feature extraction technique to effectively retrieve the time-frequency information from the source excitation. Upon pitch synchronous analysis, the wavelet transform is applied to the LP residual signal with exactly 2 pitch cycles. The feature set WOCOR$_w$ with $\alpha > 1$ characterizes the speaker-specific spectro-temporal properties of each pitch cycle, as well as the time evolution over successive pitch cycles. The recognition tests show the discriminative power of WOCOR$_w$ for speaker recognition. Especially, when employing WOCOR$_w$ as a complementary characteristic to the vocal tract feature (LPCC_D_A), speaker recognition system with fused information relatively outperforms that with only MFCC_D_A by 17.2% for identification and 8.87% for verification.

### 5. Acknowledgements

This effort was partially supported by a research grant awarded by the Hong Kong Research Grant Council.

The first author wishes to acknowledge Dr. Frank Soong for instructive discussions and suggestions during this work.

### 6. References