Workshop on Spoken Language Processing
An ISCA-Supported Event
Mumbai, India
January 9-11, 2003

A STUDY INTO FRONT-END SIGNAL PROCESSING FOR AUTOMATIC SPEECH RECOGNITION

Rohit Sinha and S. Umesh

Department of Electrical Engineering
Indian Institute of Technology
Kanpur, 208 016, INDIA
{srohit, sumesh}@iitk.ac.in

ABSTRACT

In our recently reported work, we have observed some difference in recognition performance using our proposed method of feature computation when compared to features computed using the traditional mel-filterbank analysis. In the alternate method of feature computation, we use a spectral smoothing procedure which is very similar to weighted overlapped segment averaging (WOSA) method of spectral estimation. In this paper, we study the signal processing of the above mentioned feature computation methods, and point out to the differences between the two methods, and the effect of these differences on the recognition performance.

Contact Address: S. Umesh
Room # 207-A, ACES Building
Department of Electrical Engineering
Indian Institute of Technology
Kanpur - 208 016
INDIA
Tel: +91-512-59-7855(Office)
+91-512-59-7846(Lab)
Fax: +91-512-59-0063

AN STUDY INTO FRONT-END SIGNAL PROCESSING FOR AUTOMATIC SPEECH RECOGNITION

Rohit Sinha and S. Umesh

Department of Electrical Engineering
Indian Institute of Technology
Kanpur, 208 016, INDIA
{srohit, sumesh}@iitk.ac.in

ABSTRACT

In our recently reported work, we have observed some difference in recognition performance using our proposed method of feature computation when compared to features computed using the traditional mel-filterbank analysis. In the alternate method of feature computation, we use a spectral smoothing procedure which is very similar to weighted overlapped segment averaging (WOSA) method of spectral estimation. In this paper, we study the signal processing of the above mentioned feature computation methods, and point out to the differences between the two methods, and the effect of these differences on the recognition performance.

1. INTRODUCTION

The Mel-frequency cepstral coefficients (MFCCs) are the most commonly used features in automatic speech recognition (ASR) systems. MFCCs are computed with filterbank consisting of overlapping triangular filters. In [1, 2], we have proposed an alternate approach to compute features using a smoothing technique which is very similar to the averaged periodogram method of spectral estimation which is also known as WOSA [3].

There are variations in the position of spectral peaks (formants) of a given sound among different speakers due to differences in their vocal tract lengths. Speaker normalization methods try to reduce this variability which causes significant performance degradation in speaker independent speech recognition systems.

A commonly used approximation is that for same sound the spectral envelopes between two speakers, $S_A(f)$ and $S_B(f)$ are linearly scaled versions of one another, i.e.,

$$S_A(f) \approx S_B(\alpha_{AB} f) \tag{1}$$

Using this approximation one of the most commonly used method of speaker normalization is based on the estimation of the linear scaling parameter, α_{AB}, in the maximum likelihood (ML) sense. The linear scaling of the frequency axis (for normalization) can be implemented by re-sampling the speech waveform in the time domain [4] or more efficiently by modifying the filter-bank for each scaling parameter in the mel-frequency cepstral coefficient (MFCC) feature analysis [5]. The estimation of frequency scaling parameter is done with respect to a reference speech model and we do not have access to the reference speech model at the outset. It is obtained by iterative procedure of estimating warping factors and updating speech models.

Recently, we have used the proposed WOSA based features on a digit recognition task with maximum-likelihood based speaker normalization and have observed some improvement for children when compared to the MFCC based features [6, 7]. We have also noticed that the recognition performance after normalization showed a decreasing trend with increasing iteration particularly for children with filterbank based MFCC features as shown in Table 1. On the other hand, no such trend was noticed in the recognition performance

Iteration Number	MFCC features		WOSA based features	
	Adults	Children	Adults	Children
Unnorm.	96.58	84.65	96.78	85.17
1	97.30	90.48	97.31	90.99
2	97.40	88.75	97.32	91.03
3	97.38	87.72	97.30	91.13

Table 1: *Word recognition accuracy (%) for mel-filterbank based MFCC features and WOSA based features for adults and children. The table also shows the recognition performance for* 3 *consecutive iterations of speaker normalization for models built on adult data.* (Table adapted from [7])

after normalization with WOSA based features for either adults or children.

