Robust Speaker Recognition Based on High Order Cumulant

Jiyong Ma and Wen Gao

Institute of Computing Technology,
Chinese Academy of Sciences, 100080, China
{jyma, wgao}@ict.ac.cn

Abstract

LP-derived cepstral coefficients are sensitive to additive noise in speech signal. In this paper, an approach to extracting speech feature based on the high-order cumulant is proposed to depress the effect of additive noise in speech signal. The performance of this approach is evaluated using a text-prompt speaker verification system. Experimental results show that this approach is effective to increase the robustness of the speaker recognition system.

1. Introduction

The research of speaker recognition began in 60's. B.S.Atal[1] showed that the recognition performance using LPC coefficients as speech features is better than that of the spectrum of filter bank and the performance of the Cepstral of LPC analysis is the best. SIMAI[2] developed the Frequency Cepstrum Coefficients and applied in speaker recognition. H.Hermansky[3] introduced the Perceptually based Linear Predictive analysis, experiments showed that PLP was successful used in speech recognition and speaker recognition. The recent development of speaker recognition was reported by Furui[4].

Linear Predictive analysis is one of the most widely used front-end processors for various recognition tasks in speech processing. It is widely accepted that the most reliable LP-derived feature set for speaker recognition is the cepstral coefficients. Cepstral features are found to yield excellent performance for speaker recognition when training and testing speech signal collected under the same environments. However, in practical applications, the speech used by the recognition system is subject to various sources of degradations such as background noise. Such degradations often result in the reduced recognition rates. This is due to the mismatch created among corresponding reference and testing features.

The conventional LPC are derived by the Yule-Walker equations with respect to autoregressive coefficients. The second order statistics like autoregressive coefficients are sensitive to noises in speech signals. To depress the effect of additive noises on autoregressive coefficients, an approach to extracting speech feature based on the theory that high order cumulant can depress the additive Gaussian noises in signals is proposed in this paper.

This paper is organized as follows. In Section 3, speech feature extraction approach based on high-order cumulant is presented. In Section 5 the performance of the speaker verification system based on the feature extraction approach is demonstrated. Finally, the summary and conclusion are given.

2. Feature Extraction Based on High-order Cumulant

The most popular feature extraction approach in speaker recognition is LPC cepstral approach. The conventional LPC are derived by the Yule-Walker equations. However, the second order statistics like autoregressive coefficients are sensitive to noises in speech signals. To depress the effect of additive noises on autoregressive coefficients, this paper proposes an approach to extracting speech feature based on the theory that high-order cumulant can depress the additive Gaussian noises in signals. The theoretical basis is that if the noises are Gaussian noises, their cumulants above third order are zero. This means that high-order cumulants are insensitive to Gaussian noises. From high-order cumulants, we can reconstruct the second order statistics that will be used in calculating the LPC.

First, let me review some basic concepts. Suppose that $x$ is random variable, its first characteristic function is

$$
\Phi(\omega) = E\left\{ e^{j\omega x} \right\}
$$

Its k-order moment is as the following
\[ m_k = \frac{d^k \Phi(\omega)}{d\omega^k} \bigg|_{\omega=0} \]

And its second characteristic function is
\[ \Psi(\omega) = \ln \Phi(\omega) \]

The k-order cumulant is defined as the following
\[ c_k = \frac{d^k \Psi(\omega)}{d\omega^k} \bigg|_{\omega=0} \]

Theoretically, if x is a random variable with zero mean, and its distribution is Gaussian, then its second characteristic function is
\[ \Psi(\omega) = -\frac{1}{2} \omega^2 \sigma^2 \]

Therefore, its cumulants are
\[ c_1 = 0, c_2 = \sigma^2, c_k = 0, k \geq 3 \]

This shows that if the noise is Gaussian, its cumulant above third order are zero.

For speech signal, first, the speech signal is emphasized with the filter \(1 - 0.97z^{-1}\). Then the procedure to extract speech feature is as follows.

For a stationary random process with zero mean \(\{s(n)\}\), its forth-order cumulant is defined as
\[ C_v(s, \tau, \tau', \tau'') = E(s(n)s(n+\tau)s(n+\tau)s(n+\tau')) - R(s(n))R(s(n+\tau))R(s(n+\tau'))R(s(n+\tau'')) \]

Where \(R(s) = E\{s(n)s(n+\tau)\}\)

is the autoregressive coefficients of the random process \(\{s(n)\}\).

The consistent estimation of the forth-order cumulant is as
\[ \hat{C}_v(s, \tau, \tau', \tau'') = \frac{1}{N} \sum_{n=0}^{N} s(n)s(n+\tau)s(n+\tau)s(n+\tau') - R(s(n))R(s(n+\tau))R(s(n+\tau'))R(s(n+\tau'')) \]

\[ \hat{R}(s) = \frac{1}{N} \sum_{n=0}^{N} s(n)s(n+\tau), \quad \tau = 0.1\Lambda, \quad \hat{R}(s; -\tau) = \hat{R}(s) \]

Where, for \(n < 0\) and \(n > N - 1\), \(s(n) = 0\).

