Comparison of Spectral Analysis Methods for Automatic Speech Recognition

Venkata Neelima Parinam, Chandra Vootkuri, Stephen A. Zahorian

Department of Electrical and Computer Engineering
Binghamton University, Binghamton, NY 13902, USA
{vparina1, cvootku1, zahorian}@binghamton.edu

Abstract

In this paper, we evaluate the front-end of Automatic Speech Recognition (ASR) systems, with respect to different types of spectral processing methods that are extensively used. Experimentally, we show that direct use of FFT spectral values is just as effective as using either Mel or Gammatone filter banks, as an intermediate processing stage, if the cosine basis vectors used for dimensionality reduction are appropriately modified. Furthermore it is shown that trajectory features computed over intervals of approximately 300ms are considerably more effective, in terms of ASR accuracy, than are delta and delta-delta terms often used for ASR. Although there is no major performance disadvantage if a filter bank is used, simplicity of analysis is a reason to eliminate this step in speech processing. The experimental results which confirm the above assertions are based on the TIMIT phonetically labeled database. The assertions hold for both clean and noisy speech.

Index Terms: DCTC/DCSC, MFCC, Gammatone filter bank, Mel filter bank, ASR.

1. Introduction

All automatic speech recognizers perform spectral analysis at the front end which converts the speech signal, possibly noisy and/or degraded, into values from which useful features can be easily computed. The front end spectral analysis is performed by calculating the short time Fourier transform (STFT) of the speech signal, either using an FFT, a filter bank, or a combination of the two methods. For the combination method, the filter bank is approximated by summing weighted combinations of FFT magnitude values. The filter bank approach, even if derived from FFT values, is thought to be advantageous since it can be designed to mimic the functionality of the cochlea of the human auditory system, such as a nonlinear ("warped") frequency scale.

The majority of ASR systems are implemented using a Mel filter bank as the spectral analysis front end, followed by a cosine transform based feature extraction which is shown to outperform other signal processing methods [1]. Very recently, another filter bank has been presented as a superior alternative to the triangular-shaped Mel filters called the Gammatone filter bank, which simulates the motion of the basilar membrane within the cochlea of the human auditory system. It was first introduced by Johannsma (1972) to describe the shape of the impulse response function of the auditory system as estimated by the reverse correlation function of neural firing times. The general thinking is that since the Gammatone filter bank approximates the human auditory system better than the Mel filter bank, it should also be superior for ASR applications [2].

The Gammatone filter is defined in the time domain (impulse response function) as:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \emptyset) \quad (1)$$

where f is the frequency, $\emptyset$ is the phase of the carrier, a is the amplitude, n is the filter order, b is the bandwidth and t is time.

Front-end spectral analysis can also be performed without using any filter bank, but simply using an FFT directly. In either case, spectral values (that is FFT values or filter bank outputs, both converted to log magnitudes), are typically reduced in dimensionality using some type of cosine transform. If the filter bank step is used, cosine basis vectors can be used directly. However, if the FFT magnitudes are used as the direct input to the cosine transform, the cosine basis vectors should be modified to account for the non-uniform frequency resolution. In order to incorporate spectral trajectory information into ASR feature sets, additional terms are generally computed from blocks of frame-based features, such as delta terms.

In the following sections we compare spectral features computed as cosine transforms of filter bank outputs with features computed as modified cosine transforms (DCTCs) of FFT spectral log magnitudes directly. We also compare delta type trajectory features with trajectory features computed over much longer time intervals using another set of modified cosine basis vectors (DCSCs). More details of the more common spectral and feature calculation method (MFCCs with delta and delta-delta terms are given in [3] and [4]. More details of the DCTC/DCSC general method are given in [5], [6] and [12]. All the methods are evaluated using as much similarity of parameters and recognizer as feasible (such as frequency range, # of HMM mixtures, etc.) in order to make comparisons most meaningful.

2. FFT Based Spectral Analysis

The most common spectral analysis method for speech recognition uses a frame-based approach in which the time varying speech signal is described by a stream of feature vectors, with each vector reflecting the spectral magnitude properties of a relatively short (10-30ms) segment (frame) of the signal. For experimental results reported in this paper, 16 kHz sampling rate speech signals are short-time Fourier transform (STFT) analyzed using a 10ms Kaiser window with a frame space of 2ms. The spectrogram of a typical speech signal is as shown in Figure 1. The FFT spectral values are used as the front-end for DCTC/DCSC feature extraction, as described later. The frame length and frame spacing mentioned were empirically determined as providing most accurate ASR results.

