Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition

Mark D. Skowronski[a] and John G. Harris[b]
Computational Neuro-Engineering Laboratory, Electrical and Computer Engineering, University of Florida, Gainesville, Florida 32611

(Received 29 September 2003; revised 4 June 2004; accepted 8 June 2004)

Mel frequency cepstral coefficients (MFCC) are the most widely used speech features in automatic speech recognition systems, primarily because the coefficients fit well with the assumptions used in hidden Markov models and because of the superior noise robustness of MFCC over alternative feature sets such as linear prediction-based coefficients. The authors have recently introduced human factor cepstral coefficients (HFCC), a modification of MFCC that uses the known relationship between center frequency and critical bandwidth from human psychoacoustics to decouple filter bandwidth from filter spacing. In this work, the authors introduce a variation of HFCC called HFCC-E in which filter bandwidth is linearly scaled in order to investigate the effects of wider filter bandwidth on noise robustness. Experimental results show an increase in signal-to-noise ratio of 7 dB over traditional MFCC algorithms when filter bandwidth increases in HFCC-E. An important attribute of both HFCC and HFCC-E is that the algorithms only differ from MFCC in the filter bank coefficients: increased noise robustness using wider filters is achieved with no additional computational cost. © *2004 Acoustical Society of America.*

[DOI: 10.1121/1.1777872]

PACS numbers: 43.72.Ar, 43.72.Ne, 43.72.Dv [DDO] Pages: 1774–1780

I. INTRODUCTION

Automatic speech recognition (ASR) has not found general widespread use, primarily due to the degraded performance of ASR in noisy environments. The first step in ASR is to form a compact representation of the speech signal, emphasizing phonetic information over variations due to speaker, channel, and background noise sources. The mel frequency cepstral coefficient (MFCC) front end is currently the most common, primarily because MFCC is well-suited for the assumptions of uncorrelated features used in Hidden Markov Models (HMM) to estimate the state distributions and also because of the superior noise robustness of MFCC as compared to linear prediction-based feature extractors.[1] In the work of Jankowski *et al.*, the authors suggest that the filter banks for MFCC, as well as the Seneff and EIH auditory models, are the primary elements of the algorithms which provide noise robustness, while others have shown that the linear prediction algorithm does not accurately model the zeros that additive white noise creates in the speech spectrum.[2] The bandwidth of the triangular filters in MFCC is determined by the *spacing* of filter center frequencies which is a function of sampling rate and the number of filters. That is, if the number of filters in the filter bank increases, the bandwidth of each filter decreases unintentionally.

The authors have previously introduced a novel modification to MFCC called human factor cepstral coefficients (HFCC). In HFCC, filter bandwidth is decoupled from filter spacing and is determined by the *known* relationship between center frequency and critical bandwidth of the human auditory system.[3] Using HFCC, the authors have demonstrated improved robustness over MFCC in various noise environments with filter bandwidth determined by Moore and Glasberg's critical band equivalent rectangular bandwidth (ERB) expression.[4] While HFCC uses a particular expression for filter bandwidth, the important point to note is that HFCC is part of a framework in which filter bandwidth is a free design parameter, opening up an unexplored tangent of research in ASR.

The ability to control bandwidth independent of filter spacing in filter bank design is important for two reasons: (1) it eliminates errors in bandwidth caused by extreme choices in the number of filters or filter bank frequency range, and (2) it allows optimization of bandwidth beyond Moore and Glasberg's ERB expression. Though Moore and Glasberg's ERB expression describes the critical bands of the human auditory system, HFCC is ultimately a preprocessor for an artificial classifier. Different bandwidth expressions may better condition the speech features for use in artificial classifiers. The current work considers a linear scale factor of ERB, called the E-factor, used in HFCC. Since HFCC refers to the algorithm using Moore and Glasberg's original ERB function with unity scaling factor, HFCC with E-factor is referred to as HFCC-E.

Several researchers have investigated feature extraction algorithms for automatic speech recognition that employ various filter bank designs. Hermansky's perceptual linear predictive (PLP) analysis uses a bark-scaled filter bank and cube root compression before estimating linear prediction coefficients.[5] PLP analysis uses engineering approximations to psychophysical laws, such as equal loudness preemphasis and the intensity-to-loudness power law, and admittedly al-

[a] Electronic mail: markskow@cnel.ufl.edu
[b] Electronic mail: harris@cnel.ufl.edu

lows for variations in deriving the auditory spectrum. Chan *et al.* used multiresolution analysis to design a mel-spaced wavelet filter bank.[6] The filter bank design allowed them to include Weiner filtering in the feature extraction process, and improved noise robust speech recognition was reported.

