INCREASED MFCC FILTER BANDWIDTH FOR NOISE-ROBUST PHONEME RECOGNITION

Mark D. Skowronski and John G. Harris

Computational Neuro-Engineering Laboratory, University of Florida

ABSTRACT

Many speech recognition systems use mel-frequency cepstral coefficient (mfcc) feature extraction as a front end. In the algorithm, a speech spectrum passes through a filter bank of mel-spaced triangular filters, and the filter output energies are log-compressed and transformed to the cepstral domain by the DCT. The spacing of filter bank center frequencies mimics the known warped-frequency characteristics of the human auditory system, yet the bandwidths of these filters is not chosen through biological inspiration. Instead they are set by aligning endpoints of the triangle, which is itself an arbitrary shape. It is surprising that for such a popular speech recognition front end, proper analysis or optimization of the filter bandwidths has not been performed. With complex cochlear models, realistic filter shapes that more closely approximate critical bands are used. And these filters, compared to the filters used in mfcc, are considerably wider and overlap with neighboring filters more. We have extended this filter characteristic to the mfcc algorithm and found that the increased filter bandwidth improves recognition performance in clean speech and provides added noise robustness as well.

1. INTRODUCTION

For the past 20 years, the feature extraction algorithm of Davis and Mermelstein [1] has seen widespread use in the field of speech recognition. Mel frequency cepstral coefficients (mfcc) are used in systems with various languages and various tasks (phonemes, isolated words, continuous speech) [2, 3]. The algorithm stems from two ideas: 1) vocal tract modeling, and 2) homomorphic filtering. In the vocal tract model, speech is produced by passing an excitation through a filter whose response models the effects of the vocal tract on the excitation. Furthermore, all relevant speech recognition information is contained in the filter. Homomorphic filtering is an operation transformer. In this case, the convolution of the excitation with the vocal tract response is transformed into addition, where linear filtering techniques are applied to remove the excitation from the filter response. However, instead of transforming the log-compressed spectrum directly, mfcc transforms the log-energies of the spectrum passed through a bank of bandpass filters. This transformation is similar to PCA since the cosine basis vectors of the DCT are similar to the principle eigenvectors of mel-spaced filter bank energies extracted from a typical training population of speech [1].

The mfcc filter bank is composed of triangular filters spaced on a linear-logarithm scale. The spacing of filters follows the mel frequency scale, which is inspired by critical band measurements of the human auditory system, yet the bandwidth of each filters is chosen by aligning the triangle base with the center frequencies of the neighboring filters. Thus neighboring filters intersect at each's quarter power point, which appears too small from an engineering standpoint. When one considers cochlear models of the human auditory system, such as that used in the Ensemble Interval Histogram [4], the filters in the model's filter banks are much wider and overlap with neighboring filters more so than mfcc filters. Inspired by more accurate models of the human auditory system, we investigate the effects of wider mfcc filters in a recognition task. In Section 2, we detail two schemes for increasing filter bandwidth, and in Section 3 we describe two recognition experiments using these schemes. Section 4 includes a discussion of experimental results, and Section 5 concludes with a summary of our results.

2. FILTER WIDENING SCHEMES

The standard mfcc filters are equally spaced in mel frequency space, and their shape is triangular in linear frequency space with the base of each triangle defined by the center frequencies of adjacent filters. Thus, the base of each filter overlaps with its neighbor by 50%. The first scheme for widening the mfcc filters increases this overlap while maintaining the bandwidth of the entire filter bank. Thus the triangle base length L for each filter is

$$L = \frac{\hat{f}_{max} - \hat{f}_{min}}{N(1-m)+m} \tag{1}$$

where $\hat{f}_{max}$ and $\hat{f}_{min}$ are the maximum and minimum frequencies of the filter bank, respectively, in mel frequency space, N is the number of filters in the filter bank, and

m is the percent overlap between adjacent filters bases ($0 \leq m \leq 1$). Since $\hat{f}_{max}$ and $\hat{f}_{min}$ are constant in this scheme, filter bank center frequency $\hat{f}_n$ is a function of m:

$$\hat{f}_n = \frac{L}{2} + (n-1)\frac{\hat{f}_{max} - \hat{f}_{min} - L}{N-1} + \hat{f}_{min} \quad (2)$$

for the n^{th} filter in the filter bank ($1 \leq n \leq N$). Thus, as $m \rightarrow 1$, $B \rightarrow \hat{f}_{max} - \hat{f}_{min}$ and all filters converge to the same center frequency. Yet even for $m = 90\%$ overlap, $f_N - f_1$ in linear frequency is still 80% of that for the traditional mfcc filter bank. We refer to this algorithm as mfccVW since the filters are *variable width* according to the free parameter m.

