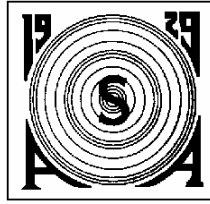


FIRST PAN-AMERICAN/IBERIAN MEETING ON ACOUSTICS
PRIMERA REUNION PAN-AMERICANA/IBERICA DE ACUSTICA
2-6 Dec. 2002 **Cancún, Q.R. México**



Human Factor Cepstral Coefficients

Mark D. Skowronsk and John G. Harris
Computational Neuro-Engineering Laboratory
University of Florida, Gainesville, FL, USA

Abstract

Automatic speech recognition (ASR) is an emerging field with the goal of creating a more natural man/machine interface. The single largest obstacle to widespread use of ASR technology is robustness to noise. Since human speech recognition greatly outperforms current ASR systems in noisy environments, ASR systems seek to improve noise robustness by drawing on biological inspiration. Most ASR front ends employ mel frequency cepstral coefficients (mfcc) which is a filter bank-based algorithm whose filters are spaced on a perceptually-motivated linear-log frequency scale. However, filter bandwidth is set by filter spacing and not through biological motivation. The coupling of filter bandwidth to other filter bank parameters (frequency range, number of filters) has led to variations of the original algorithm with different filter bandwidths. In this work, a novel extension to mfcc is introduced which decouples filter bandwidth from the other filter bank parameters by employing the known relationship between filter center frequency and critical bandwidth of the human auditory system. The new algorithm, called human factor cepstral coefficients (hfcc), is shown to outperform the original mfcc and two popular variations in ASR experiments and various noise sources.

Introduction

Automatic speech recognition (ASR) is an emerging technology for enhancing the man/machine interface. Commercial products currently exist in niche applications, yet widespread usage has not been realized, primarily due to the mismatch between test and train environments. All ASR systems contain a classifier trained on known speech signals (usually noise-free) with the goal of learning the phonetic content of the speech in order to recognize it in an unknown (test) utterance of speech. The test environment may be different from the train environment (different speakers, background noise, channel effects), and providing robustness to mismatch is currently an unsolved problem in the ASR community. Methods for reducing the effects of noise can be classified into three categories [1]: 1) warp noisy test features to the clean feature space, 2) warp clean train space to noisy feature space, 3) design noise-robust characteristics into the feature extractor/classifier. Methods 1 and 2 rely on accurate noise estimates and assumptions about the noise source. Noise-robust features are attractive in that the method makes few assumptions about the noise while offering a biologically-plausible avenue of research.

In this paper we outline a novel algorithm for noise-robust feature extraction—human factor cepstral coefficients (hfcc)—which is derived from the popular mel frequency cepstral coefficient (mfcc) feature extraction algorithm. A description of the original mfcc algorithm is given as well as the limitations of the algorithm which led to the development of hfcc. We then describe our novel modification of mfcc in detail

along with the design for our ASR experiments. We conclude with comments on the experimental results as well as the role of hfcc in future ASR systems.

Mel frequency cepstral coefficients

A filter bank-based method incorporating characteristics of the human auditory system, mfcc is widely used in ASR applications. A segment of speech (typically 20 ms) is windowed and transformed to the frequency domain via the fast Fourier transform (FFT), then the magnitude spectrum of the utterance is passed through a bank of triangular-shaped filters whose center frequencies are spaced along the perceptually-motivated mel frequency scale. The energy output from each filter is then log-compressed and transformed to the cepstral domain via the DCT. The first coefficient, c_0 (the energy term), is typically discarded, and the remaining 10-15 coefficients are used to describe the current window of speech. First described by Davis and Mermelstein (D&M) in 1980 [2], mfcc is widely used in ASR primarily for three reasons: 1) The cepstral features are roughly orthogonal because of the DCT, 2) cepstral mean subtraction [3] eliminates static channel noise, and 3) mfcc is less sensitive to additive noise than linear prediction cepstral coefficients (LPCC) [4]. Linear prediction models the spectrum with an all-pole model, yet additive noise introduces zeros into the spectrum which are not accounted for by the model [5]. The key component of mfcc responsible for noise robustness is the filter bank—the filters smooth the spectrum, reducing variation due to additive noise across the bandwidth of each filter. We detail the filter bank design of mfcc in the following section.