Other researchers have modified MFCC in order to improve noise robust performance. Tchorz and Kollmeier developed an auditory model-based preprocessor, similar to MFCC.[7] The static log compression in MFCC was replaced with adaptive compression. In an isolated digits experiment, the authors showed improved robustness in different noise environments compared to their model with a static log compressor, though clean-speech recognition accuracy using the adaptive compression was lower. Strope and Alwan directly modified the MFCC algorithm by adding an adaptive masker before the discrete cosine transform (DCT).[8] The masker, along with a peak-picking algorithm, improved robustness to additive noise shaped to match the long-term average spectrum of speech in an isolated digits experiment.

Research has also focused on filter bandwidth in MFCC. The 2000 SPINE evaluation of Singh *et al.* reported using filters in MFCC with double the bandwidth while maintaining filter center frequency and number of filters.[9] Improved noise-robust performance was achieved in large-vocabulary experiments, yet MFCC filter bandwidth was not investigated further. Sinha and Umesh showed improvements over baseline MFCC recognition by increasing the number of filters while maintaining the original filter bandwidth, though no explanation for the performance gain was reported.[10] Furthermore, the authors previously demonstrated improved noise robustness in MFCC by changing the *overlap* between adjacent filters.[11]

HFCC, like MFCC or PLP, is not a perceptual model of the human auditory system but rather a biologically inspired feature extraction algorithm. While MFCC and PLP draw from the theories of auditory transduction and perception laid down by psychoacousticians and neurobiologists, the algorithms concomitantly transform the speech into a representation suitable for the artificial classifiers currently used for speech recognition (hidden Markov models, dynamic time warping, or neural networks). The discrete cosine transform, an integral part of MFCC, is not biological. However, the DCT decorrelates the log energy outputs from the filter bank, strengthening the assumption of diagonal covariance matrices typically used for the Gaussian distribution of features in each state of an HMM. In this paper, the decoupled filter bandwidth in HFCC is exploited by increasing the width of the filters beyond their biologically inspired values. The experiments that follow show that HFCC-E better conditions the cepstral features for HMM classification in noisy environments.

The rest of the paper is organized as follows: the MFCC algorithm is presented, followed by a description of HFCC. Next, details for the ASR experiments using MFCC (the original as well as two modern variations) and HFCC with the isolated English digits from the TI-46 corpus are elaborated. A discussion of the ASR results and the effects wider filters have on noise robust performance concludes the paper.

II. MEL FREQUENCY CEPSTRAL COEFFICIENTS

The original MFCC algorithm introduced by Davis and Mermelstein (DM)[12] combined perceptually spaced filters with the DCT (shown to be similar to principal eigenvectors of Dutch vowels)[13] in a mainstream speech processing publication. The algorithm can be summarized as follows: a time signal is windowed, Fourier transformed to the frequency domain, and scaled by a bank of triangular filters, equally spaced on a linear-log frequency axis, and the sum of magnitude coefficients scaled by each filter is log-compressed and transformed via the DCT to cepstral coefficients. The filter bank is comprised of triangular filters, which are a coarse approximation to the shape of the critical band bandpass response of the human auditory system. The base of each triangle is determined by the center frequencies of the adjacent filters; that is, filter bandwidth in MFCC is determined by the frequency range of the filter bank as well as the number of filters in the bank.

However, coupling bandwidth to other filter bank design parameters creates two problems. As ASR experimenters adapt the algorithm to their own desired frequency range (typically set by the sampling rate of the speech corpus under study), they *unintentionally* change filter bandwidth. MFCC came to the forefront of filter bank-based speech feature extractors because of the perceptually motivated filter spacing, yet the well-known relationship between frequency and critical bandwidth of the human auditory system is not incorporated. This shortcoming is rectified in HFCC. The second problem with coupling bandwidth to the number of filters and to frequency range is that filter bandwidth is not subject to optimization with respect to experimental recognition accuracy.

The following section details the filter bank design in HFCC.