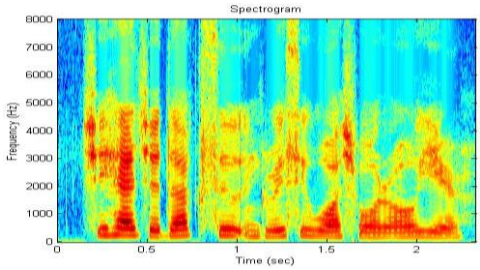

Figure 1: FFT spectrogram

3. Filter bank based Spectral analysis

A filter bank can be regarded as a crude model for the initial stages of transduction in the human auditory system. A set of band pass filters is designed so that the desired range of the speech band is entirely covered by the combined pass bands of the filters composing the filter bank. The output of the band pass filters are considered to be the time varying spectral representation of the speech signal.

For the experiments given in this paper, we evaluate two commonly used filter banks: the Mel filter bank and Gammatone filter bank. Either the DCTC/DCSC method (but without frequency warping) or the more common method used for MFCC features (i.e., delta terms rather than DCSCs) are used. Results are compared for the filter bank approaches versus the FFT-only spectral method.

3.1. Mel filter bank

The Mel filter bank is a series of triangular band pass filters, as depicted in Figure 2, designed to simulate the band pass filtering believed to be similar to that occurring in the auditory system.

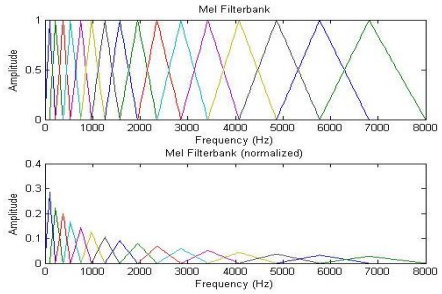

Figure 2: Frequency response of 16 channel Mel filter bank and the normalized versions of the filters, as used for MFCCs.

To convert the frequency in Hz into frequency in Mels the following equation is used:

$$m = 1127.01048 * log_e\left(1 + \frac{f}{700}\right) \quad (2)$$

On a linear frequency scale, the filter spacing is approximately linear up to 1000 Hz and approximately logarithmic at higher frequencies. For actual implementation, the Mel filter bank is computed by first computing the power spectrum with an FFT, and then multiplying the power spectrum by the Mel filter bank coefficients. In Figure 3 is shown a spectrogram based on 32 Mel filters. Note that this spectrogram is qualitatively similar to the direct FFT spectrogram shown in Figure 1. The details of the two spectrograms are quite different since the frequency range is more quantized in Figure 3 and the frequency scale is effectively in Mels rather than linear. However, it should be noted that the Mel spectrogram, or Mel filters, are derived from the FFT spectral values and thus are simply an intermediate step in processing.

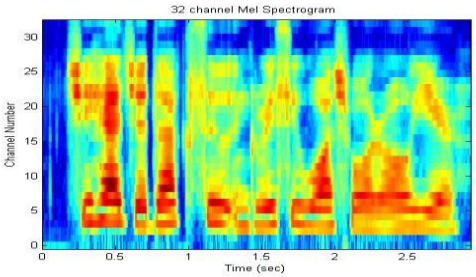

Figure 3: 32 channel Mel spectrogram

3.2. Gammatone Filter Bank

A Gammatone filter is a linear filter with impulse response described as the product of a (gamma) distribution and sinusoidal (tone), hence the name Gammatone. The filter bank is a combination of individual Gammatone filters with varying bandwidth based on the Equivalent Rectangular Bandwidth (ERB) scale. For moderate sound pressure levels, Moore et al [7] [8] estimated the size of ERBs for humans as:

$$ERB[f] = 24.7 + 0.108 * f_c \quad (3)$$

The value ERB[f] is used as the unit of center frequency f_c on the ERB scale. For example, the value of ERB[f] for a center frequency of 1 kHz is about 132.64, so an increase in frequency from 975 to 985 Hz represents a step of one ERB[f]. Each step in ERB roughly corresponds to a constant distance of about 0.89 mm on the basilar membrane [9].

As the center frequency increases the bandwidth of the filter bank increases. A 16 channel Gammatone FFT based filter bank frequency response is shown in Figure 4.