Mfcc filter bank

D&M describe the filter bank used in mfcc by including a plot of the filter bank as well as by stating that ‘mel scale’ means linear below 1 KHz and logarithmic above 1 KHz. Observation of the plot shows 20 filters: 10 with center frequencies equally-spaced between 100 and 1000 Hz, 5 log-spaced between 1 and 2 KHz, and 5 log-spaced between 2 and 4 KHz. The filters are triangular in shape, and the base of each triangle is defined by the center frequencies of the neighboring filters. Used with speech sampled at 8 KHz, extension of mfcc to other sampling rates is not obvious and has led to variations among ASR researchers in implementation. A serious limitation of the original algorithm, present in nearly all variations used in practice, is the following: *filter bandwidth is not an independent design parameter* but instead is determined by the frequency range of the filter bank and the number of filters used. Thus, as researchers add or subtract filters to accommodate the sampling frequency of their experiments, they inadvertently change the bandwidth of their triangular filters. In order to decouple filter bandwidth from other design parameters, we introduce a novel scheme for designing each filter by incorporating knowledge about the critical bands of the human auditory system.

Hfcc filter bank

Perceptual studies reveal a filter bank structure of the human auditory system. The relationship between center frequency and bandwidth of these critical bands is well known, yet this relationship is ignored in mfcc filter bank design. Hfcc determines filter bandwidth by using the following expression for critical band equivalent rectangular bandwidth (ERB) from Moore and Glasberg [6]:

$$\text{ERB}(f_c) = 6.23 f_c^2 + 93.39 f_c + 28.52 \text{ Hz} \quad (1)$$

for filter center frequency f_c in KHz. See [7] for filter bank construction details.

Experiments

We evaluate the performance of hfcc and mfcc through ASR experiments. HMM word models for each of the English 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 [8], and 3) HTK’s [9] C++ function using Fant’s expression for the mel frequency scale [10]. Slaney’s and HTK’s versions are popular in the literature and are included to show variations among mfcc variations as well as between mfcc and hfcc. Figure 1 shows the ERB for each of the four filter banks.

Cepstral mean subtraction [3] is applied to all feature vectors, and Δ coefficients [11] are appended to the original 13 cepstral coefficients as well (26 coefficients total per window). HMMs using Gaussian mixture models are constructed with Compaq's Probabilistic Model Toolkit for Matlab [12]. 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 [13]. For our experiments, we chose white, pink, and babble noise sources.

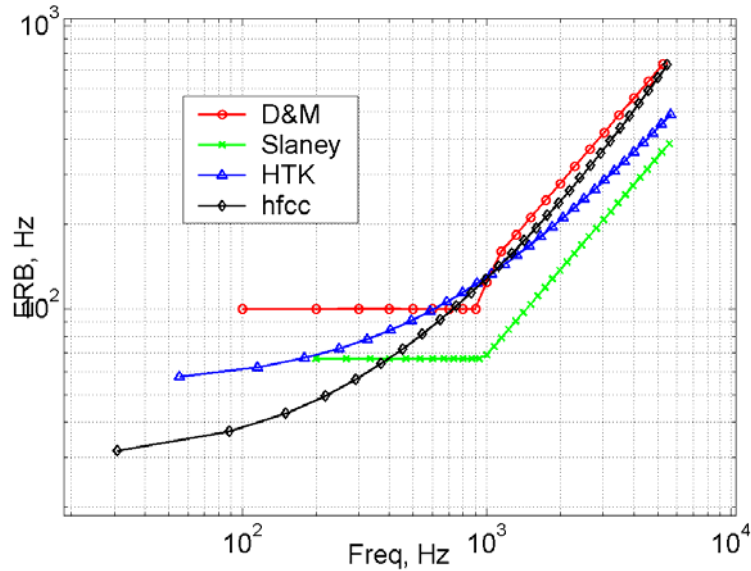


Figure 1. ERB for hfcc and three versions of mfcc: 1) D&M's original, 2) Slaney's Matlab function, and 3) HTK's C++ function.

The experiments test speaker-independent performance for additive noise over the male speakers (8 speakers total). For each trial (10 trials total), four speakers are chosen at random to train the HMMs while the other four have noise added before feature extraction. Figures 2 and 3 show absolute and relative performance for all four algorithms in white noise. Hfcc shows significant improvement in robustness to white noise as compared to the three versions of mfcc. The relative results of Figure 3 reduce inter-trial variation due to speakers by comparing the difference between each method and D&M for each trial. Results with babble noise show little difference between the four methods, as seen in Figures 4 and 5. For pink noise, D&M outperforms all other tested methods, as seen in Figures 6 and 7.