III. HUMAN FACTOR CEPSTRAL COEFFICIENT FILTER BANK

Bandwidth in HFCC is a design parameter independent of filter spacing. To determine the bandwidth, HFCC employs Moore and Glasberg's approximation of critical bandwidth, measured in equivalent rectangular bandwidth (ERB):

$$\mathrm{ERB}=6.23f_c^2+93.39f_c+28.52 \quad \mathrm{Hz}, \tag{1}$$

where center frequency f_c is in kHz with the curve fit valid between 0.1 and 6.5 kHz.[14] The ERB of a bandpass filter is the width of a rectangle whose height is the maximum of the filter magnitude response and whose area is the same as the filter response. ERB is used as an alternative to 3 dB points to describe filter bandwidth. Filter center frequencies are equally spaced in mel frequency using Fant's expression relating mel frequency $\hat{f}$ to linear frequency f,[15]

$$\hat{f}=2595\log_{10}\left(1+\frac{f}{700}\right). \tag{2}$$

Complete details of the HFCC filter bank construction can be found in the Appendix.

The authors have shown previously that, using Moore and Glasberg's ERB curve to define filter bandwidth, HFCC

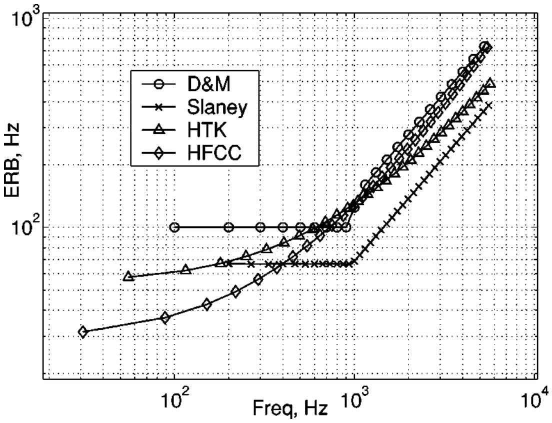

FIG. 1. ERB vs center frequency for HFCC and various MFCC implementations. Each marker represents the center frequency of a filter.

shows improved noise robustness in automatic speech recognition experiments over MFCC.[3] The current work exploits the decoupled filter bandwidth in HFCC by investigating the effects of linearly scaling the ERB bandwidth defined in Eq. (1) on noise robustness.

IV. EXPERIMENTS AND RESULTS

Filter bandwidth in HFCC-E is scaled by multiplying the ERB expression in Eq. (1) by a constant that called the ERB scale factor, or E-factor. The performance characteristics of HFCC-E with various E-factors are determined using ASR experiments in noisy environments. As a standard, three variations of MFCC common in current ASR systems are included. The next section briefly describes the three versions of MFCC before detailing the experiments.

A. Variations of MFCC

The description of the filter bank in DM's original paper has been interpreted differently by researchers using the algorithm. In the experiments, two popular versions commonly used in ASR are included along with DM's original algorithm for direct comparison with HFCC-E. The first variation is from Slaney's Auditory Toolbox for Matlab.[16] The function MFCC.m from the toolbox uses twice as many filters as in DM with a smaller spacing between center frequencies and a wider frequency range. Filters are spaced linearly below 1 kHz and log-spaced above 1 kHz. All triangular filters are equal in area.

The second variation of MFCC is from the Cambridge HMM Toolkit (HTK),[17] which is a popular C library for implementing HMMs for large vocabulary continuous speech recognition. The HTK function has many design parameters (number of filters, frequency range, vocal tract length normalization, the use of signal magnitude or square magnitude)—Young's paper is referenced for parameter values.[17] The HTK version uses Fant's definition of mel frequency, while DM and Slaney use a linear–log approximation. All filters in HTK and DM are the same height. Figure 1 summarizes the bandwidth of filters used in the three variations of MFCC and HFCC (unity E-factor), and Table I summarizes the filter bank parameters for MFCC and HFCC at a 12.5 kHz sampling rate.

TABLE I. Summary of filter bank parameters.

Method	f_{range}, Hz	# filters	Spacing	Amplitude
DM	0–6063	22	lin–log	Equal height
Slaney	133.3–5973	38	lin–log	Equal area
HTK	0–6179	29	mel warp	Equal height
HFCC	0–6250	29	mel warp	Equal height

For all three implementations of MFCC, care is taken *to preserve the original filter bandwidth* while accounting for the sampling rate of 12.5 kHz used in the reported ASR experiments. For example, DM's original sampling rate was 10 kHz.[12] They used 10 filters with center frequencies equally spaced between 100 and 1000 Hz. Above 1 kHz, center frequencies were spaced five per octave (scaling factor of $2^{1/5} \approx 1.149$) for a total of 20 filters. The next few center frequencies above 4 kHz are $4k \times 2^{1/5} \approx 4.595$ kHz, $4k \times 2^{2/5} \approx 5.278$ kHz, $4k \times 2^{3/5} \approx 6.063$ kHz, and $4k \times 2^{4/5} \approx 6.964$ kHz. The last frequency is above the Nyquist rate of 6.25 kHz, so the third frequency represents the upper frequency of the last filter. Thus, two filters are added to the DM filter bank, resulting in 22 filters.

