IMPROVING THE FILTER BANK OF A CLASSIC SPEECH FEATURE EXTRACTION ALGORITHM

Mark D. Skowronski and John G. Harris

Computational Neuro-Engineering Lab
University of Florida, Gainesville, FL, USA
markskow,harris@cnel.ufl.edu

ABSTRACT

The most popular speech feature extractor used in automatic speech recognition (ASR) systems today is the mel frequency cepstral coefficient (mfcc) algorithm. Introduced in 1980, the filter bank-based algorithm eventually replaced linear prediction cepstral coefficients (lpcc) as the premier front end, primarily because of mfcc's superior robustness to additive noise. However, mfcc does not approximate the critical bandwidth of the human auditory system. We propose a novel scheme for decoupling filter bandwidth from other filter bank parameters, and we demonstrate improved noise robustness over three versions of mfcc through HMM-based experiments with the English digits in various noise environments.

1. INTRODUCTION

Davis and Mermelstein (D&M) coined the term 'mel frequency cepstral coefficients' (mfcc) in 1980 when they combined nonuniformly-spaced filters with the discrete cosine transform (DCT) as a front-end algorithm for automatic speech recognition (ASR) [1]. The algorithm can be summarized as follows: a signal passes through a triangular filter bank, spaced on a linear-log frequency axis, and the energy output from each filter is log-compressed and transformed via the DCT to cepstral coefficients. Previous work by Pols [2] showed that the eigenvectors of the log-energy output from his filter bank resembled cosine basis vectors for Dutch vowel data; hence, the DCT provides a quasi-PCA decorrelation of the log-energy, which allows for low-time liftering in the cepstral domain (dimension reduction). The log function provides compression of the dynamic range of filter bank output energies while also making the distribution of output energies more Gaussian. Therefore, all other functionality of mfcc is attributed to characteristics of the bank of triangular filters (see Figure 1). The energy calculation of each filter smooths the speech spectrum, reducing the effects of pitch, while the warped frequency scale provides variable sensitivity to the speech spectrum inspired by the human auditory system.

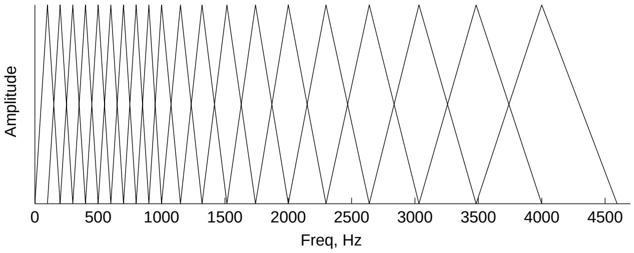
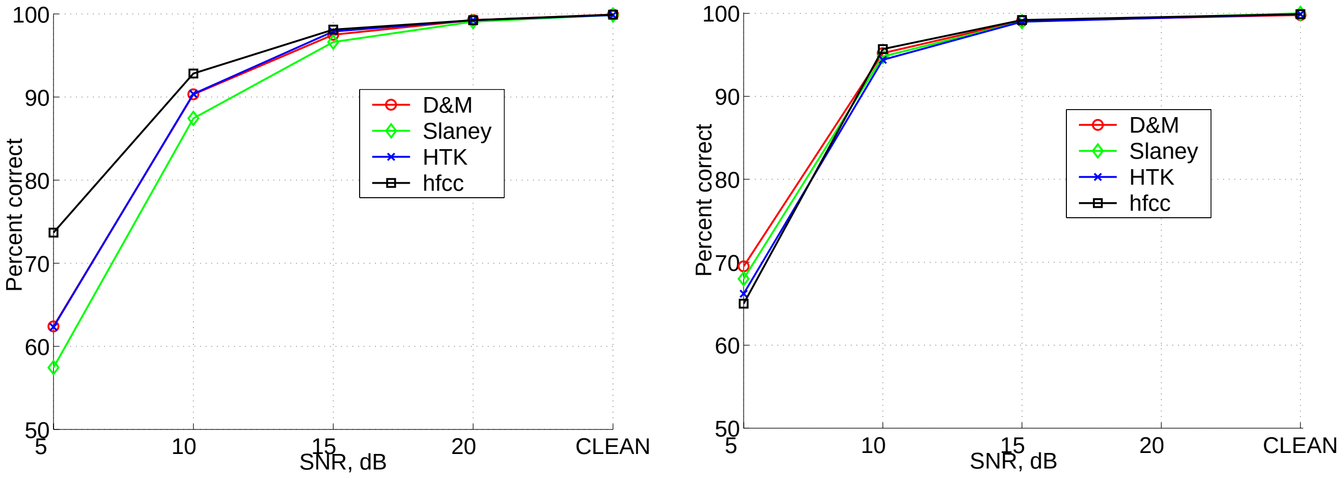