The linear predictive error of speech signals is as
\[ \sum_{j=0}^{k} a_j s(t - j) = e(t) \]

In general, the filter coefficients \(a_0, a_1, \ldots, a_p\) are obtained by whitening the speech with additive noises, namely, it enables \(E[e^2(n)]\) to minimize, it is not to whiten the pure speech signals. To whiten the pure speech without contaminating with additive noises, we use the following technique [5]

\[ C\alpha = \pi \]

Where
\[ \alpha = (a_0, a_1, \ldots, a_p) \]

\[ C = (c(i-j))_{0 \leq i, j \leq p} \]

\[ \pi = (E_p, 0.1\Lambda, 0) \]

\[ c(i) = \sum_{i=0}^{p} C_v(0.1\Lambda, \tau + i) \]

\[ E_p = c(0) - \sum_{i=1}^{p} C_v(0.1\Lambda, i) \]

From the above equations, it is seen that the form of the equation is the same with that of the conventional linear predictive approach, the difference is the autoregressive coefficients \(\hat{R}(s)\) are replaced with autoregressive coefficients reconstructed from forth-order cumulant \(c(i)\). After the linear coefficients \(a_0, a_1, \ldots, a_p\) are solved by the above equations, the cepstral coefficients are obtained in the same way as the conventional approach.

The spectral slope distance as the following is used as a measure
\[ d^2 = \sum_{m=1}^{M} m^2 \left( c_m - c'_m \right)^2 \]

Where the coefficients \(c_m, c'_m\) are the test cepstral coefficients and reference coefficients respectively.

3. Performance Evaluation

3.1 The Speaker Verification System

A screen saver based on human face detection and text-prompt speaker verification is designed to test the techniques proposed in this paper. When the human face module detects that there is human face in the front of a computer, the module notifies the speaker verification module to prompt a password consisting of a digital string. Every user has at most three tries to enter the system during the test the performance of the system. The recognition approach used in this system is the conventional continuous HMM. For each digit,
The left-to-right HMM is trained with training samples. Each HMM has six states. The length of digital string is 6.

3.2 Database

The training and testing conditions: The speech signal is sampled at 8 kHz, digitized with 16 bits. The emphasis coefficient is 0.97, the frame period is 8ms and the frame length is 32ms. The speaker set consists of 40 speakers, the number of female speakers is 20, and the number of male speakers is the same. The training database consists of speech utterances of each Chinese digit 3 times. And the test database consists of speech utterances of digital string with length 6, which are recorded five times in different periods. Thus, each client has 5 times to register the system, and the 245 times for an imposter test with the cross test among the speakers. Another way to test the system performance is on-line test. We asked users to record the false rejection times and false acceptance times when they entered the system.

3.3 Comparison with LPCC and HC-LPCC

First we examine the robustness of the approach of High-order Cumulant-Linear Predictive Cepstral Coefficients. For a male speaker, his speech utterance of digit “one” is added with simulated Gaussian noises. The experimental results are shown in the Fig.2. In the figure, HC-LPCC denotes the high-order cumulant linear predictive cepstral coefficients without adding any simulated additive noises, H30dB and H10db are the results when the signal to noise is 30 dB and 10 dB. Similarly, for the conventional cepstral coefficients, the result is shown in Fig 1.

Table 1 gives the relative change rates of the maximum cepstral coefficient when the signal to noise ratio is 30dB.

<table>
<thead>
<tr>
<th></th>
<th>HC-LPCC</th>
<th>LPCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative</td>
<td>20.6%</td>
<td>30.1%</td>
</tr>
</tbody>
</table>

Table 1. The Relative Change Rate of the Maximum Peak

From the table, we know that the relative change rate of the maximum cepstral coefficient based on high order cumulant is smaller than that of traditional cepstral approach. Therefore, HC-LPCC is more robust than LPCC.

The cohort normalization is used for the experiment. In our experiment, the normalized likelihood is as the following

\[ L(x) = \frac{1}{C} \sum_{c=1}^{C} L(x|S_c) \]

Where \( S_t \) is the target speaker, \( S_c (c=1,2,..,C) \) are cohort speakers, and \( x \) is the feature vector. The number of cohort speakers is set to 8 in the experiments.

![Figure 1. The Average Cepstral Coefficients of Digit “One” calculated by the traditional approach](image1)

![Figure 2. The Average Cepstral Coefficients of Digit “One” Calculated by High-order Cumulant Approach.](image2)
For different order of cepstral coefficient and different ratios of signal to noise, two experiments were conducted. For the case of traditional LPCC, the Equal Error Rates are shown in the Fig.3 for different order of cepstral coefficient. For the case of HC-LPCC, the Equal Error Rates are shown in the Fig.4 for different orders of cepstral coefficients. The relatively reduced error rates are listed in the table 2. From the table, it is seen that HC-LPCC is very effective to reduce the ERR.

Table 2. Error Reduced Rates

<table>
<thead>
<tr>
<th></th>
<th>30dB</th>
<th>10dB</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>14.9%</td>
<td>20.1%</td>
<td>12</td>
</tr>
<tr>
<td>14</td>
<td>15.3%</td>
<td>20.2%</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>15.8%</td>
<td>21.9%</td>
<td>16</td>
</tr>
<tr>
<td>18</td>
<td>15.8%</td>
<td>22.3%</td>
<td>18</td>
</tr>
</tbody>
</table>

4. Summary and Conclusions

We have developed a speech feature extraction approach based on the high-order cumulant for speaker recognition. The performance of the approach was evaluated using off-line test speech and a screen saver based on face detection and speaker verification. Experimental result show that approach is capable of improving both the verification performance and the robustness of the system. The HC-LPCC can depress the additive noises in speech, this performance is very important for a speaker verification system in practical applications because the environment is usually noisy.

References