B. TI-46 digits experiment

To characterize the performance of HFCC-E, features are extracted from the isolated English digits "zero"–"nine" from the TI-46 speech corpus in noisy environments (white and pink noise from the Noisex92 database).[18] This speech corpus offers an interesting vocabulary while requiring a classifier of only modest complexity. Clean speech recognition is nearperfect, meaning the classifiers used are adequate for the task at hand. The purpose of the experiments in this work is to demonstrate the robustness of the various features to additive noise, and larger speech databases (Aurora, Switchboard) would require further techniques whose performance would obscure the overall recognition score (endpoint detection/word spotting, prosody effects, bigram/trigram modelling, out-of-vocabulary handling, language modelling, conversational speech idiosyncrasies).

Features are extracted from an utterance by analyzing the speech with 20 ms frames at 100 frames/s. Each frame is Hamming windowed, filtered by a first-order pre-emphasis filter ($\alpha = 0.95$), and the magnitude spectrum is computed and scaled by the triangular filter bank. The sum of samples scaled by each filter in the bank (output energy) is then log-compressed and transformed via the DCT to cepstral coefficients. Thirteen coefficients per frame are computed. The first cepstral coefficient is replaced by the log of the Parseval energy of the speech frame. Cepstral mean subtraction is applied,[19] and delta coefficients (±4 frames) are computed and appended to the cepstral coefficients (26 coefficients total).

HMM word models with eight states and diagonal covariance matrices are used to classify the digits. For each of eight trials, the models are trained with clean speech from seven of the eight male speakers from the TI-46 corpus and tested with noisy speech from the remaining male speaker

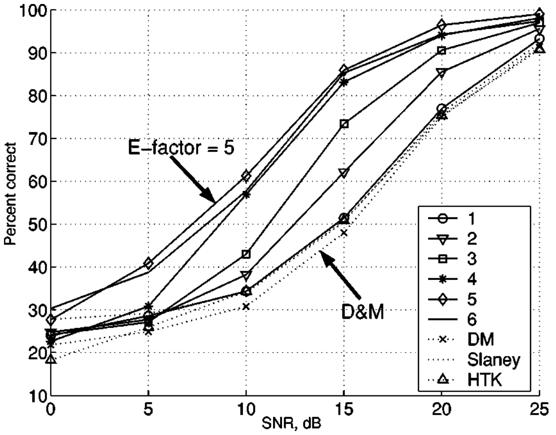

FIG. 2. Recognition accuracy vs white noise global SNR, averaged over eight trials. E-factors range from 1 to 6, with highest recognition at E-factor=5. See Table II for error bars.

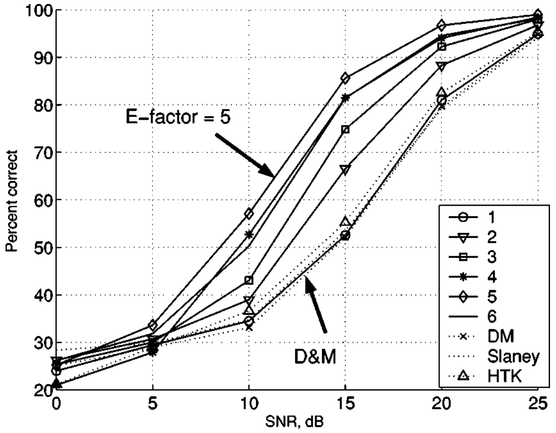

FIG. 3. Recognition accuracy vs pink noise global SNR, averaged over eight trials. E-factors range from 1 to 6, with highest recognition at E-factor=5. See Table III for error bars.

(speaker-independent experiment). The experiment trial is repeated eight times so that each of the eight male speakers is the test speaker. Preliminary experiments showed that male and female speakers are highly separable, so an accurate gender detector is assumed to be available for these experiments.[20] Separate sets of models using male and female speakers are trained and tested on the appropriate set of models after gender has been (ideally) determined. For brevity, only the results from male speakers are reported. Results using female speakers are similar, though the absolute recognition accuracy is lower for all feature extractors.

