AUDITORY SPECTRUM BASED FEATURES (ASBF) FOR ROBUST SPEECH RECOGNITION

Chi H. Yim, Oscar C. Au, Wanggen Wan, Cyan L. Keung, Carrson C. Fung

Human Language Technology Center
Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
Email: eeyim@ust.hk, eeau@ust.hk, wanwg@hotmail.com, keunglui@ust.hk, c.fung@ieee.org

ABSTRACT
MFCC are features commonly used in speech recognition systems today. The recognition accuracy of systems using MFCC is known to be high in clean speech environment, but it drops greatly in noisy environment. In this paper, we propose new features called the auditory spectrum based features (ASBF) that are based on the cochlear model of the human auditory system. These new features can track the formants and the selection scheme of these features is based on the second order difference cochlear model and the primary auditory nerve processing model. In our experiment, the performance of MFCC and the ASBF are compared in clean and noisy environments. The results suggest that the ASBF are much more robust to noise than MFCC.

1. INTRODUCTION
Some common acoustic features for speech recognition are mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC). These cepstral coefficients are well known to be sensitive to additive noise [1]. In the case of additive white noise, the speech signal is affected evenly and uniformly at all frequencies regardless of the signal strength distribution in the frequency domain. With constant noise power at all frequency, the signal-to-noise ratio (SNR) is higher at the frequency components with strong signal strength than at those with lower signal strength. If only the frequency components with strong signal strength are used in the recognition, the recognition scheme might perform better in noisy situations. The formants are one type of features that have high signal strength.

In this paper, we propose to perform robust speech recognition using formant-related features. These features are called auditory spectrum based features (ASBF). Using our feature extraction algorithm, the extracted features would track the formant positions within a speech frame. The features are based on the cochlear model of the human auditory system. The cochlear model is applicable to automatic speech recognition [2]. Our simulation results suggest that reasonably good and robust recognition is possible using ASBF.

2. ALGORITHM OVERVIEW
There are basically six main processing steps in the algorithm to extract the new features. A block diagram shown in Figure 1 illustrates the general structure of the algorithm.

Figure 1: ASBF Extraction Algorithm Overview

2.1. Endpoint Detection and Pre-emphasis
The endpoint detector in [3] is built for the pre-processing. We observed that a good endpoint detection of high accuracy is important to our systems. Otherwise the extracted features may be inconsistent. Pre-emphasis is implemented using Hamming window with pre-emphasis coefficient of value from 0.95 to 1. Each frame is then Fourier transformed into the frequency domain for the next step in which a bank of cochlear filters is applied.

2.2. Second order difference cochlear model (SDCM)
After each frame is transformed by FFT, it is passed to the SDCM cochlear filter banks whose frequency responses are shown in Figure 2. The second order difference cochlear model (SDCM) is a mathematical model of the basilar membrane (BM) [5]. It was applied in [4] for low-bit-rate speech coding. The SDCM is composed of a bank of filters to mimic the cochlear in the human auditory system. The frequency responses in Figure 2 reflect the human response to auditory stimulus.

The mathematical model of SDCM is [5]:

\[
\text{SDCM}(f) = \frac{1}{2} \left( 1 + \cos(2\pi f T) \right)
\]

where \(T\) is the characteristic frequency of the filter and \(f\) is the frequency of the input signal. The mathematical model of SDCM is used to mimic the frequency response of the human cochlea.

Figure 2 shows the frequency responses of the SDCM filters. The filters are designed to have a peak at the characteristic frequency of the filters, which is defined as the frequency at which the filter response is maximum. The filters are also designed to have a low-pass characteristic, which means that the frequency response of the filter decreases with increasing frequency. The filters are used to extract the formant frequencies from the speech signal. The formant frequencies are the frequencies at which the filter response is maximum. The formant frequencies are used to represent the speech signal in a more compact form.
The corresponding transfer function is [5]:

\[ H_i(z) = A_i a_n (1 - z^{-1}) / (1 + b_1 z^{-1} + b_2 z^{-2}) \]  

In Eqn. (1), \( u(n) \) is the stape’s velocity, \( y(n) \) is the BM displacement in position \( x_k \), parameters \( b_{1k}, b_{2k}, A_k \) and \( a_{nk} \) are coefficients with respect to position \( x_k \) along the BM. Eqn. (2) is the transfer function form of Eqn. (1).

