Comparison of Features Extracted Using Time-Frequency and Frequency-Time Analysis Approach for Text-Independent Speaker Identification

Nirmalya Sen¹, Tapan Basu²
¹²Signal Processing Research Group, CET
IIT Kharagpur, India
Email¹: nirmalya75@rediffmail.com
Email²: basutk02@yahoo.co.in

Sandipan Chakroborty
Lead Engineer
SAMSUNG
Bangalore, India
Email: mail2sandi@gmail.com

Abstract—This paper compares the feature sets extracted using time-frequency analysis approach and frequency-time analysis approach for text-independent speaker identification. Mel-frequency cepstral coefficient (MFCC) feature set and Inverted Mel-frequency cepstral coefficient (IMFCC) feature set are extracted using time-frequency analysis approach. Temporal energy subband cepstral coefficient (TESBCC) feature set is extracted using frequency time analysis approach. Time-bandwidth product of MFCC filter bank and TESBCC filter bank has been compared. RV coefficient has been used to calculate the correlation between the feature sets. Experimental evaluation was conducted on POLYCOST database with 130 speakers using Gaussian mixture speaker model. The TESBCC feature set has 9.5% higher average accuracy compared to the MFCC feature set. It is found that, the feature set extracted using time-frequency analysis approach is practically uncorrelated with the feature set extracted using frequency-time analysis approach. It is also demonstrated that IMFCC feature set has important role in fusion.

Keywords— Feature extraction ; time-frequency analysis; GMM; Nyquist filter; frequency-time analysis; POLYCOST database

I. INTRODUCTION

We know that the physiological structure of a vocal tract is different for different persons. Due to this, we can differentiate one person’s voice from others. This difference in vocal tract structure is reflected in the frequency spectrum of speech signal. People have been using this speech spectrum for speaker identification for a very long time [1]. The Mel-frequency cepstral coefficients (MFCC) feature was first proposed for speech recognition [2]. This is a filter bank based approach but implemented using time-frequency analysis technique. Here, first time analysis is done through framing operation and then frequency analysis is done by passing that frame through filter bank. As the time analysis is done first, to handle the quasi periodicity of the speech signal, MFCC needs overlapping frames. Filters are designed in such a way that they resemble the human auditory frequency perception. Later people used MFCC satisfactorily for speaker recognition also [3]. Presently MFCC is the most widely used feature set for speaker recognition. In the MFCC filter bank, the low frequencies are given more importance compared to the high frequencies. This structure is very much suitable for speech recognition. But for speaker recognition point of view, researchers performed various experiments through which it is evident that high frequency zone should also be given similar importance as the low frequency zone.

S. Hayakawa and F. Itakura found that, the speaker recognition rate of the frequency band from 0 to 4 kHz was roughly the same as that of the frequency band from 4 to 10 kHz [4]. They observed that a relatively small amount of speaker specific information is available in the frequency region between 500 Hz to 2 kHz. They concluded that a rich amount of speaker specific information is contained in the higher frequency band, and it is useful for speaker recognition. Similarly L. Besacier and J.F. Bonastre from their detailed investigation using subband architecture concluded that, the low-frequency (under 600 Hz) and the high-frequency (over 3 kHz) contain more speaker specific information than the middle-frequency subbands [5]. Recently X. Lu and J. Dang demonstrated that the speaker specific information is concentrated mainly in three regions in the frequency domain [6]. Glottal information is in the region from 50Hz to 300 Hz. Piriform fossa information is in the region from 4kHz to 5.5kHz. Third region is from 6.5 kHz to 7.8 kHz which may be related to the consonants. They observed that comparatively less information was available from 500Hz to 3.5 kHz.

S. Chakroborty et al. proposed a flipped MFCC filter bank using the inverted Mel scale [7]. After that we calculated a new feature set named Inverted Mel Frequency Cepstral Coefficients (IMFCC), following the same procedure as normal MFCC but using the reversed filter bank structure. Identification accuracy of features calculated using flipped MFCC filter bank was comparable with conventional MFCC features. We also used the fusion of IMFCC features with MFCC features. Identification accuracy, after fusion, was higher compared to MFCC features. Later H. Lei and E.L. Gonzalo also observed the importance of fusion using IMFCC features for speaker verification task [8].

But all the above works were based on the time-frequency analysis approach. In our earlier works [9]-[11], we proposed a new Nyquist window [12] and using the cosine modulation...
of that window we constructed a filter bank. We used that filter bank for text-independent speaker identification in a frequency-time analysis approach. Here first frequency analysis is done through filter bank and then time analysis is done through framing the output signal of each filter. It is easy to visualize that, due to the inner product operation of the speech signal and the filter impulse response; there is no need to take overlapping frames to handle the quasi periodicity of the speech signal in case of frequency-time analysis approach.