Figure 2 shows recognition accuracy vs global SNR for white noise and MFCC along with HFCC-E with various E-factors, while Table II shows the same results relative to DM for each trial. The relative results reduce the variation in recognition accuracy due to differences among test speakers. At 15 dB SNR, HFCC-E with an E-factor of 5 shows a recognition accuracy 38±4 percentage points higher than DM. In the transition region between 40% and 80% correct recognition, the curve for DM is about 7 dB to the right of the curve for HFCC-E with and E-factor of 5. That is, HFCC-E with an E-factor of 5 improves recognition by about 7 dB SNR over DM.

Figure 3 and Table III show the same results for pink noise. At 15 dB SNR, HFCC-E with an E-factor of 5 shows a recognition accuracy 33±3 percentage points higher than DM. In the transition region between 40% and 80% correct recognition, the curve for DM is about 6 dB to the right of the curve for HFCC-E with and E-factor of 5.

Figure 4 shows the relative performance of HFCC-E compared to DM at 15 dB SNR for several E-factors and for both white and pink additive noise. Plots for both noise sources peak at an E-factor of 5.

V. DISCUSSION

In both white and pink noise, recognition results at a given SNR increase as the E-factor increases from 1 to 5. Beyond an E-factor of 5, the results decrease. This trade-off can be better understood by considering the effects of filter bandwidth on frame SNR as well as the correlation among the filter outputs.

A. Frame SNR

Frame SNR for the *i*th frame of noisy speech is defined as follows:

$$\text{frame SNR} \equiv 10 \log_{10}\left(\frac{E_X(i)}{E_N(i)}\right) \text{ dB}, \tag{3}$$

where $E_X(i)$ and $E_N(i)$ are the sum of output energies for all M filters for speech and noise, respectively, of the *i*th frame,

TABLE II. Relative performance of HFCC-E and MFCC in white noise averaged over eight trials: mean ±95% confidence interval of percentage-point difference relative to DM. E-factor ranges from 1 to 6. Global SNR of 5–20 dB. Largest difference in bold.

	SNR=5	10	15	20 dB
E-factor=1	3.7±1	3.5±1	3.5±3	1.9±2
2	2.8±2	7.4±2	14.1±5	10.4±2
3	2.3±2	12.2±3	25.4±6	15.5±5
4	5.9±4	26.1±7	35.2±6	19.0±4
5	**16.0±4**	**30.4±6**	**38.0±4**	**21.4±6**
6	13.9±2	26.9±4	37.3±3	19.2±6
Slaney	4.3±1	3.1±1	2.5±3	1.0±2
HTK	1.1±3	3.5±2	2.9±4	0.3±2

TABLE III. Relative performance of HFCC-E and MFCC in pink noise averaged over eight trials: mean ±95% confidence interval of percentage-point difference relative to DM. E-factor ranges from 1 to 6. Global SNR of 5−20 dB. Largest difference in bold. Results at 15 dB SNR plotted in Fig. 4.

	SNR=5	10	15	20 dB
E-factor=1	0.5± 0.6	1.4±2	0.4±1	1.3±2
2	1.7± 1	5.9±2	14.4±4	8.6±2
3	1.1± 1	9.9±3	22.8±5	12.6±4
4	−1.1± 2	19.6±7	29.4±5	14.4±4
5	**4.7± 3**	**24.0±7**	**33.5±3**	**17.1±6**
6	2.9± 2	17.2±4	29.3±4	14.9±5
Slaney	0.9± 0.7	1.7±1	1.3±2	−0.4±2
HTK	0± 2	3.6±1	3.1±1	0.5±2

$$E_X(i)=\frac{1}{M}\sum_{m=1}^{M}\sum_{k\in K_m}A_m(k)|X_i(k)|,$$
$$E_N(i)=\frac{1}{M}\sum_{m=1}^{M}\sum_{k\in K_m}A_m(k)|N_i(k)|. \tag{4}$$

In Eq. (4), M is the number of filters in the filter bank, A_m is the triangular function of the mth filter, K_m is the domain of the mth filter, X_i and N_i are the DFT coefficients of the speech and noise signals of the ith frame of noisy speech, respectively, and k is the DFT index.

Frame SNR is the ratio of average speech output to average noise output, where averaging is performed over the M filter outputs for a frame of noisy speech. By contrast, global SNR is defined as the ratio of the signal energy to the noise energy over the duration of the entire utterance. Figure 5 shows the frame SNR of the utterance "seven" for several E-factors.