2.3. Primary auditory nerve processing model (PANPM)

After the cochlear filter banks, this stage determines the dominating frequencies of the auditory input. The primary auditory nerve processing model (PANPM) is a model to simulate the selection process of frequency components from a speech signal by the auditory nerves. The PANPM stimulates the strength effect of each auditory nerve and ensemble effect of all the auditory nerves, and it selects the most significant frequency components according to the following mechanism as illustrated in Figure 3.

Each speech frame is filtered by each of the \( N \) cochlear filters. For each cochlear filter, the output has a FFT which is the product of the FFT of the filter response and the FFT of the input signal and this signal has different values at different frequencies in general. Which frequency components are larger than the others would depend on the input signal spectrum and on which cochlear filter it is. The frequency components that generate the largest output power in each filter is identified and is considered as the dominant frequency. A counter is set up for each frequency to count the number of times the particular frequency being one of the dominant frequencies in the \( N \) cochlear filters. Some frequency components with very small signal strength would have counts of zero, while those around the formant frequencies would tend to have a large count. When the count of a particular frequency exceeds a threshold \( T_{cf} \), the frequency is extracted as one of the features.

In human auditory perception, a frequency component with high energy can mask off neighboring frequency components with energy lower than a certain profile. Such perceptual masking effects are used heavily in audio coding. In our case, when the extracted frequencies are very closed to each other, redundant frequencies are eliminated to account for the masking effects. The remaining features are called the auditory spectrum based features (ASBF).

2.5. Smoothing

To model the slowly-varying articulation effect of voice production mechanism, a smoothing function is applied to smooth the feature fluctuations. The smoothed frequency value is computed by taking a linear combination of the nearest extracted frequency in previous frame with respect to the position of the current feature, extracted frequency in current frame and a dc offset frequency.

\[ x(n) = c_0 x(n-1) + c_1 y(n) + c_2 d(n) \]  

In Eqn. (3), \( n \) is the time index, \( x(n) \) is the smoothed feature, \( y(n) \) is the original extracted feature, \( d(n) \) is the dc offset frequency for \( x(n) \), and \( c_0, c_1 \) and \( c_2 \) are
coefficients for averaging the terms $x(n-1)$, $y(n)$ and $dc(n)$ such that the sum of the three coefficients is unity. A simplified idea of choosing the value of the dc offset frequencies is by considering the features coming from different subbands of equal bandwidth and the dc offset frequency is the center frequency in each subband. The subband bandwidth can also be determined in Mel scale. To illustrate the idea of our algorithm discussed above, an example is shown in Figure 4.

The example is a speech frame from the English word “Two” and the extracted features in frequency domain are indicated by the vertical lines. These are the dominant frequencies whose dominance counts exceed a threshold. These are also frequency values at which the local maxima in the frequency spectrum occurs. The upper figure shows that 8 features are extracted before frequency masking. Since some extracted features are very near to the others, frequency masking can remove them effectively. As a result, the number of features extracted are reduced.

3. RECOGNITION RESULTS
The performances of the ASBF and MFCC as recognition features are compared in clean and noisy environments. The experiments are conducted on the isolated digit database from TIDIGITS. There are 1210 utterances from 55 speakers for training data set and 1232 utterances from 56 speakers for testing data set. Each digit is modeled using a 6-state left-to-right HMM with 7 continuous Gaussian mixtures per state. The frame length is 10 ms with 5 ms overlapping between consecutive frames. In our experiment, we use 50 cochlear filters ($N=50$). White noise is added to all frames of the clean speech at signal-to-noise ratios (SNR) of 0dB, 5dB, 10dB, 15dB and 20dB. The MFCC feature vector has 39 dimensions, including 13 cepstral coefficients, 13 cepstral derivatives and 13 cepstral acceleration. In all experiments, the bandwidth is limited from 100 to 4.4 kHz for generating both ASBF and MFCC and the pre-emphasis coefficient is 0.97. The results are shown in Table 1 and Table 2. The number enclosed in brackets in Table 1 and Table 2 is the dimension of the feature vector being used.