In this paper, we present a study into the two methods signal processing to compute features and point out to the differences between the two methods and the effect of the differences in recognition performance.

The remainder of this paper is organized as follows. In the next section we review the traditional filterbank based MFCCs computation. This is followed by description of our recently proposed WOSA based feature computation. In Section 4, we provide insights into the two methods of front-end signal processing which help us to understand the differences between the two methods. Based on the understanding of differences we do some modifications in traditional filterbank based MFCC computation method which help reduce the performance difference. Finally we present recognition results based on telephone based connected digit recognition task, and conclude by discussing the effect on the performance caused by the variation in the front-end signal processing.

2. FILTERBANK BASED MFCC COMPUTATION

Figure 1 shows the various signal processing steps involved in the computation of filterbank based MFCCs. The speech waveform, sampled at 8 kHz, is first preemphasized and cut into a number of overlapping segments, each 20 ms long and shifted by 10 ms. A Hamming window is multiplied and Fourier transform (FFT) is computed for each frame.

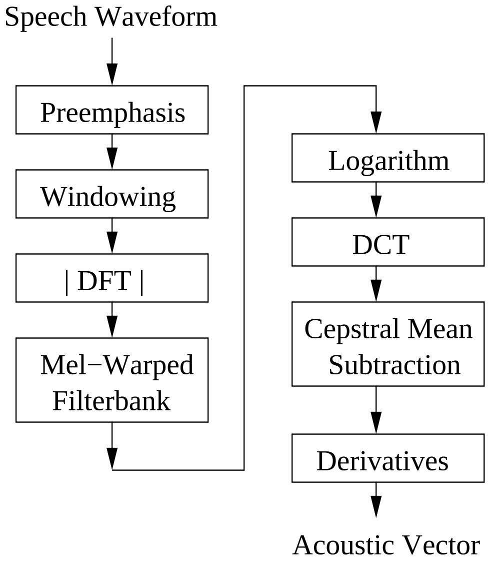

Figure 1: *Signal processing steps in traditional filterbank based MFCC computation.*

The magnitude spectra is then multiplied with a Mel-warped filterbank consisting of overlapping triangular filters. The Mel-warping is done to adapt the frequency resolution to the properties of the human ear and the filterbank segments the spectrum into a number of critical bands. A discrete cosine transform (DCT) is applied to the logarithm of the filterbank outputs and 12 cepstral coefficients ($C_1 \ldots C_{12}$) along with normalized logarithm of signal energy are used as the raw MFCC vector The mean of each cepstral component is subtracted. Finally, the MFCC vector is augmented with first and second derivatives.

In this paper, we have used 21 channel filterbank spaced uniformly on Mel-scale between 200-3452 Hz.

3. PROPOSED WOSA BASED FEATURE COMPUTATION

In speech recognition we are basically interested in modeling only the spectral envelope of the speech signal and hence we would like to reduce the effect of pitch excitations. In traditional filterbank based MFCC computation the pitch smoothing is effected by the averaging of triangular shaped filterbank over a number of DFT coefficients.

The signal processing steps involved in our recently proposed method to compute MFCCs are shown in Figure 2. On comparison with Figure 1, one can easily see that our proposed feature computation method differs from filterbank based MFCC computation method *only* in the procedure used for smoothing the pitch excitations. We have used a procedure to smooth pitch excitation which has a marked similarity with WOSA. WOSA is a widely used technique of spectral estimation based on the time averaging of periodograms [3].

In our smoothing procedure, each 20 msec long preemphasized frame of speech (which corresponds to 160 samples for 8 kHz sampling) is segmented into 6 *overlapping* sub-frames. Each sub-frame is chosen to be 64 samples long and the overlap between sub-frame is 45 samples. The subframes are then Hamming windowed and sample autocorrelation estimates are computed. An averaged autocorrelation estimate is obtained by averaging over the available 6 autocorrelation estimates. The Fourier transform of the averaged autocorrelation estimate is essentially the *smoothed* spectral envelope.

Pitch is effectively suppressed since duration of each sub-frame is less than the expected pitch interval of an average adult male. For every subframe that does contain an individual pitch-pulse there is a broadband energy contribution to the spectrum of that sub-frame but not to that of any other subframe. The result is that the averaged spectrum contains all of the formant structure but almost none of the pitch structure.