In this paper we have given a brief introduction of IMFCC feature set and TESBCC feature set. The concept of RV analysis is introduced. We have used the standard database POLYCOST for closed set Text-Independent speaker identification experiments. For classification we have used Gaussian mixture speaker model [3]. Split v-q has been used to initialize the GMM. The detailed results of MFCC, IMFCC and TESBCC feature sets and their combinations of fusion are given in TABLE I to TABLE V.

II. IMFCC FEATURE CALCULATION

In MFCC feature extraction process which was mainly derived for speech recognition, overlapped triangular filters are placed on the Mel scale at equal distance. To increase the frequency resolution in the high frequency range, we derived the inverted Mel warping function (Figure 1.) and placed overlapped triangular filters on the inverted Mel scale at equal distance. As a result, in usual frequency scale, filters are placed densely in the high frequency range and sparsely in the low frequency range (Figure 2.). Cepstral coefficients are calculated using the inverted Mel filter bank in place of the Mel filter bank. The detail procedure is given in our earlier publication [7]. The inverted Mel warping function is given below,

$$w[n] = \frac{2\gamma^2 \sin(2\pi n/N)}{\pi(4\gamma n - n')} \quad \text{for } n \neq 0, \pm (N/2)$$

$$= (1/N) \quad \text{for } n = 0$$

$$= (1/2N) \quad \text{for } n = \pm (N/2)$$

Here \( \gamma = (N/4) \)

\( \alpha(k) = 1 \) when \( k = 0 \) and \( (N/2) \)

\( = 2 \) when \( k = 1, 2, ..., (N/2) - 1 \)

Equation (4) is necessary to keep the magnitude response of the overall filter bank all pass. \( N=12 \) implies the decomposition of the speech signal into seven sequences. The Fourier transform of window function is given below,

$$FT(w[n]) = W(e^{j\omega}) = \cos^2(\omega) - (2\pi/N) \leq \omega \leq (2\pi/N)$$

$$= 0 \quad \text{when } |\omega| > (2\pi/N)$$

Figure 3. shows a typical window function \( w[n] \) for \( N=12 \) and length \( L = 61 \). Centre lobe of the window extends from \(- (N-1)\) to \( (N-1) \) and contains more than 99.948% of total energy. Figure 4. shows the magnitude spectrum of the window function for \( N=12 \) and \( L = 61 \). Figure 5. shows the filter bank generated using the window function \( w[n] \) for \( N=12 \) and \( L = 61 \). There are total seven filters in the filter bank as evident from equation (2). The all pass property of the overall magnitude response of the filter bank does not depend on the window length, because our window function is a Nyquist filter [12].

III. TEMPORAL ENERGY SUBBAND CEPSTRAL COEFFICIENT

This feature TESBCC is extracted using frequency-time analysis approach [9]-[11]. This is implemented using an FIR allpass complementary Nyquist filter bank as shown in Fig.6. Here the speech signal \( x[n] \) is decomposed into some finite \((l + (N/2))\) number of sequences \( x_k[n] \) as follows,

$$x_k[n] = \alpha(k) \sum_{m=-\infty}^{\infty} x[m]w[n-m] \cos[2\pi k(n-m)/N] \quad (2)$$

Here \( N \) is an even number. The Nyquist window function \( w[n] \) is defined as below,
The energy of each frame is calculated. Log compression and DCT is applied to get the cepstral coefficients.

IV. COMPARISON OF THE TIME-BANDWIDTH PRODUCT

It is known that for a given function the time-bandwidth product is unique [13]. It is also known that, the Gaussian function has the lowest time-bandwidth product which is equal to 0.5. But the Gaussian function is not a Nyquist filter. Hence it cannot produce an allpass complementary filter bank. As a result some frequency will lose [9]. Hence we have not used the Gaussian window.

In our earlier publication [9] it was shown that the frequency domain of the window function \( w[n] \) is differentiable at each point throughout its domain. The time-bandwidth product of the proposed window function is 0.513 which is very near to the Gaussian window. It was also shown that [11], the time-bandwidth product of MFCC filter is 0.55968. As the IMFCC filter bank is derived from the MFCC filter bank by interchanging the filters positions, therefore the time-bandwidth product of IMFCC filter is also 0.55968.

Therefore our proposed window provides 9.1% improved time-bandwidth product compared to MFCC filter and IMFCC filter. Hence proposed transform will perform better time-frequency analysis compared to MFCC and IMFCC filter bank in an allpass manner.