### Table 1: Recognition accuracy on clean and noisy speech at SNR=20dB and SNR=15dB using various features

<table>
<thead>
<tr>
<th>Feature Vector (dimension)</th>
<th>Clean</th>
<th>Addition of white noise SNR=20dB</th>
<th>Addition of white noise SNR=15dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC (39)</td>
<td>97.65</td>
<td>82.06</td>
<td>71.51</td>
</tr>
<tr>
<td>ASBF (5)</td>
<td>94.40</td>
<td>87.18</td>
<td>81.49</td>
</tr>
<tr>
<td>ASBF (6)</td>
<td>93.83</td>
<td>84.09</td>
<td>79.63</td>
</tr>
<tr>
<td>ASBF (7)</td>
<td>96.75</td>
<td>88.31</td>
<td>82.31</td>
</tr>
<tr>
<td>ASBF (10)</td>
<td>93.75</td>
<td>85.06</td>
<td>76.95</td>
</tr>
<tr>
<td>ASBF (15)</td>
<td>93.75</td>
<td>82.47</td>
<td>77.35</td>
</tr>
</tbody>
</table>

### Table 2: Recognition accuracy on noisy speech at SNR=10dB, 5dB and 0dB using various features

<table>
<thead>
<tr>
<th>Feature Vector (dimension)</th>
<th>Addition of white noise SNR=10dB</th>
<th>Addition of white noise SNR=5dB</th>
<th>Addition of white noise SNR=0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC (39)</td>
<td>57.71</td>
<td>44.56</td>
<td>31.66</td>
</tr>
<tr>
<td>ASBF (5)</td>
<td>70.29</td>
<td>55.52</td>
<td>37.82</td>
</tr>
<tr>
<td>ASBF (6)</td>
<td>69.89</td>
<td>51.87</td>
<td>37.91</td>
</tr>
<tr>
<td>ASBF (7)</td>
<td>68.67</td>
<td>48.38</td>
<td>27.70</td>
</tr>
<tr>
<td>ASBF (10)</td>
<td>60.31</td>
<td>34.25</td>
<td>17.69</td>
</tr>
<tr>
<td>ASBF (15)</td>
<td>66.48</td>
<td>50.65</td>
<td>32.63</td>
</tr>
</tbody>
</table>

In Table 1, we can see that the recognition accuracy is higher for the 39-feature MFCC than for our proposed ASBF in clean speech conditions. In Table 1 and Table 2, when white noise is added to the speech signal, the results suggest that using ASBF as features can be significantly better than using MFCC from 5 dB to 20 dB except the case of using 10-feature ASBF. When the white noise level is at 0 dB, the recognition accuracy for 39-feature MFCC is higher than for 7-feature and 10-feature ASBF and is lower than for 5-feature, 6-feature, 15-feature ASBF.
To illustrate the performance in the recognition task using MFCC and ASBF, the results of using 39-feature MFCC and 5-feature ASBF are shown in Figure 5. When white noise is added to the speech signal, the results suggest that using ASBF as features can be significantly better than using MFCC. In clean environment, the traditional 39 dimension MFCC performs 3.4% better than our proposed ASBF. At 20 dB SNR, our proposed ASBF achieves 6.2% higher recognition accuracy than the MFCC with 39 features. At 15 dB SNR, our proposed ASBF achieves 14.0% higher recognition accuracy than the MFCC with 39 features. At 10 dB SNR, our proposed ASBF achieves 21.8% higher recognition accuracy than the MFCC with 39 features. At 5 dB SNR, our proposed ASBF achieves 24.6% higher recognition accuracy than the MFCC with 39 features. At 0 dB SNR, our proposed ASBF achieves 19.5% higher recognition accuracy than the MFCC with 39 features.

These simulation results suggest that, for the TIDIGIT recognition task, our proposed ASBF are robust features that can achieve slightly worse recognition accuracy than MFCC in clean speech, and significantly better recognition accuracy in noisy speech.

The results in Table 1 and Table 2 also suggest that when the noise level is high, our proposed features with fewer dimension perform better than those with more dimension. It suggests that some dominant features extracted are from noise and these features should not be taken into our proposed ASBF feature set. In particular ASBF with dimension from 15 to 4 is significantly better than MFCC with 39 features in noisy situations.

Because the values of $c_0$, $c_1$, and $c_2$ in Eqn.(3) are chosen to be the same for the smoothing function in all our experiments, different set of values may need to be applied to get a higher recognition accuracy using ASBF with smoothing than using ASBF without smoothing.

4. CONCLUSION

In this paper, sinusoidal representation and auditory model based features are proposed for robust speech recognition. Our experimental results show that our proposed features, the auditory spectrum based features ASBF, are quite robust to additive white noise in speech recognition. A small number of ASBF features can achieve significantly higher recognition rates than a much larger number of MFCC features. The smoothing and masking effect in our results need further work.

ACKNOWLEDGMENT

This work is funded in part by a CAG grant and an HKTIIT grant. We acknowledge the collaboration of Yuan Jingxian of Shanghai University and Wilson Tam Yik Cheung of Hong Kong University of Science and Technology.

5. REFERENCES