The choice of subframe length of 64 is suitable only for the adult male speakers so we are presently investigating into the effect of different subframe lengths on the recognition performance for both adults and children.

For the discrete implementation of Mel-warping we compute non-uniformally spaced DFT of the averaged autocorrelation estimates. The nonuniform DFT is computed on 21 frequencies that are exactly *same* as the 21 center frequencies of the mel-warped filterbank in the traditional MFCC computation method. To obtain cepstral coefficients we compute DCT of the logarithm of melwarped power spectra and 12 cepstral coefficients $(C_1 \dots C_{12})$ along with normalized logarithm of signal energy are used as the raw feature vector.

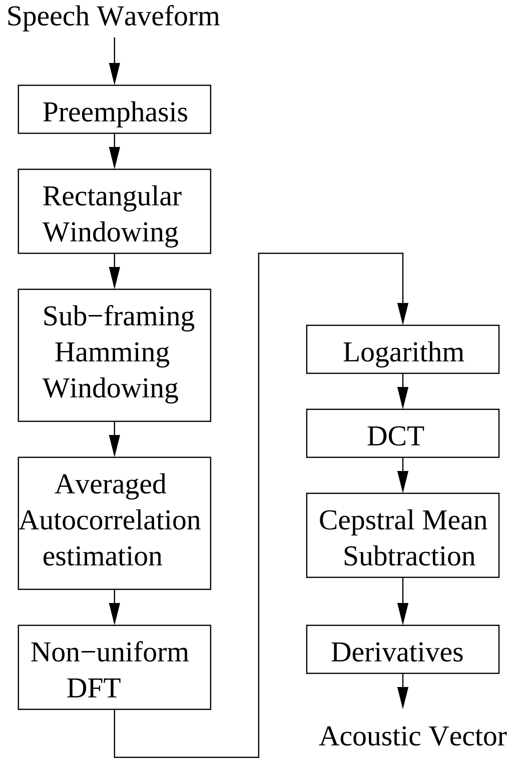

Figure 2: *Signal processing steps in our proposed WOSA based feature computation.*

Similar to the filterbank based MFCC feature vector, in this case also cepstral mean subtraction is performed and the feature vector is augmented with first and second derivatives. Thus the resulting feature vector dimension is same as that of filterbank based MFCC feature vector.

4. RELATIONSHIP BETWEEN THE TWO METHODS OF FEATURE COMPUTATION

In this section, we provide some insights into the two method of feature computation so as to find some relationship between the two.

In filterbank based MFCC computation method, the purpose of using filterbank is to estimate the smoothed magnitude spectra which is free from the effect of pitch excitation, apart from dividing the spectra into a number of critical bands. Since the triangle shaped overlapping filters are used as shown in Figure 3 on the top, so the smoothing increases with increasing center frequency in the filterbank.

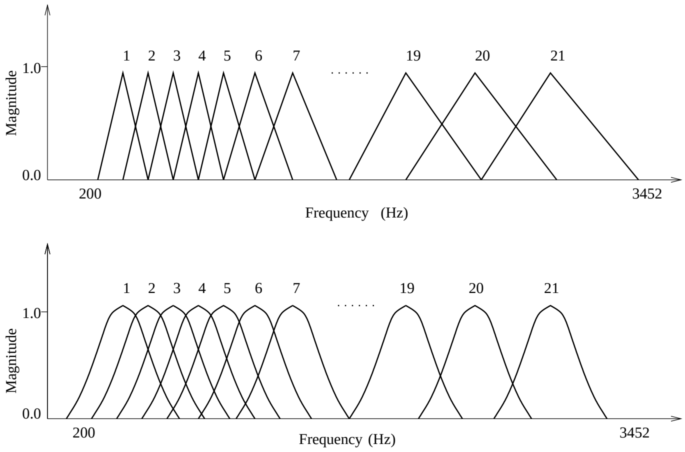

Figure 3: *Filterbank with triangular shape filter used in MFCC computation (shown on top) and filterbank with filters of uniform bandwidth and having frequency response equal to the square of Fourier transform of Hamming window used in proposed WOSA based feature computation (shown on bottom).*

In the WOSA based feature computation the smoothed power spectra is computed using WOSA method. The WOSA method can also be given a filtering interpretation which enables us to find some relationship between WOSA based and filterbank based MFCC computation.