V. RV COEFFICIENT ANALYSIS

Let \( X \) be an \( n\times p \) feature matrix extracted from speech signal. Here each row is a feature vector which represents a point of a \( p \) dimensional vector space. Assume all the \( p \) features have been centered to have means equal to zero. The matrix \( S_x = XX^T / [\text{trace}(XX^T)]^{\frac{1}{2}} \) represents the sample configuration and it is independent of scaling. Let \( Y \) be another \( n\times q \) feature matrix extracted from the same speech signal. Here each row is a feature vector which represents a point in a \( q \) dimensional vector space. The corresponding sample configuration matrix is \( S_y = YY^T / [\text{trace}(YY^T)]^{\frac{1}{2}} \). Here \( \|S_x\|_2 = \|S_y\|_2 = 1 \).

Hence the distance between \( X \) and \( Y \) is given by,

\[
\|S_x - S_y\|_2 = \|S_x\|_2 + \|S_y\|_2 - 2 \langle S_x, S_y \rangle
\]

\[
= 2 - 2 \frac{\text{trace}(XX^T YY^T)}{[\text{trace}(XX^T)]^{\frac{1}{2}} [\text{trace}(YY^T)]^{\frac{1}{2}}}
\]

As \( \text{trace} \) is a commutative operation, we simplify the above expression as below,

\[
\|S_x - S_y\|_2 = 2 - 2 \frac{\text{trace}(C_{xx} C_{yy})}{[\text{trace}(C_{xx}) [\text{trace}(C_{yy})]^{\frac{1}{2}}}
\]

In the TESBCC feature calculation, first frequency analysis is done by taking the inner product of speech signal with the impulse responses of the FIR filters. The output of each filter is then passed through a framing operation, to get a limited time zone for time analysis. Unlike MFCC or IMFCC this TESBCC technique does not need any overlapping frames to handle the quasi periodicity of the speech signal because it belongs to frequency-time analysis approach. Here the last point of a frame and first point of the next frame represent two positions of speech signal which are shifted by one sample point, therefore in the frequency-time analysis approach the impulse response of an FIR filter acts as an overlapping window with maximum possible overlap (i.e. window length minus one). Hence frequency-time analysis approach is very efficient for quasi periodic signals and here optimal time-frequency resolution is also possible.
Here $C_{XX} = (X'X)/(n-1)$ and $C_{YY} = (Y'Y)/(n-1)$ are the covariance matrices of feature matrix $X$ and $Y$ respectively. $C_{XY} = (X'Y)/(n-1)$ and $C_{YX} = (Y'X)/(n-1)$ represent the cross covariance matrix between $X$ and $Y$. The quantity $\frac{\text{trace}(C_{XX}C_{YY})}{\text{trace}(C_{XX})\text{trace}(C_{YY})} = \frac{||C_{XX}||^2}{||C_{XX}|| ||C_{YY}||}$ is called the RV coefficient which is always positive, as norm is always a positive quantity. It is easy to verify that $RV(X,Y)\in [0,1]$. The RV coefficient is a measure of similarity, between $n \times p$ dimensional and $n \times q$ dimensional feature matrices extracted from the same speech signal. This can be seen as a generalized correlation coefficient between the two feature sets. When the value of RV coefficient is very high the two feature sets are highly correlated. RV coefficient is zero if and only if $C_{XX}$ is zero.

VI. EXPERIMENTAL EVALUATION

We have used the standard database POLYCOST for experimental evaluation of the performances of the MFCC, IMFCC and TESBCC feature sets. We have used 20 filters. The first coefficient is discarded since it contains only the energy of the spectrum and the resulting 19 dimensional vector is used.

A. Database Description

The POLYCOST database [14] was recorded as a common initiative within the COST 250 action during January-March 1996. It contains around 10 sessions recorded by 134 subjects from 14 countries. Each session consists of 14 items. The database was collected through the European telephone network. The recording has been performed with ISDN cards on two XTL SUN platforms with an 8 kHz sampling rate. Four speakers (M042, M045, M058 and F035) are not included in our experiments as they provide sessions which are lower than 6. All speakers (130 after deletion of four speakers) in the database were registered as clients. For training the speaker model, we have concatenated the speech form first five sessions and used to train GMM model. All the data for each speaker from session six to last available session for an individual speaker were used for testing. A total eleven hours of data were put under test.

B. Comparison through GMM Classifier

We took three types of model order (8 components GMM, 16 components GMM and 32 components GMM). In all the cases, we have used diagonal covariance matrix. Training speech length was 90 seconds. To test the model, we took two different types of test segments which were 1 second and 5 seconds long.