Figure 1: Filter bank of Davis and Mermelstein's mfcc algorithm. The filter bank is comprised of 10 linearly-spaced centers below 1 KHz and 10 log-spaced filters above 1 KHz. The base of each triangular filter is determined by the center frequencies of the neighboring filters, and all filters are unity height.

As seen in Figure 1, the bandwidth of each filter (the principle factor determining spectral smoothing) is arbitrarily set by fixing the base of each triangular filter by the center frequencies of the neighboring filters. Furthermore, popular variations of the mfcc filter bank, in an effort to accommodate data of sampling frequencies greater than 8 KHz, have increased the number of filters present and changed the function for frequency warping *without regard to changes in filter bandwidth that these modifications incur*. For example, Malcolm Slaney's Matlab version of mfcc [3] doubles the number of filters, effectively halving the bandwidth of D&M's filters, and Steve Young's HMM Toolkit (HTK) [4], a principle tool in C/C++ for large vocabular ASR for labs throughout the world, features an mfcc function that allows the user to select frequency range and number of filters for the filter bank (but not bandwidth!). These methods, as well as D&M's original version, are limited by the fact that filter bandwidth is not an independent design parameter; instead, bandwidth is determined by the filter spacing. Bandwidth should at least be related to filter center frequency, as inspired by the critical bands of the human auditory system.

In this paper we introduce a novel scheme for determining filter bandwidth, based on the approximation of critical



(a)

(b)

Figure 2: Absolute recognition results for male-only speakers (50% test/train), averaged over 10 trials, for (a) white noise and (b) pink noise. See text for details about the various mfcc algorithms.

band equivalent rectangular bandwidth (ERB) from Moore and Glasberg [5]. The new scheme, called human factor cepstral coefficients (hfcc), decouples bandwidth from other filter bank design parameters (frequency range, number of filters), allowing for independent design and optimization of bandwidth. We show through ASR experiments that hfcc, using Moore and Glasberg's expression for critical band ERB, produces recognition results at or above those of three popular versions of mfcc (D&M, Slaney, and HTK) in various noise environments. We go on to show that, since bandwidth is an independent design parameter in hfcc, we can improve performance even more by *increasing* filter bandwidth beyond that of Moore and Glasberg's ERB expression.

2. FILTER DESIGN IN HFCC

The key aspect to hfcc is the decoupling of filter bandwidth. The given design parameters are sampling frequency f_s (frequency range $[f_{\min}, f_{\max}]$), number of filters N, and frequency warping function. We use Fant's expression [6] relating mel frequency $\hat{f}$ to linear frequency f:

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

Let f_{l_i}, f_{c_i}, and f_{h_i} be the low, center, and high frequencies for the i^{th} filter in linear frequency, and let $f_{\min}$ and $f_{\max}$ define the frequency range for the entire filter bank. We require that center frequencies are equally-spaced in mel frequency and that the filters are equilateral in mel frequency. That is,

$$\hat{f}_{c_i} = \frac{1}{2}(\hat{f}_{h_i} + \hat{f}_{l_i}). \qquad (2)$$

The steps for filter bank design are summarized as follows:

1. Determine the first and last filter's center frequency. The two equations needed to solve for f_{c_i} come from Equation 2 as well as from the expression of ERB for a triangular function and Moore and Glasberg's ERB expression:

$$\begin{aligned} (700 + f_{c_i})^2 &= (700 + f_{h_i})(700 + f_{l_i}) \\ a f_{c_i}^2 + b f_{c_i} + c &= \frac{1}{2}(f_{h_i} - f_{l_i}) \qquad (3) \end{aligned}$$

   where f_{c_i} in Hz and $a = 6.23\ \ 10^{-6}$, $b = 93.39\ \ 10^{-3}$, and $c = 28.52$ [5].