From Nuttall and Carter [3], in WOSA method, the relationship between power spectra $\hat{G}_{av}(f)$ computed by Fourier transform of averaged autocorrelation estimate and the true power power spectra $G(f)$ of the speech frame can be expressed as follows,

$$\hat{G}_{av}(f) = G(f) \circledast \{\mathcal{F}[w(t)]\}^2 \qquad (2)$$

where '$\circledast$' denotes convolution and $\mathcal{F}[w(t)]$ denotes Fourier transform of Hamming window which is used on each sub-frame.

Thus the estimated average power spectra is the convolution of the true power spectra with the square of Fourier transform of Hamming window.

Therefore, WOSA method estimates the smooth power spectra at any frequency by bandpass filtering the true power spectrum at that frequency where the bandpass filter has the frequency response equal to the square of Fourier transform of Hamming window.

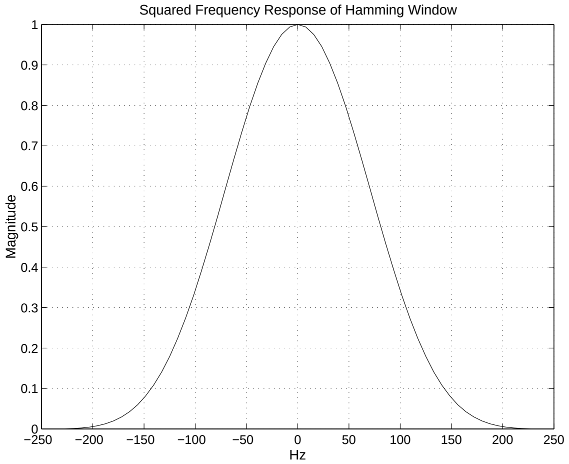

Figure 4: *Frequency response of the square of Fourier transform of* 64 *point Hamming window.*

In order to compute MFCCs, we have computed the power spectra at frequencies spaced on non-uniform (Mel) scale so it can be argued that WOSA based feature also uses a filterbank, similar to filterbank based MFCC computation, except the constituent filters are of *uniform bandwidth* and have frequency response equal to the *square of Fourier transform of Hamming window* as shown in Figure 3 on the bottom.

From the filtering interpretation of WOSA, it is clear that the following are the main differences in the filterbank and WOSA signal processing steps:

- WOSA results in estimate of smooth power spectra while in MFCC filterbank smooth magnitude spectra is commonly estimated.
- The shape and bandwidth of the filters used in two methods are not identical.

In WOSA based method, the filters of argued filterbank are of constant bandwidth with frequency response equal to the square of Fourier transform of Hamming window. On the other hand, the filterbank based MFCC computation method uses triangular shaped filters and the bandwidths of the filters are not constant and increase with increasing center frequencies.

5. MODIFICATIONS IN FILTERBANK BASED MFCC COMPUTATION

As discussed in previous section that the main differences in signal processing of the two feature computation method are in the shape and bandwidth of filters and the use of power spectra. Motivated by these we tried following modifications in the filterbank based MFCC computation:

- Power spectra instead of magnitude spectra is used while filterbank remained unmodified.
- Power spectra is used and the filters of MFCC filterbank are modified such that each filter is now of constant bandwidth approximately equal to that of WOSA filter.

In the computation of WOSA based feature, we have segmented each speech frame into subframes of length 64 samples so the frequency response of the WOSA filter will be equal to the square of Fourier transform of 64 point Hamming window and is shown in Figure 4. The width of the main lobe of the WOSA filter is around 350 Hz.

So the triangular filters, in the traditional filterbank based MFCC computation , are modified to have a constant bandwidth of 350 Hz. Since the shape of WOSA filter is not strictly triangular, we have also experimented with varying bandwidths around 350 Hz.

6. RECOGNITION TESTS

The comparison of the MFCC computation methods is done on telephone based connected digit recognition task using HTK speech recognition toolkit. The acoustic training data is drawn from Numbers v1.1 corpus of Oregon Graduate Institute. The training set contained 6078 utterances totalling 33420 digits from adult male and female speakers. Two different testing sets are used to assess the performance of the features. The first is a *matched* test set called “Adults”, consisting of 2169 utterances totalling 12347 digits from adult male and female speakers. The other is a *mismatched* test set (where the age of speakers in the test set is very different from training set) called “Children”. The Children test set consists of 2798 utterances totaling 9974 words from predominantly children between age group of six to eighteen years recorded over telephone.