The second scheme is inspired by the equivalent rectangular bandwidth (ERB) of critical bands in the human auditory system [4]. ERB is the bandwidth of a rectangular filter centered at the center frequency of a critical band whose magnitude is the maximum magnitude of the critical band and whose energy is the same as that of the critical band, described in Equation 3.

$$\mathrm{ERB} = \frac{\int |H(f)|^2 \, \mathrm{d}f}{|H(f_0)|^2} \quad (3)$$

$|H(f)|$ is the filter response centered at f_0. From psychoacoustic experiments, Moore and Glasberg [5] used the quadratic equation in Equation 4 to fit ERB to critical band center frequency f_0 (in kHz).

$$\mathrm{ERB} = 6.23 f_0^2 + 93.39 f_0 + 28.52 \quad (4)$$

Figure 1 compares the ERBs of the mfcc triangular filters for several overlap percentages m to the ERB for humans in Equation 4. We see that the traditional mfcc filters ($m = 50\%$ overlap) have lower bandwidths than that of human critical bands by up to a factor of 3 for the widest filters, while increasing m increases ERB according to our first scheme.

The second scheme for increasing filter bandwidth is to set the bandwidth according to Equation 4. The center frequencies used by the traditional mfcc are kept constant, and the ERB for each center frequency is calculated. Lower and upper filter frequencies are symmetric about each center frequency in mel frequency space and are determined such that they produce the desired ERB when warped back to linear frequency space. Let f_H and f_L represent the high and low frequencies, respectively, of a given triangular filter in linear frequency space. The mel frequency warping function between linear frequency f and mel frequency $\hat{f}$ is

$$\hat{f} = 2595 \log_{10}(1 + \frac{f}{700}) \quad (5)$$

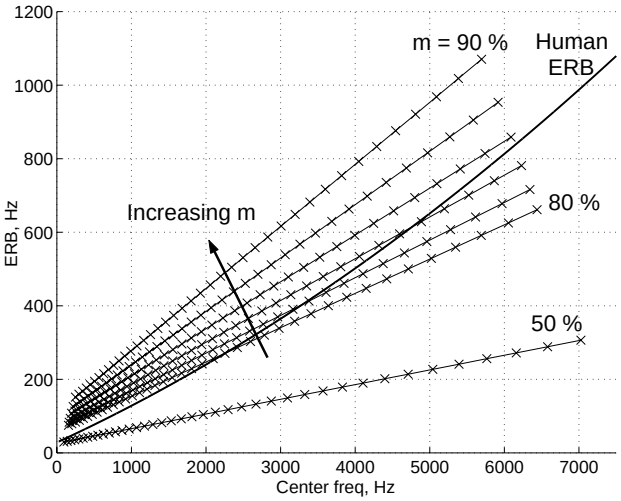


Fig. 1. ERB vs Center Frequency for mfcc ($m = 50\%$) and mfccVW ($m = 80\%, 82\%, \ldots, 90\%$) as well as for the human auditory system (See Equation 4). Each x corresponds to a filter center frequency.

and

$$\begin{aligned} \mathrm{ERB} &= \frac{1}{3}(f_H - f_L) \\ \hat{f}_0 &= \frac{1}{2}(\hat{f_H} + \hat{f_L}) \end{aligned} \quad (6)$$

are solved for f_H and f_L when f_0 is given and ERB is calculated from Equation 4. In this scheme, we choose to scale the ERB by an inflation factor to further increase filter bandwidth. Filters were truncated to fit between 0 Hz and the Nyquist rate. We refer to this scheme as mfccERB, and, for simplicity, we shall refer to the traditional algorithm as mfcc.

3. EXPERIMENT

To characterize our modified filter banks, we perform two experiments on vowels extracted from the TIMIT database. The vocabulary consists of 10 vowels (/IY/, /IH/, /EH/, /AE/, /AA/, /UH/, /UH/, /UW/, /AH/, /ER/) extracted from read sentences according to the phonetic labels provided by the corpus ($\sim 50,000$ phonemes in all). Phonemes larger than 512 samples (32 ms) are truncated about the center of the phoneme. White gaussian noise is added to produce a desired SNR. Next, a 512-point Hamming window is applied to each utterance, which is then Fourier transformed to the frequency domain. The log-energies for the various filter bank schemes are calculated, then transformed to the cepstral domain via the DCT. Filter banks of length $N = 40$ are employed, and, after dropping the first coefficient, 10 cepstral coefficients are retained for each utterance.

