Prosodic Features Based on Wavelet Analysis for Speaker Verification


Department of Electronic Science and Technology
University of Science & Technology of China
chenjixu@mail.ustc.edu.cn sunjun@ustc.edu

Abstract

Most conventional speaker recognition systems rely on short-term spectral information. But they ignore the long-term information such as prosody which also conveys speaker information. In this paper, we propose an approach that extracts prosodic features based on long-term information. First, by making wavelet analysis, we can reveal the trends of the f0 and energy contour. Subsequently, the prosodic features are extracted only from approximation coefficients. We use these features in a GMM-UBM based text-independent speaker verification system. The proposed method achieves an EER of 23.3% on the NIST2004 8sides-1side task scheme. This result is promising while the baseline system, which uses short-term f0 feature, only results in an EER of 33.49% in this task.

1. Introduction

Most current speaker recognition systems are limited to use frame-based spectral information. While these systems produce very good result, they ignore other information of the speaker, such as the prosodic information from pitch and energy contour. However, it's proved that using prosodic feature can effectively improve the performance and robustness of speaker recognition system. [1,2,3].

Previous works have utilized the prosodic features for speaker verification in two main ways. The first approach is to model global distributions of f0 and energy's frame-based features [4,5]. One problem for this global statistic approach is that it does not capture the temporal dynamic information of prosodic features. Although this can be addressed partly by using statistics of feature time derivatives, it's still based on short-term information. The second approach focus on capturing and modeling the long-term prosodic features. In [3], Dynamic Time Warping is applied to compare the trajectories of frequent words, but it is based on ASR. The more general method in [6] uses piece-wise linear approximation to stylize the f0 and energy contour, and subsequently employs bi-gram for modeling the slope and duration of the stylized segments.

The approach in [6] not only yields a satisfactory result, but also reveals some characteristics of prosodic feature. For example, the piece-wise linear approximation shows that the low-frequency component of f0 contour may convey more speaker-specific information when large amount of training and test data is available. On the contrary, the effectiveness of frame-level features which represent the details of the contour is limited in this condition.

Wavelet analysis is an effective way to remove the details and reveal the trends of a signal. In this paper, we make the simple wavelet analysis of the f0 and energy contour. Our experiments show that approximation coefficients can represent the useful low-frequency component. And the features extracted from the sequence of the coefficients can well describe the speaker-specific trend of dynamic features in long-term.

The NIST04 8sides_1side task scheme is used to evaluate our new prosodic feature and great improvement is achieved compared with the system that uses short-term f0 and energy features.

The paper is organized as follows: in section 2, the approximation of pitch contour using wavelet analysis is illustrated. In section 3, we present our method to extract prosodic features from wavelet coefficients. The experiments and results are showed in section 4.

2. Wavelet analysis of pitch contour

The wavelet decomposition tree in Fig.1 shows how the approximation coefficients (cA) are gained. We suppose that the iterated low-pass filter will omit the details and reveal trends of pitch contour, while the down-sampling can compress the temporal information and retain the long-term features.

![Figure 1 The wavelet decomposition tree](image-url)
2.1. Pitch tracker and Pre-Processing.

The Pitch tracker used in this work is based on Robust Algorithm for Pitch Tracking (RAPT) as implemented in ESPS get_f0 (Entropic Corp).

Although the Lognormal Tied Mixture (LTM)[4] is an effective way to correct doubling/halving errors, we uses an simpler hard threshold to correct these errors for each utterance. The threshold is set as the median over this utterance. After this process, a median-filter is applied to generate the estimate of f0 sequence.

2.2. Wavelet analysis

To Realize the piecewise linear approximation, we utilize a biorthogonal discrete wavelet transform (DWT) with a triangular shape scaling function. Fig.2 illustrates approximation of the prosody contour from level 2 to level 5. The level of wavelet analysis can be viewed as the scale of a map. Higher the level is, larger the approximation scale is applied and more high-frequency details are filtered. For our speaker verification experiments, we find that the level 3 approximation coefficients convey the most effective information. The first two levels contain too many details that tend to be noise. On the other hand, the higher levels omitted some useful long-term information.

As shown in Fig.2, the level 3 approximation is a good way to fit the pitch contour. Because of the low-pass filters used in the wavelet analysis, some detailed noise is removed, what's left is the slower trend of the pitch contour. Meanwhile, temporal information is greatly compressed, the number of cA3 is only about 1/8 of the pitch features. So fewer coefficients can capture long-term information.
3. Feature Extraction

3.1. Data Cleaning

As can be observed from Figure 3, although some short region of no pitch is smoothed in approximation, there are still some long unvoiced(UV) regions which contain no information of pitch contour. By setting a global threshold to remove the low coefficients, we do not take into account these long regions.