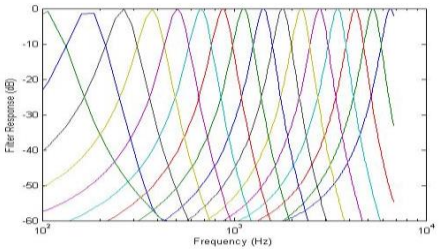

Figure 4: Frequency response of 16 channel Gammatone filter bank

The Gammatone filter bank can be implemented using sums of weighted FFT power spectrum values [10], exactly as for the Mel filter bank except using the weights corresponding to Figure 4, rather than the Mel filter weights shown in Figure 2. Alternatively, the Gammatone real filters can be implemented as actual IIR or FIR filters, followed by rectification and low pass filters, as depicted in Figure 5. Figure 6 depicts the Gammatone spectrogram of the same sentence as was used to construct the spectrograms for Figures 1 and 3.

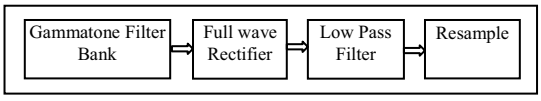

Figure 5: Block diagram of Gammatone using actual filters (difference equations) in first block

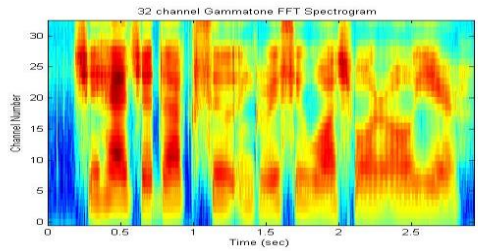

Figure 6: 32 channel Gammatone spectrogram

4. DCTCs/ DCSCs based feature extraction

Typically FFT spectral magnitudes or filter bank outputs are dimensionality reduced with a cosine or cosine-like transform for each frame of spectral values. Several frames of cosine transform coefficients are further processed in overlapping sliding blocks to form spectral trajectory features. Although both of these steps are very "standard," especially for the case of Mel filter bank spectral values for the preceding step, in this section we review these transforms especially as they relate to using FFT spectral values directly.

The first step of this feature calculation is to compute DCTC terms from the spectrum *X,* with the frequency *f* normalized to a [0, 1] range, as follows

$$DCTC(i) = \int_0^1 a\left(X\left(g(f)\right)\right)\phi_i(f)df \qquad (4)$$

In this equation, *i* is the DCTC index, *a(X)* is a nonlinear amplitude scaling and *g(f)* a nonlinear frequency warping. $\Phi_i(f)$ is the i^{th} basis vector over frequency computed as:

$$\phi_i(f) = \cos\left[\pi i g(f)\right]\frac{dg}{df} \qquad (5)$$

The crucial elements of this approach are the selection of the nonlinear amplitude scaling *a(X)* and the nonlinear frequency scaling *g(f)* so that the cosine transform is with respect to a perceptual scale. In practice, the scaling *a(X)* is typically a log, and the scaling *g(f)* is a Mel-like function unless the first step is a Mel-like filter bank, in which case g(f) = f, dg/df = 1, and the basis vectors are "regular" cosines.

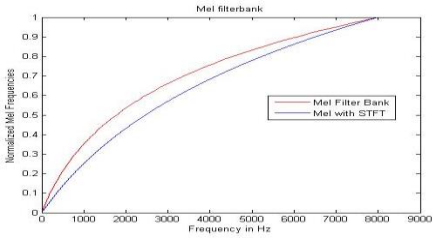

Figure 7: Mel frequency warping used for Mel filter bank center frequencies (top red curve), and "optimum" Mel frequency warping used for FFT-only/DCTC/DCSC method (bottom blue curve)

For the case of FFT-only spectral analysis frequency, g(f) is a Mel-like "warping" function, which has the effect of modifying the cosine basis vectors, according to Eq. 5. The results presented in this paper for the DCTC/DCSC expansion of FFT spectra were based on this Mel-like warping (lower blue curve in Figure 7), which was empirically found to perform better than the more precise Mel warping as given in Eq. 2 and also depicted in Figure 7.

In order to create the DCSC features that represent the spectral evolution of DCTCs over time, as an alternative to delta and delta-delta terms typically used with MFCCs, a cosine basis vector expansion over time is performed using overlapping blocks of DCTCs. That is, the DCSCs are computed as:

$$DCSC(i,j) = \int_0^1 DCTC\left(i,h(t)\right)\theta_j(t)dt \qquad (6)$$

where $\Theta_j(t)$ is the j^{th} basis vector over time computed as:

$$\theta_j(t) = \cos\left[\pi i h(t)\right]\frac{dh}{dt} \qquad (7)$$

In this equation, *h(t)* is a time warping function and *t* is normalized to [0,1] over a selected segment (a "block"). In practice, *t* is discrete, corresponding to a frame index, and the integral is computed using a sum of all frames in the block. The calculation is repeated for each overlapping block, with the block spacing some integer multiple of the frame spacing.