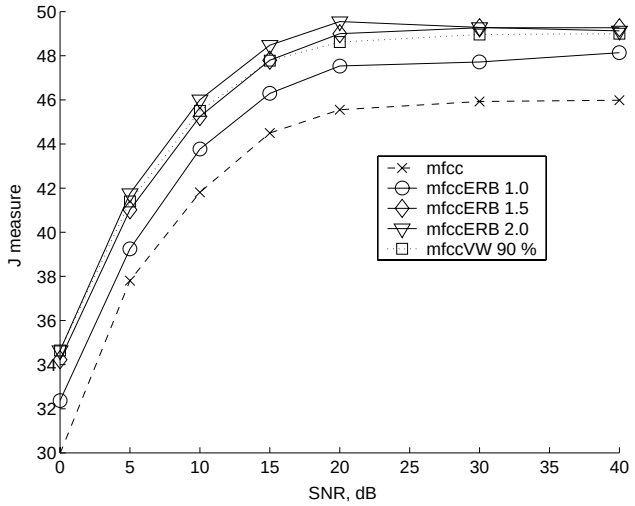


Fig. 2. J-measure vs SNR for mfcc, mfccVW, and mfccERB (inflation factors $1.0, 1.5$, and 2.0).

The first experiment measures the Fisher discriminant (J-measure) for the 10-class problem. This measure compares variance between classes to variance within each class. Larger J-measures denote greater separation between classes. The J-measure is defined as [6]

$$J = \mathrm{trace}(\mathbf{S}_W^{-1}\mathbf{S}_B) \tag{7}$$

where

$$\begin{aligned} \mathbf{S}_W &= \sum_{k=1}^{c} \mathbf{S}_k = \sum_{k=1}^{c} N_k \mathbf{\Sigma}_k \\ \mathbf{S}_B &= \sum_{k=1}^{c} N_k (\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T \end{aligned} \tag{8}$$

and $\mathbf{m}_0$ is the mean vector of the entire data set. N_k is the number of samples in the k^{th} class (c classes total). $\mathbf{S}_W$ is the within-class scatter and $\mathbf{S}_B$ is the between-class scatter. Figure 2 shows the results of this experiment. From this figure we see that increasing filter bandwidth improves the separability measure over mfcc for all noise levels. For mfccERB, J increases as ERB scale factor increases.

Encouraged by these results, we performed a second experiment using a Bayes classifier. In this experiment, each class is divided into test and train data (80% train). Features from train data are extracted with no noise, while white noise is added to test data before features are extracted. Classification is determined by maximizing the discriminant function g_k [6]

$$g_k = \mathbf{x}^T \mathbf{W}_k \mathbf{x} + \mathbf{w}_k^T \mathbf{x} + w_{k0} \tag{9}$$

where

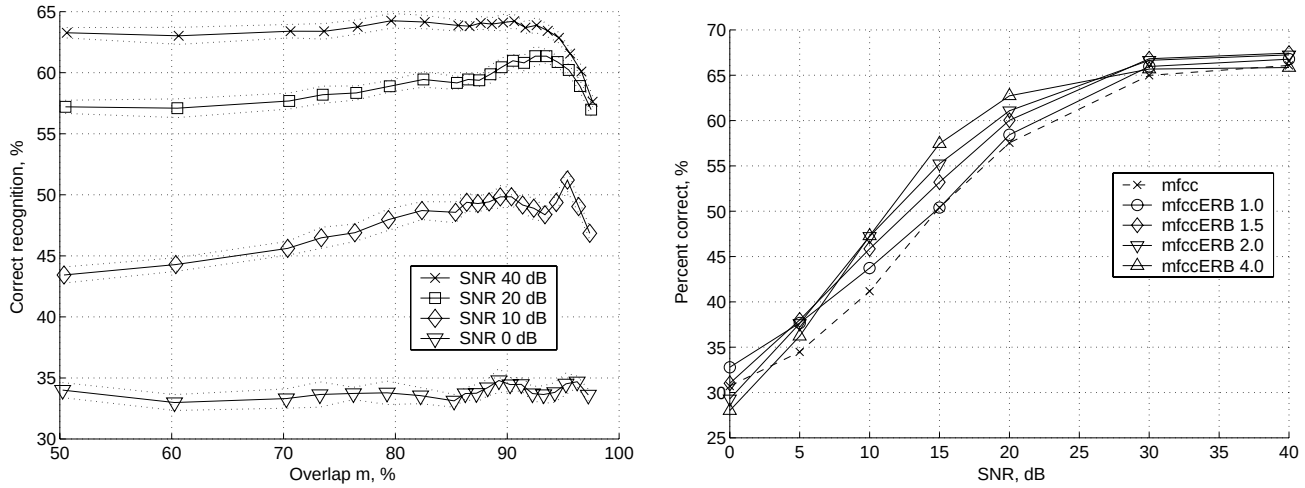
$$\begin{aligned} \mathbf{W}_k &= -\frac{1}{2}\mathbf{\Sigma}_k^{-1} \\ \mathbf{w} &= \mathbf{\Sigma}_k^{-1}\mathbf{m}_k \\ w_{k0} &= -\frac{1}{2}\mathbf{m}_k^T \mathbf{\Sigma}_k^{-1}\mathbf{m}_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log P_k \end{aligned} \tag{10}$$

for covariance $\mathbf{\Sigma}_k$ and mean $\mathbf{m}_k$ of the k^{th} class with *a priori* probability P_k. $\mathbf{x}$ is the vector of cepstral coefficients for the test data and is classified as $\arg(\max_k g_k)$. Figure 3 shows the recognition results for this experiment.