Discussion

The results can be described by considering the characteristics of the various noise sources. White noise is equal energy per Hz frequency, while the pink noise is equal energy per 1/3 octave, both of which are stationary sources. The bandwidth of the filters determines the tradeoff between noise smoothing and resolution of spectral characteristics, so stationary noise performance can be explained by the filter ERBs in Figure 1. While hfcc may find the best balance for the distribution of energy in white noise, the narrow filters below 1 KHz are too narrow for the predominately low-frequency energy of pink noise. The highest-to-lowest performance of methods in pink noise mimic the widest-to-narrowest filters below 1 KHz. Since babble noise is a nonstationary source, the tradeoff between noise smoothing and resolution averages out for the four methods.

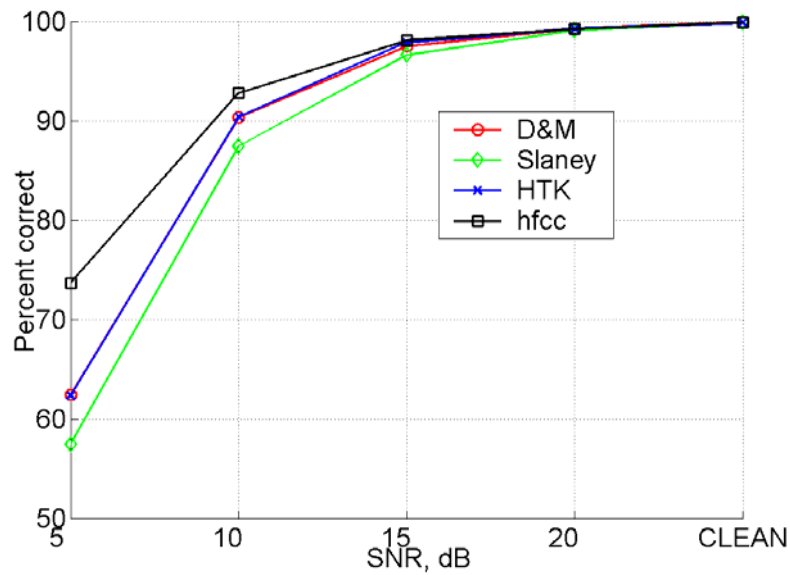


Figure 2. HMM recognition of English digits in additive white Gaussian noise, male test/train speakers. Average over 10 trials.

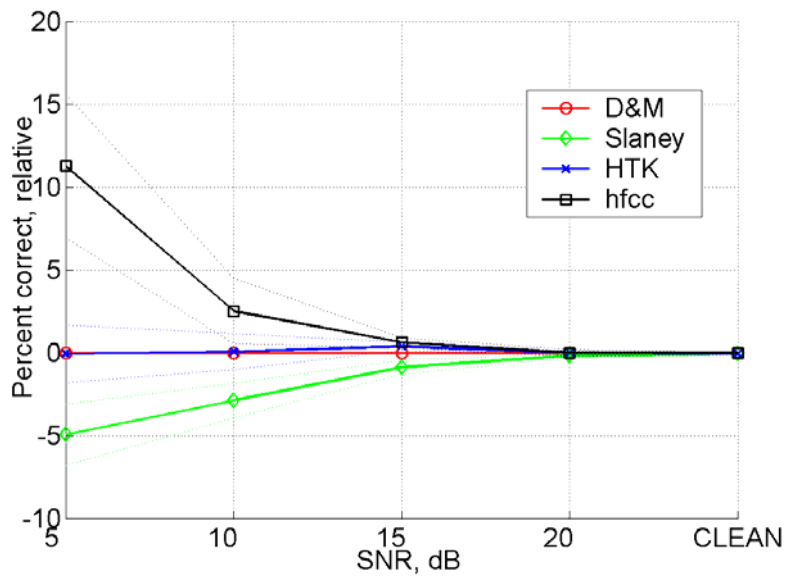


Figure 3. HMM recognition of English digits in additive white Gaussian noise, male test/train speakers, relative to D&M. Mean \pm 95% confidence interval over 10 trials.