Iteration Number	MFCC with Power Spectra	
	Adults	Children
Unnorm.	96.81	84.78
1	97.38	90.82
2	97.47	89.07
3	97.54	88.00

Table 2: *Word recognition accuracy (%) for filterbank based MFCC features using power spectra instead of magnitude spectra. The table also shows the recognition performance for* 3 *consecutive iterations of speaker normalization for models built on adult data.*

The digits are modeled with a word model. Continuous density left-to-right HMM's with 16 states and 5 multi-variate Gaussian mixtures/state with diagonal covariance matrices are used to model the digits. The silence is modeled separately with 3 states left-to-right CDHMM's with 6 mix./state along with a short pause model of single state which is tied with the middle state of silence model. 39-dimensional feature vectors are used: normalized energy, C_1-C_{12} and their first and second order derivatives. Finally cepstral features are liftered and cepstral mean subtraction is also performed.

Table 2 provides the recognition results for filterbank based MFCC computation using power spectra instead of magnitude spectra. Comparing with Table 1, we find that recognition performance for both adults and children are enhanced by the use of power spectra but the trend of decrease performance with increasing iteration for children is still present.

We then experiment with modified filterbank based MFCC computation in which individual filters are now having a constant bandwidth along with using power spectra. Table 3 provides the recognition performance for filterbank based MFCC computation with constant bandwidth filters of varying bandwidths. The performance of filterbank based for bandwidth of 350 Hz is very close to that of WOSA based MFCC features for chil-

Bandwidth (Hz)	Modified MFCC, Const. BW	
	Adults	Children
200	96.58	85.54
250	96.81	85.57
300	96.68	85.48
350	96.66	85.05
400	96.66	84.60

Table 3: *Word recognition accuracy (%) for modified filterbank based MFCC features using constant bandwidth (BW) filters and using power spectra for adults and children. (with no speaker normalization)*

Iter. #	Modified MFCC, BW-250Hz	
	Adults	Children
Unnorm.	96.81	85.57
1	97.39	91.29
2	97.41	91.51
3	97.38	91.47

Table 4: *Word recognition accuracy (%) for modified filterbank based MFCC features with filters having constant bandwidth (BW) of* 250 *Hz. The table also shows the recognition performance for* 3 *consecutive iterations of speaker normalization for models built on adult data.*

dren while the best performance is observed for bandwidth of 250 Hz.

In the next experiment we have evaluated the performance of modified filterbank based MFCC computation for speaker normalization. The details of the speaker normalization procedure used are same as described in [7]. Table 4 provides the recognition performance for speaker normalization for modified filterbank based MFCC computation with filters having a constant bandwidth of 250 Hz and using power spectra.

Comparing with Table 2, we can see that the modification has provided a statistically significant improvement in the recognition performance for children over the traditional filterbank using power spectra and also there is no peculiar trend of decreasing performance with increasing iteration for children. This improvement is obtained with no significant affect on the recognition performance for adults.

7. DISCUSSION

In this paper, we have presented a relationship between traditional filterbank based MFCC feature and our recently proposed WOSA based feature. Based on this relationship we have performed some modification in traditional filterbank based MFCC computation. Our preliminary experimental results indicate that the shape and bandwidth of filter have an effect on the recognition performance which needs further investigation.

REFERENCES

[1] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-Warping in Speech," in *Proc. of ICSLP'96*, Philadelphia,USA, 1996.

[2] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale Transform in Speech Analysis," *IEEE Transactions on Speech and Audio Processing*, January 1999.

[3] A. H. Nuttall and G. C. Carter, "Spectral Estimation using Combined Time and Lag Weighting," *Proceedings of the IEEE*, vol. 70, pp. 1115–1125, Sept. 1982.

[4] A. Andreou, T. Kamm, and J. Cohen, "Experiments in Vocal Tract Normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[5] Li Lee and Richard Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 49–59, Jan. 1998.

[6] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," in *Proc. of IEEE ICASSP'02*, May 2002, vol. 1, pp. 589–592.

[7] Rohit Sinha and S. Umesh, "Investigation into Frequency Warping and Spectral Smoothing for Vocal Tract Length Normalization," Submitted to NCC'2003.