4. DISCUSSION

Results from both experiments show improvement in separability measure J or recognition error rate when applying wider filters in place of the traditional mfcc filters. These results are surprising since feature extraction analysis seeks to produce features as orthogonal as possible. Wider filters mean increased overlap and higher correlation among the energy outputs and is usually undesired. However, a second transform is applied which fundamentally changes the problem. Consider the cepstral bandwidth of an mfcc filter. Narrow filters in the frequency domain are relatively wide low-time lifters in the quefrency domain, while wider filters transform to relatively narrow low-time lifters in the quefrency domain. The filters of mfccVW and mfccERB have lower cutoff quefrencies than the corresponding filters of mfcc, providing extra cepstral liftering which tends to remove more noise than signal. For mfccERB with scale factor 4.0 in Figure 3(b), recognition is nearly the same between 30 and 40 dB SNR while the other filter banks increased in performance over that same range. This indicates that the large inflation factor produces filters wide enough to smooth out relevant signal information, causing a relative degradation of clean-speech recognition. This also explains why the results using the largest mfccERB scale factor are highest for moderate noise ($5 - 30$ dB SNR). Below 5 dB SNR, the experiment starts approaching *a priori* values, and the results are less reliable.

The mfccVW results in Figure 3(a) shows similar performance to mfccERB at moderate noise (the dramatic change in recognition results for $m > 90\%$ results from the rapid convergence of center frequencies as $m \rightarrow 1$ which is due to the unnatural parameterization for this scheme). Performance at all SNRs is nearly maximized for m near 90%. In Figure 1, the ERB curve for $m = 90\%$ lies above the human ERB curve by about 50%, or near the ERB curve with scale factor 1.5. The results in Figure 3 (a) and (b) suggest that these are the best overall operating points for their respective schemes, indicating an optimal low-time liftering scheme is possible in the quefrency domain. This suggests that an optimal filter bandwidth scheme may be found



(a) (b)

Fig. 3. Bayes classifier results vs SNR for (a) mfcc ($m = 50\%$) and mfccVW for various SNRs vs Overlap percentage m, and (b) mfcc and mfccERB with ERB scale factors $1.0, 1.5, 2.0$, and 4.0.

by starting with quefrency domain filtering requirements for desired cepstral shaping before transforming back to the frequency domain.

5. CONCLUSIONS

We have found that widening the filters found in the mfcc filter bank increases recognition for clean speech and provides robust performance in additive white noise. The two schemes for increasing filter bandwidth, mfccVW and mfccERB, improve recognition performance over the traditional mfcc filter bank by removing noise through low-time liftering. Experiments with 80 filters in the filter bank showed no changes in performance for mfccERB but a slight decrease in performance for mfcc, especially for low SNRs. As the number of filters increases, the filter bandwidths for mfccERB remain the same since bandwidth is only a function of center frequency, while the overlap with neighboring filters remains constant for mfcc. This means that the filters in mfcc decrease in frequency bandwidth, and consequently increase in low-time cutoff in the quefrency domain. Since clean speech recognition for mfcc didn't change when increasing from 40 to 80 filters, aliasing of pitch-related information *before* cepstral truncation is not an issue even for these relative large low-time cutoffs. And the degraded performance as SNR decreases indicates more noise is passing through these low-time lifters. Since performance is due to filter properties in the transformed domain, optimum performance may be found by filter design in the transformed domain. Improved performance by employing more sophisticated filter bank schemes usually increases the computational cost of a feature extractor, yet our two schemes outperform the tradinonal mfcc algorithm and while retaining the same computational cost, with only a slight increase coming from the increased number of non-zero terms in each filter representation.

6. REFERENCES

[1] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28(4), pp. 357–366, 1980.

[2] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, 1993.

[3] V. Digalakis, M. Ostendorf, and J. Rohlicek, "Fast algorithms for phone classification and recognition using segment-based models," 1992.

[4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," in *IEEE Trans. on Speech and Audio Processing*, 1994, vol. 2, pp. 115–132.

[5] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," in *J. Acoust. Soc. America.*, 1983, vol. V74, pp. 750–753.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Academic Press, New York, 1973.