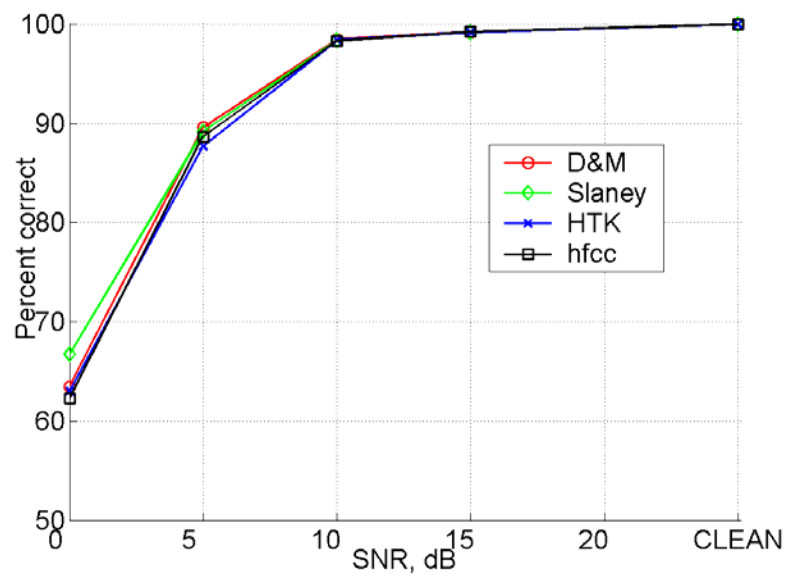


Figure 4. HMM recognition of English digits in additive babble noise, male test/train speakers. Average over 10 trials.

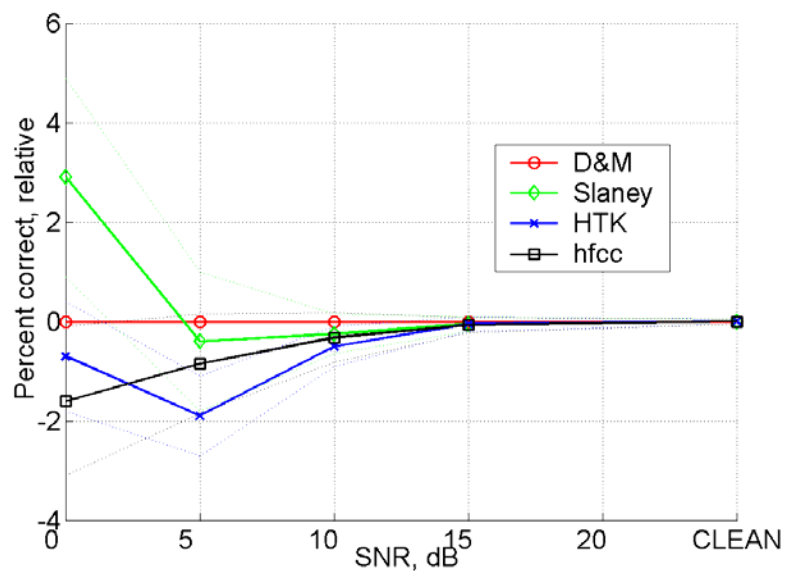


Figure 5. HMM recognition of English digits in additive babble noise, male test/train speakers, relative to D&M. Mean \pm 95% confidence interval over 10 trials.

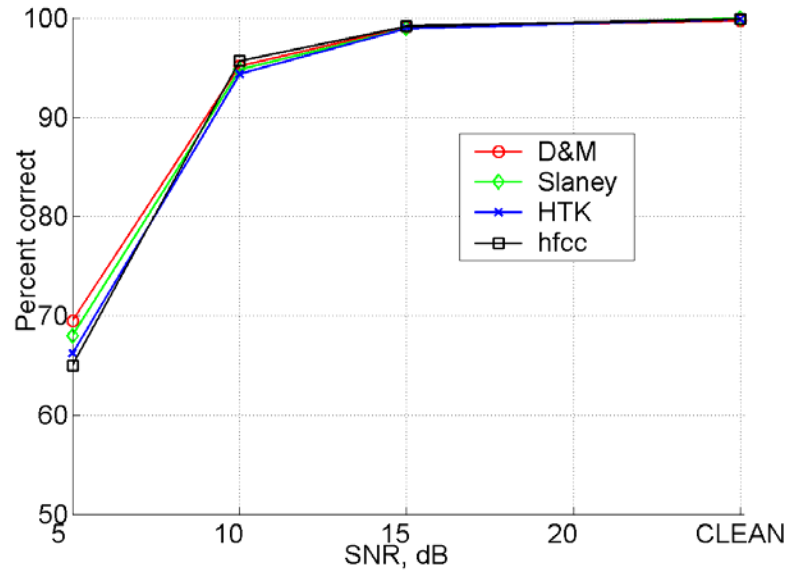


Figure 6. HMM recognition of English digits in additive pink noise, male test/train speakers. Average over 10 trials.