2. Find the remaining center frequencies:

$$\hat{f}_{c_i} = \hat{f}_{c_1} + (i-1)\frac{\hat{f}_{\max} - \hat{f}_{\min}}{N-1} \qquad (4)$$

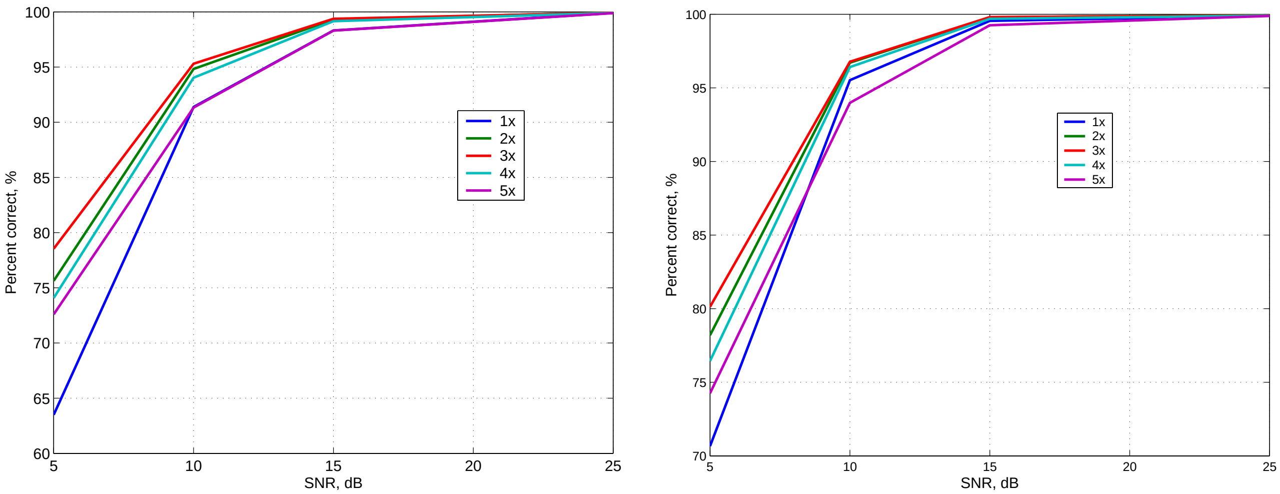
3. Find lower and upper frequencies:

$$\begin{aligned} (700 + f_{c_i})^2 &= (700 + f_{l_i} + 2\mathrm{ERB}_i)(700 + f_{l_i}) \\ f_{h_i} &= f_{l_i} + 2\mathrm{ERB}_i \qquad (5) \end{aligned}$$

4. Construct filter in frequency domain by connecting straight lines between f_{l_i} and f_{c_i} and between f_{c_i} and f_{h_i}. The triangle has zero height at each end and unity height at f_{c_i}.

3. EXPERIMENTS

We evaluate the performance of hfcc and mfcc through ASR experiments. HMM word models for each of the English



(a)

(b)

Figure 3: Absolute recognition results for male-only speakers (50% test/train), averaged over 6 trials for (a) white noise and (b) pink noise. Filter bandwidth (ERB expression) is scaled by a constant (1x, 2x, ...) during filter bank construction.

digits ‘zero’ through ‘nine’ are constructed from utterances taken from the TI-46 corpus of isolated digits. Three versions of filter banks for mfcc are included in the tests: 1) D&M’s original scheme, 2) Malcolm Slaney’s lin-log-spaced Matlab function, and 3) HTK’s C++ function.