For a fixed global SNR, Fig. 5 shows a steady *increase* in frame SNR in the voiced regions (/EH/, /v/, and /N/) as E-factor increases, while the opposite trend occurs for the unvoiced /s/. Wider filters pass more signal energy as well as noise energy, but the local SNR of each filter only increases as bandwidth increases if the *ratio* of the added signal to added noise is greater than the local SNR of the narrower filter. Since voiced speech is peaky relative to unvoiced speech due to the fundamental harmonics, wider filters tend to pass more harmonics, which tend to have much more signal energy than noise energy. Hence, local SNR of voiced speech increases for increasing E-factor. For the unvoiced /s/, no harmonic peaks are available to increase the added signal to added noise ratio. For the case of the utterance in Fig. 5, the added noise dominates the local SNR calculation.

In summary, wider filters increase frame SNR for voiced frames of speech while they also decrease frame SNR in other parts of the utterance (naturally, since the global SNR is fixed). Thus, wider filters emphasize higher-energy regions of the speech signal over lower-energy regions of speech, which are typically the first regions of speech corrupted by additive noise. Recognition performance in noisy environments relies primarily on high-energy peaks of the speech signal.[8] The noise robustness of HFCC-E is expected to increase as E-factor increases since wider filters increase frame SNR for voiced frames, and voiced frames typically contain high-energy peaks.

B. Pairwise correlation coefficient

The limiting factor concerning wider filters is the fact that neighboring filters overlap more. That is, as the E-factor increases in HFCC, the correlation of filter outputs increases and adjacent filters operate on increasingly similar regions of the speech spectrum. In the limit where each filter spans the

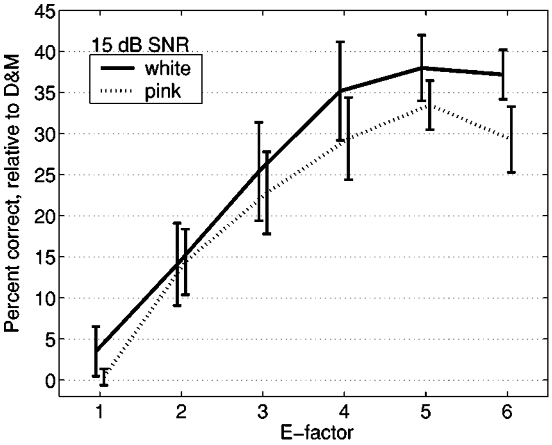

FIG. 4. Relative performance of HFCC-E compared to DM at 15 dB SNR, averaged over eight trials: mean ±95% confidence interval of percentage-point difference relative to DM. E-factor ranges from 1 to 6 for both white and pink additive noise.

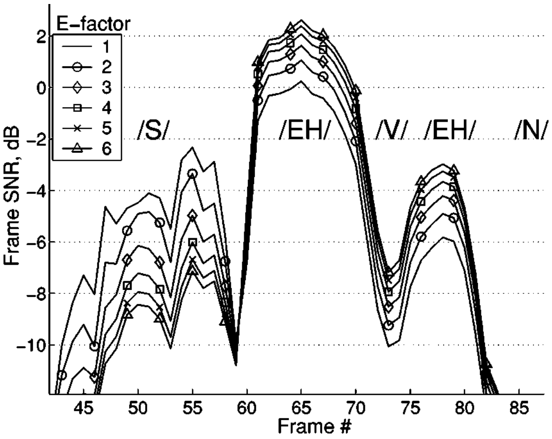

FIG. 5. Frame SNR for a typical utterance of the word "seven" for several E-factors. The phonetic labels denote the regions of each phoneme in the utterance. Frame SNR increases in voiced regions as E-factor increases, the opposite trend is true for the unvoiced /s/ phoneme.

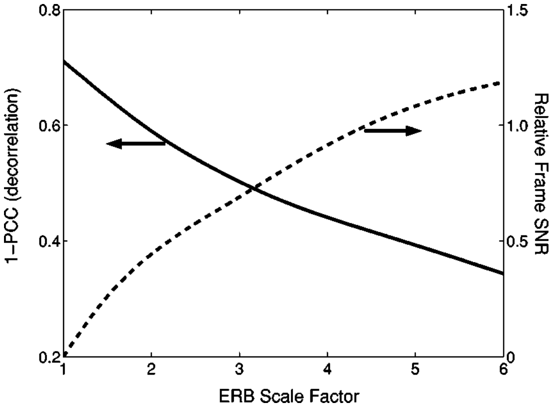

FIG. 6. Left axis: 1-PCC of filter outputs (decorrelation) vs E-factor for clean speech. Right axis: frame SNR, relative to unity E-factor, for the voiced regions in Fig. 5.

entire frequency range of interest, all filters are identical, operate on the entire speech spectrum, and output the same signal (same spectral information). Increased correlation represents a loss of spectral information, and this loss is quantified by calculating the pairwise correlation coefficient (PCC) over all pairs of filter outputs for clean speech.