5. Phonetic recognition experiments

Phonetic recognition experiments were conducted using the TIMIT phonetically-labeled database. 3296 sentences from 462 speakers were used for training and 1344 sentences from 168 speakers were used for test. SA sentences were excluded. A frequency range of 100 to 8000 Hz was used for all experiments. Experiments were conducted with clean, 20 dB SNR, 10 dB SNR, and 0dB SNR speech. For all conditions, training and test conditions were matched with respect to noise; additive white Gaussian noise was used for noise.

The objective of the experiments was to compare phoneme recognition accuracy for four spectral analysis methods, as depicted in Figure 8, and also to compare to a control case (13 MFCCs with delta and acceleration terms, or 39 total terms, derived from a Mel filter bank, as implemented in HTK).

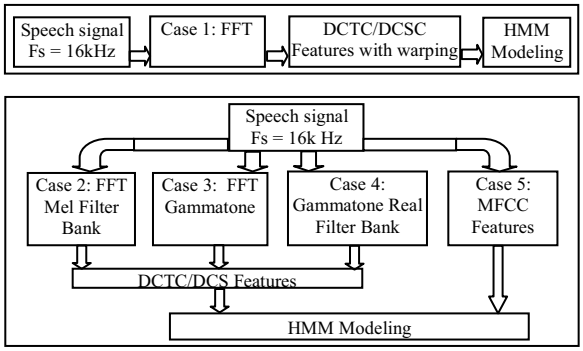

Figure 8: Block diagram of phonetic speech recognition process

Five cases, as depicted in Figure 8, and outlined below were tested.

Case 1: FFT spectrum directly used as front end for DCTC/DCSC feature, using frequency warping (Figure 7).

Case 2: DCTC/DCSC feature extraction applied to Mel filter bank spectrum. Since the filter bank already has warping in it, the DCTC basis vectors have no warping.

Case 3&4: Gammatone filter banks (FFT-based and actual filters cases) used as front end for DCTC/DCSC features, with no frequency warping used for DCTCs.

Case 5: HTK MFCC features with delta terms.

For all experiments with DCTC/DCSC features, a frame spacing of 2ms (500 frames per second) was used. Blocks were comprised of 150 frames (300ms) and spaced 8ms apart (125 blocks per second). Experiments were conducted with both 78 features (13 DCTCs times 6 DCSCs), and the more standard 39 features (13 DCTCs times 3 DCSCs).

HMMs with 3 hidden states from left to right with 16 Gaussian mixtures were used for phonetic recognition experiments. A total of 48 (eventually reduced to 39 phones) context independent monophone HMMs were created using the HTK toolbox (Ver3.4) [12]. The bigram phone information extracted from the training data was used as the language model.

6. Results

Phonetic recognition accuracy (based on 39 phones) obtained for all 5 cases is given in Table 1. It can be seen that there is negligible or no improvement when filter bank techniques are used. For results in Table 1, 39 features were used. The experiment was repeated with 78 features for all cases except MFCC, and results are given in Table 2.

Table 1: Accuracy (%) comparison for 39 features

SNR (dB)	FFT only	Mel FB	Gammatone FFT FB	Gammatone Real FB	MFCC
Clean	69.2	68.5	69.8	69.1	62.8
20 dB	64.2	63.5	63.7	63.4	
10 dB	56.3	55.0	55.8	55.0	
0 dB	42.2	41.5	41.4	40.5	

Table 2: Accuracy (%) comparison for 78 features

SNR (dB)	FFT only	Mel FB	Gammatone FFT FB	Gammatone Real FB
Clean	71.2	69.7	71.1	70.1
20 dB	65.8	64.7	65.8	64.9
10 dB	58.0	58.1	58.1	56.9
0 dB	43.4	42.5	42.8	41.8

Both case 2 and case 5 in Table 1 used Mel warping, but there is a considerable difference in the performance of the two. To investigate the possible reason for this, the delta terms and the DCSC terms were removed from MFCC using HTK and "our" Mel filter bank respectively, and the results shown in Table 3 were obtained.

Table 3: Performance comparison of MFCC and Mel filter bank.