Cepstral mean subtraction is applied to all feature vectors, and Δ coefficients with a lag of ± 4 frames are appended to the original 13 cepstral coefficients (26 coefficients total per window). HMMs using Gaussian mixture models are constructed with Compaq’s Probabilistic Model Toolkit for Matlab [7]. Models are trained with noise-free utterances, while noisy test utterances are generated at various signal-to-noise ratios (SNR) by adding noise from the Noisex92 database [8]. For our experiments, we chose white, pink, and babble noise sources.

4. RESULTS

Figure 2(a) shows the absolute results for male-only speakers in white noise, while (b) shows the absolute results in pink noise. Results in babble noise showed no significant difference between all feature extraction algorithms and are not included here. This is expected, since smoothing of the spectrum does not reduce the effects of the non-stationary babble noise. When compared relative to D&M over each trial in white noise, hfcc increases recognition by up to 12 ± 5 percentage points at 5 dB SNR (10 trials of random test/train speakers). Notice that all algorithms perform near-perfectly when no noise is present in the utterance for this vocabulary. Also, hfcc has made no assumptions about the various noise sources–robustness is increased due to the emphasis of frequency ranges in the noisy spectrum. The hfcc algorithm, using Moore and Glasberg’s ERB, shows significant robustness to white noise over the other mfcc methods, while the performance of hfcc in pink noise is slightly worse than that of mfcc, particularly D&M’s version.

In a second experiment, we scaled Moore and Glasberg’s ERB by integer factors between 1.0 and 5.0. Previous experiments with widening schemes showed improved noise robustness [9] for the wider filters,yet with the log compression in the algorithms, analytic analysis is difficult. Figure 3 shows recognition results for the same ASR experimental setup as before using various ERB scale factors in white and pink noise. Marked improvement, up to 15 ± 6 percentage points for white noise and 10 ± 5 percentage points for pink noise above the filter bank of unscaled bandwidth (1x) relative for each trial were achieved with a scale factor of 3 (6 trials).

5. CONCLUSIONS

We have introduced a novel scheme for designing the filter bank in mfcc that decouples filter bandwidth from other filter bank design parameters. By creating filter bandwidth as an independent design parameter, hfcc allows one to increase ASR performance by improving the tradeoff between noise smoothing and resolution of spectral characteristics [10]. As filter bandwidth increases, the more samples are present in the filter to estimate log energy. Therefore, the variance of the estimate decreases. Concurrently, spectral details are blurred by wide filters. Without analytic expressions for the distributions (due to the nonlinear complexity of the feature extraction algorithm), ASR experiments are one of the few useful tools in characterizing this tradeoff. Experiments with simplified vowel models indicate that filter bandwidth affects the trajectory of the feature pdf as SNR decreases, though it currently remains a topic of research to characterize the trajectory behavior as a function of filter bank design parameters.

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] L. C. W. Pols, *Spectral analysis and identification of Dutch vowels in monosyllabic words*, Ph.D. thesis, Free University, Amsterdam, The Netherlands, 1977.

[3] M. Slaney, *Auditory Toolbox, Version 2, Technical Report No: 1998-010*, Interval Research Corporation, 1998.

[4] S. J. Young et al., *The HTK Book*, Entropics Cambridge Research Lab, Cambridge, UK, 1995.

[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] C. G. M. Fant, "Acoustic description and classification of phonetic units," *Ericsson Technics*, vol. 15, no. 1, 1959, reprinted in *Speech Sound and Features*, MIT Press, Cambridge, 1973.

[7] "Compaq probability model toolbox," http://research.compaq.com/downloads.html.

[8] "NOISEX92 noise database," http://spib.rice.edu/spib/select_noise.html.

[9] M. D. Skowronski and J. G. Harris, "Increased mfcc filter bandwidth for noise-robust phoneme recognition," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 801–4, 2002.

[10] M. D. Skowronski and J. G. Harris, "Human factor cepstral coefficients," December 2002, Cancun, Mexico.