PCC is defined as the average of the off-diagonal terms of the correlation coefficient matrix R_{pq}:

$$\mathrm{PCC} \equiv \frac{1}{M(M-1)} \sum_{\substack{p=1,\\ p\neq q}}^{M} \sum_{q=1}^{M} R_{pq}, \tag{5}$$

where

$$R_{pq} = \frac{C_{pq}}{\sqrt{C_{pp}C_{qq}}} \tag{6}$$

and C_{pq} is the covariance of the pth and qth filter outputs. For uncorrelated filter outputs, $R_{pq}=0$, while $R_{pq}=1$ when the outputs are identical. Thus, PCC is bounded between [0,1].

The left axis of Fig. 6 shows the expression 1-PCC, indicating that the filters become less *decorrelated* as E-factor increases. Using the same utterance of ‘‘seven’’ as in Fig. 5, Fig. 6 depicts the trade-off between the increase in frame SNR and the decrease in decorrelation as E-factor increases in HFCC-E. Both high frame SNR and high decorrelation are desired, and the E-factor is tuned to balance these two objectives for a particular experiment.

VI. CONCLUSIONS

The current work demonstrates the positive benefits of decoupling filter bandwidth from the number of filters used in the HFCC filter bank. HFCC provides a new avenue of research in feature extraction in that filter bandwidth can now be treated as an independent parameter to be optimized. Furthermore, the current work investigates one such possibility by introducing a linear ERB scale factor to increase the bandwidth of filters used in HFCC. The wider filters in HFCC-E provide increased noise robustness over the traditional MFCC algorithm: for white noise, 38 percentage points higher, or 7 dB increased SNR, and for pink noise, the increase is 33 percentage points, or 6 dB increased SNR. HFCC-E allows control over the tradeoff between frame SNR and decorrelation of the filter outputs for a particular application.

Untapped lines of research using HFCC abound: different bandwidth expressions besides ERB, such as approaches due to Patterson[21] or Zwicker,[22] adaptive E-factors, multiple classifiers trained on various E-factors with an intelligent switching mechanism during classification, filters that adaptively move to regions of high signal energy to improve local SNR, vocal tract normalization by scaling filter center frequencies, wavelet-like filters that vary in frequency *and* time to control decorrelation in both domains.

An important practical characteristic of HFCC is its simple elegance: HFCC and HFCC-E are nearly the same as MFCC, except for the coefficients that describe the filter banks. Existing MFCC code can be easily changed to HFCC or HFCC-E by simply replacing the filter bank coefficients. No additional overhead or computation is necessary for implementation.

APPENDIX A: ERB OF THE TRIANGLE FILTER

Equivalent rectangular bandwidth (ERB) is used as an alternative to the 3 dB point when describing the bandwidth of a filter. ERB is defined as the following:

$$\mathrm{ERB} = \frac{\int |H(f)|^2\, df}{|H(f_c)|^2}, \tag{A1}$$

where $|H(f)|$ is the amplitude of the filter transfer function with peak amplitude at f_c. For HFCC and all MFCC implementations, the triangular filter bank is defined such that $|H(f)|^2$ has a triangular passband. Define f_l, f_c, and f_h as the low, center, and high frequencies of the triangular filter in linear frequency. Without loss of generality, assume $|H(f_c)| = 1$. Then

$$\mathrm{ERB} = \int_{f_l}^{f_c} \frac{f-f_l}{f_c-f_l}\, df + \int_{f_c}^{f_h} \frac{f-f_h}{f_c-f_h}\, df = \frac{1}{2}(f_h - f_l). \tag{A2}$$

APPENDIX B: FINDING THE CENTER FREQUENCIES OF THE FIRST AND LAST FILTERS

After specifying the frequency range of the filter bank between $f_{\min}$ and $f_{\max}$, the center frequencies of the first and last filters are determined by these limits. Define f_{l_i}, f_{c_i}, and f_{h_i} as the low, center, and high frequencies for the ith filter in linear frequency, respectively. The triangular filters are equilateral in mel frequency $\hat{f}$. That is,