C. Measurement of Correlation using RV Coefficient

We have calculated the RV coefficient between feature sets for all 130 speakers. We consider the average value for easy understanding. Results are given in TABLE I.

<table>
<thead>
<tr>
<th></th>
<th>IMFCC</th>
<th>TESBCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>.54</td>
<td>.0025</td>
</tr>
<tr>
<td>IMFCC</td>
<td>.0024</td>
<td></td>
</tr>
</tbody>
</table>

D. Experimental Results using GMM Classifier

The identification accuracies for MFCC, IMFCC and TESBCC feature sets for 1 second test speech length are given in TABLE II. The identification accuracies after score level fusion of the above feature sets are given in the TABLE III. We have used equal weight sum rule for score level fusion.

<table>
<thead>
<tr>
<th>GMM Model Order</th>
<th>MFCC</th>
<th>IMFCC</th>
<th>TESBCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>63.09</td>
<td>53.24</td>
<td>71.64</td>
</tr>
<tr>
<td>16</td>
<td>68.31</td>
<td>58.93</td>
<td>77.75</td>
</tr>
<tr>
<td>32</td>
<td>71.06</td>
<td>62.02</td>
<td>80.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GMM Model Order</th>
<th>MFCC and IMFCC</th>
<th>MFCC and TESBCC</th>
<th>TESBCC and IMFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>66.68</td>
<td>81.07</td>
<td>82.39</td>
</tr>
<tr>
<td>16</td>
<td>70.93</td>
<td>84.53</td>
<td>86.36</td>
</tr>
<tr>
<td>32</td>
<td>73.38</td>
<td>85.80</td>
<td>87.94</td>
</tr>
</tbody>
</table>

The identification accuracies of the above feature sets for 5 seconds test speech length are given in the TABLE IV. The results after score level fusion are given in the TABLE V.

<table>
<thead>
<tr>
<th>GMM Model Order</th>
<th>MFCC</th>
<th>IMFCC</th>
<th>TESBCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>76.91</td>
<td>76.26</td>
<td>87.05</td>
</tr>
<tr>
<td>16</td>
<td>79.58</td>
<td>78.73</td>
<td>89.35</td>
</tr>
<tr>
<td>32</td>
<td>81.06</td>
<td>80.46</td>
<td>90.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GMM Model Order</th>
<th>MFCC and IMFCC</th>
<th>MFCC and TESBCC</th>
<th>TESBCC and IMFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>79.49</td>
<td>87.98</td>
<td>91.08</td>
</tr>
<tr>
<td>16</td>
<td>81.62</td>
<td>89.16</td>
<td>92.03</td>
</tr>
<tr>
<td>32</td>
<td>82.48</td>
<td>89.43</td>
<td>92.80</td>
</tr>
</tbody>
</table>

From the above five tables (TABLE I to TABLE V) we have the following critical observations.
From TABLE I it is evident that the TESBCC feature set is practically uncorrelated with MFCC and IMFCC feature sets.

From TABLE II and TABLE IV it is clear that, the identification accuracy of the TESBCC feature set is significantly higher compared to MFCC and IMFCC feature sets. The average improvement in the identification accuracy using TESBCC feature set is 9.5% compared to the MFCC feature set.

Again, from TABLE II and TABLE IV it is evident that, for small test segments (i.e. 1 second test speech) the identification accuracy of the IMFCC feature set is poor compared to the MFCC feature set. But for large test segments (i.e. 5 second test speech) the identification accuracy of the IMFCC feature set is comparable with the MFCC feature set.

The results after fusion of MFCC and IMFCC feature sets are higher compared to the accuracy of MFCC alone. This is true for both test segment lengths.

For small test segments the accuracy is higher after fusion of MFCC and TESBCC feature sets. But for large test segment there is no improvement after fusion.

The results, after fusion of TESBCC and IMFCC feature sets, are higher compared to the accuracy of TESBCC alone. For small test segments the improvement is significantly higher.

VII. CONCLUSION

This paper is an extended version of our previous works where we proposed IMFCC feature set and a new feature set based on the frequency-time analysis approach called TESBCC using the cosine modulated filter bank of a newly designed Nyquist filter for text-independent speaker identification. In this paper we have three main conclusions.

The feature set based on the frequency-time analysis approach, TESBCC is practically uncorrelated with the MFCC and IMFCC feature sets which are based on the time-frequency analysis approach.

The identification accuracy of TESBCC feature set is far better compared to the MFCC and IMFCC feature sets.

The IMFCC feature set is important for fusion.

REFERENCES


Figure 6. Block diagram of the proposed TESBCC feature extraction technique using frequency-time analysis approach.