3.2. Feature extraction

When using the cA3 coefficients as features, we found that a first-order Markov (bi-gram) is not enough to capture the variational patterns of pitch. So every four successive cA3 coefficients are combined into one four dimensional feature vector which can represent 3rd-order Markov dependence of the cA3 sequence. Meanwhile, we follow the same wavelet analysis of \((\log)\)Energy contour. Since we found that the absolute energy and the short-term dynamic patterns of energy are not reliable in speaker verification task, we perform a linear fitting for each four successive cA3 coefficients \([A_{k1}, A_{k2}, A_{k3}, A_{k4}]\). Based on MSE, we can get the slope of fitting line as:

\[
ESlope_k = (3A_{k4} + A_{k3} - A_{k2} - 3A_{k1})/10
\]

The energy slope is combined with the corresponding pitch feature to form a new five dimensional feature vector. The experiments show that the added energy slope greatly improve the performance.

4. Experiments and results

4.1. Database

The performance of the proposed method is evaluated on 2004 NIST Speaker Recognition Evaluation (SRE) 8sides_1side task [7] which uses eight single channel conversational sides of one speaker for training and then test on one single channel conversation side. Each conversation side is about 5 minute including silence duration. Our experiment only uses 170 male speakers in the database, along with a total of 8088 verification trials.

4.2. Baseline

Although the baseline system in [3] uses the f0 and energy features, the frame-based energy feature is experimentally not reliable in our task. So, our baseline system only uses global distribution of f0 features. For each voiced frame, a two-dimensional feature vector composed of \(\log f0\) and its derivative estimated over a 5-frame context is created. The universal background model (UBM) is a 256-component Gaussian mixture model (GMM) trained with 112 male speakers from the NIST01 SRE database. Given the UBM, the target speaker models are then derived using Bayesian adaptation [8]. The Equal Error Rate (EER) of this system is 33.49% for the nist04 8sides_1side task.

4.3. Experiments and results

We adopt the same GMM-UBM scheme the baseline to model the distribution of prosodic features. Considering the fewer features, we use a UBM of 128 mixtures which is also trained from nist01 SRE database. To estimate the effect of long term energy slope, we also did the experiment without it.

Figure 4 shows the DET curves [9] when using the two different kinds of features and the performance of the baseline system. The features extracted only from pitch contour result in an EER of 26.43%, a relative reduction of 21.08% compared with baseline system. This shows that our prosodic features based on wavelet analysis is useful for capturing speaker-specific information.

Using Energy Slope combined with our pitch contour features, the EER is reduced to 23.33%, indicating that the energy slope based on four approximation coefficients is, indeed, adding new information.

We also explore the effect of decreasing the training data. Table 1 shows the performance for the system using prosodic features when training with 1, 2, 4, 8 conversation sides. It shows that system only using pitch feature outperforms the baseline no matter the number of training sides.
Table 1: Systems performance (EER) per number of training sides

<table>
<thead>
<tr>
<th>Training Sides</th>
<th>Pitch Feature EER(%)</th>
<th>Pitch Feature+ Energy Slope EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.26</td>
<td>27.25</td>
</tr>
<tr>
<td>2</td>
<td>28.36</td>
<td>26.16</td>
</tr>
<tr>
<td>4</td>
<td>27.29</td>
<td>24.89</td>
</tr>
<tr>
<td>8</td>
<td>26.43</td>
<td>23.33</td>
</tr>
</tbody>
</table>

5. Conclusion and future directions

In this paper, we utilize the simple wavelet analysis to reveal the trends of pitch and energy contour. Then, the prosodic features are extracted from approximation coefficients. Experiments on the NIST04 8side_1side task showed that these features can represent the speaker-specific information. And the proposed features result in better performance than the frame-based pitch features.

However, this is just the beginning of using wavelet analysis to capture prosodic features. First, we need to examine better ways of feature extraction from different DWT levels which are expected to deliver information in different scale. Secondly, we are interested in incorporating additional features such as the frame-based Prosodic features and unvoiced region duration. Thirdly, we need to combine the prosodic features with spectral features to further improve the speaker verification performance.

6. Acknowledgements

This work has been supported by the National Science Foundation of China (60272039)

7. References