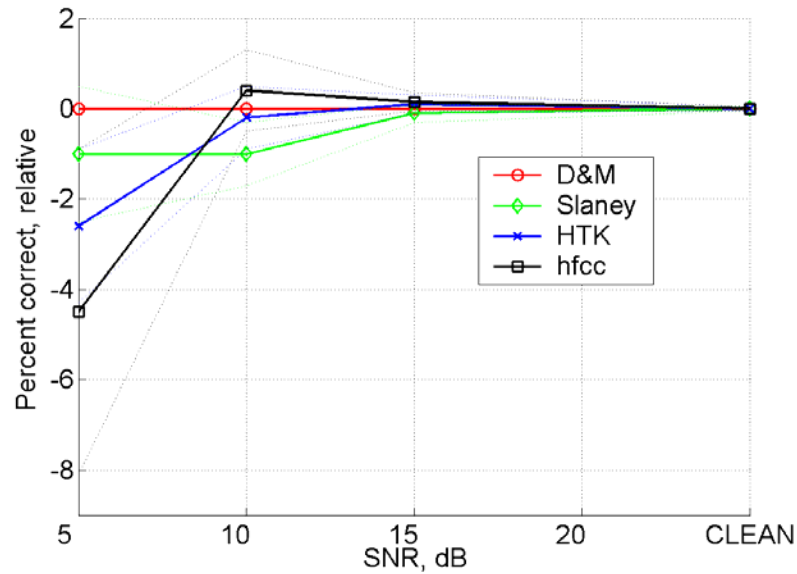


Figure 7. HMM recognition of English digits in additive pink noise, male test/train speakers, relative to D&M. Mean \pm 95% confidence interval over 10 trials.

Conclusions

We have introduced a novel approach—human factor cepstral coefficients—to designing the filter bank in mfcc that decouples filter bandwidth from other design parameters. In mfcc, bandwidth of a filter is determined by the center frequencies of the neighboring filters. Yet it is well known that critical bandwidth of the human auditory system is a function of critical band center frequency. In our new filter bank design,

we use the critical band ERB approximation of Moore and Glasberg to relate filter bandwidth to center frequency. The new scheme allows us to set bandwidth independently from other filter bank parameters. In ASR experiments with various additive noise sources, we characterized the tradeoff between noise smoothing and spectral resolution for hfcc as well as three popular versions of mfcc, each version having a different filter bank design scheme. We found that performance varied significantly across the different noise sources. This shows that bandwidth of filters in mfcc or hfcc is a significant factor in determining noise-robust performance. Unlike mfcc, hfcc allows for independent design of filter bandwidth in order to optimize the tradeoff between noise smoothing and spectral resolution.

Bibliography

1. Y. Gong, Speech Recognition in Noisy Environments: A Survey, *Speech Communication*, vol. 12, no. 3, pp. 231-239, June, 1995.
2. S. B. Davis and P. 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.
3. H. Hermansky and N. Morgan, RASTA Processing of Speech, *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, Oct. 1994.
4. C. R. Jankowski Jr., H.-D. H. Vo, R. P. Lippmann, A Comparison of Signal Processing Front Ends for Automatic Word Recognition, *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, July 1995.
5. Chan, Y. T., Lavoie, J. M. M., and Plan, J. B., A Parameter Estimation Approach to Estimation of Frequencies of Sinusoids, *IEEE Transactions on Acoustics, Speech and Signal processing*, vol. 29, no. 2, p. 214-219, April, 1981.
6. B. C. J. Moore and B. R. Glasberg, Suggested formula for calculating auditory-filter bandwidth and excitation patterns, *J. Acoust. Soc. America.*, vol. V74, pp. 750-753, 1983.
7. M. D. Skowronski and J. G. Harris, Human factor cepstral coefficients, submitted to *IEEE Trans. Speech and Audio Processing*.
8. M. Slaney, *Auditory Toolbox, Version 2, Technical Report No: 1998-010*, Internal Research Corporation, 1998.
9. S. J. Young et. al., *The HTK Book*, Entropics Cambridge Research Lab, Cambridge, UK, 1995.
10. 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.
11. S. Furui, On the role of spectral transition for speech perception, *J. Acoustical Society of America*, pp. 1016-1025, 1986.
12. <http://research.compaq.com/downloads.html>.
13. <http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html>.