# Channels	MEL FB	MFCC{HTK}
32 {FL=10ms, FS =2ms}	53.9	53.2
32 {FL=25ms, FS =10ms}	49.1	50.2
20 {FL=10ms, FS =2ms}	53.3	53.4
20 {FL=25ms, FS =5ms}	48.9	50.6
26 {FL=10ms, FS =2ms}	53.9	50.5
26 {FL=25ms, FS =10ms}	50.3	50.4

'FL' is the frame length and 'FS' is the frame spacing that is used. The results show that when the delta terms and the DCSC terms are removed, the performance of MFCC computed using HTK is similar to that of the Mel filter bank implemented in our code. Thus, presumably, the advantage of our Mel filter bank versus the HTK filter bank is due to the difference in the way the spectral change information was represented.

As yet another test, Table 4 shows the accuracy obtained with the Gammatone filter bank as the number of channels is varied from 8 to 128. Although there is a very slight improvement when using 64 channels, this comes at the expense of more computational time and complexity, so we considered the "standard" as 32 channels for the Gammatone filters, and used 32 channels for all the results (except for Table 4 results) in this paper.

Table 4: FFT Gammatone performance as number of filters is varied.

SNR (dB)/Channels	8	16	32	48	64	128
Clean	64.5	69.4	71.7	71.3	71.4	71.1
20 dB	60.1	64.8	65.8	65.8	65.9	65.9
10 dB	50.8	56.0	58.1	58.1	59.3	58.1

To test the statistical significance of the differences in accuracy for the results given in this paper, we performed several t-tests by dividing the 1344 sentences of test into sets of 96 sentences each. Using the means and variances of the groups of 14 independent tests, and using standard statistical hypothesis testing methods [13], it was determined that 2% differences are significant at the 97.5% confidence level, and 1% differences are significant at the 90% confidence level. Thus, for example, in Table 1, for a fixed SNR, many of the results are statistically similar, except for MFCC results, which are lower than for all the other methods shown.

7. Conclusion

From the experimental data, we conclude that FFT-based spectral analysis in both clean and noisy conditions with a Mel-like frequency scale incorporated using frequency warping for DCTC features performs nearly identically to cochlea-motivated filter bank spectral analysis. Directly using the FFT spectrum, without the intermediate filter bank prior to feature calculations, has the advantage of simplicity and would appear to be a better front end strategy for spectral front end calculations for speech processing. The DCSC method for computing spectral trajectory features is experimentally shown to result in much higher ASR accuracy than obtained with delta and delta-delta terms.

8. Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA87501210093. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

9. References

[1] S. B. D and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP- 28, no. 4, pp. 357-366, 1980

[2] Yuxuan Wang, Kun Han, DeLiang Wang "Exploring Monaural Features for Classification-Based Speech Segregation," *IEEE transactions on audio, speech and language processing, Vol. 21, No. 2, February 201.*

[3] Md. Afzal Hossan, S. Memon, M A Gregory, "A novel approach of MFCC feature extraction," *IEEE Trans. On Signal Processing and Communication 2010 4th international conference.*

[4] Wu Junqin, Yu Junjun, "An Improved Arithmetic of MFCC in Speech Recognition System," *IEEE 201, pp 719-722*

[5] S.A. Zahorian, Silsbee, P., and Wang, X., "Phone Classification with Segmental Features and a Binary-Pair partitioned Neural Network Classifier," *Proc. ICASSP* 1997, pp.1011-1014, 1997.

[6] M. Karjanadecha and S.A. Zahorian, "Signal Modeling for High-Performance Isolated Word Recognition," *IEEE Trans. on Speech and Audio Processing*, 9(6), pp.647-654, 2001.

[7] S. Strahl, "Analysis and design of Gammatone signal models," *J. Acoust. Soc. Am.* 126, pp. 2379-2389, 2009.

[8] B. Moore, R. Peters, and B. Glasberg, "Auditory filter shapes at low center frequencies," *J. Acoust. Soc. Am*. 88, 132–140, 1990.

[9] B. Moore and B. Glasberg, "A revision of Zwicker's loudness model*," Acta. Acust. Acust*. 82, 335–345, 1996

[10] Holdsworth J. et al. "Implementing a Gamma Tone Filter Bank," in SVOS Final Report – *Part A: The Auditory Filter bank, MRC Applied Psychology Unit*, Cambridge, England, 1988.

[11] L. Rabiner, B.H. Juang, "Fundamentals of speech Recognition," *Prentice Hall Signal Processing Series*, 1993.

[12] S.A. Zahorian, Hongbing Hu, Zhengqing Chen, Jiang Wu, "Spectral and Temporal Modulation Features for Phonetic Recognition," *Interspeech* 2009.

[13] Will Thalheimer, Samantha Cook, "How to calculate effect sizes from published research: A simplified methodology,"*A Work-Learning Research Publication, Published August 2002.*