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

Unwarping to linear frequency using Eq. (2) yields

$$\begin{aligned} \log\left(1+\frac{f_{c_i}}{700}\right) &= \frac{1}{2}\left(\log\left(1+\frac{f_{h_i}}{700}\right) + \log\left(1+\frac{f_{l_i}}{700}\right)\right), \\ \log\left(1+\frac{f_{c_i}}{700}\right) &= \frac{1}{2}\log\left[\left(1+\frac{f_{h_i}}{700}\right)\left(1+\frac{f_{l_i}}{700}\right)\right], \qquad \text{(B1)} \\ (700+f_{c_i})^2 &= (700+f_{h_i})(700+f_{l_i}), \end{aligned}$$

where, given $f_{l_1}=f_{\min}$ or $f_{h_N}=f_{\max}$, f_{h_1} or f_{l_N} can be solved for in terms of f_{c_1} or f_{c_N} respectively. Note: all linear frequencies are in Hz. From Eq. (A2) and Eq. (1),

$$af_{c_i}^2+bf_{c_i}+c=\tfrac{1}{2}(f_{h_i}-f_{l_i}) \tag{B2}$$

for f_{c_i} in Hz, where the parameters for Moore and Glasberg's ERB expression are

$$a=6.23\times 10^{-6}, \quad b=93.39\times 10^{-3}, \quad \mathrm{c}=28.52. \tag{B3}$$

Given $f_{\min}$ or $f_{\max}$, Eq. (B1) and Eq. (B2) can be written as

$$af_{c_i}^2+bf_{c_i}+c=\hat{a}f_{c_i}^2+\hat{b}f_{c_i}+\hat{c} \tag{B4}$$

or

$$f_{c_i}^2+\overline{b}f_{c_i}+\overline{c}=0, \tag{B5}$$

where

$$\overline{b}=\frac{b-\hat{b}}{a-\hat{a}}, \quad \overline{\mathrm{c}}=\frac{\mathrm{c}-\hat{\mathrm{c}}}{\mathrm{a}-\hat{\mathrm{a}}}. \tag{B6}$$

For the first filter ($f_{l_1}=f_{\min}$),

$$\hat{a}=\frac{1}{2}\,\frac{1}{700+f_{\min}}, \quad \hat{b}=\frac{700}{700+f_{\min}},$$

$$\hat{c}=-\frac{f_{\min}}{2}\left(1+\frac{700}{700+f_{\min}}\right), \tag{B7}$$

and, for the last filter ($f_{h_N}=f_{\max}$),

$$\hat{a}=-\frac{1}{2}\,\frac{1}{700+f_{\max}}, \quad \hat{b}=-\frac{700}{700+f_{\max}},$$

$$\hat{c}=\frac{f_{\max}}{2}\left(1+\frac{700}{700+f_{\max}}\right). \tag{B8}$$

Then

$$f_{c_i}=\tfrac{1}{2}(-\overline{b}\pm\sqrt{\overline{b}^2-4\overline{c}}), \tag{B9}$$

where only a + sign leads to meaningful center frequencies.

APPENDIX C: FINDING THE UPPER AND LOWER FILTER FREQUENCIES GIVEN THE CENTER FREQUENCY

From Eq. (A2),

$$f_{h_i}=f_{l_i}+2\,\mathrm{ERB}_i, \tag{C1}$$

where ERB_i (and f_{c_i}) have been determined. Equation (B1) yields

$$(700+f_{c_i})^2=(700+f_{l_i}+2\,\mathrm{ERB}_i)(700+f_{l_i}) \tag{C2}$$

which is quadratic in f_{l_i}. Solving the quadratic for f_{l_i} yields

$$f_{l_i}=-(700+\mathrm{ERB}_i) \pm\sqrt{(700+\mathrm{ERB}_i)^2+f_{c_i}(f_{c_i}+1400)}, \tag{C3}$$

where only the + sign leads to meaningful results. Equation (C1) is then used to find f_{h_i}.

[1] C. R. Jankowski, H. D. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," IEEE Trans. Speech Audio Process. **3**, 1 (1995).

[2] Y. T. Chan, J. M. M. Lavoie, and J. B. Plan, "A parameter estimation approach to estimation of frequencies of sinusoids," IEEE Trans. Acoust., Speech, Signal Process. **29**, 214–219 (1981).

[3] M. D. Skowronski and J. G. Harris, "Human factor cepstral coefficients," J. Acoust. Soc. Am. **112**, 2279 (2002).

[4] M. D. Skowronski and J. G. Harris, "Improving the filter bank of a classic speech feature extraction algorithm," *International Symposium on Circulatory Systems* (IEEE, Bangkok, Thailand, 2003), Vol. IV, pp. 281–284.

[5] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," J. Acoust. Soc. Am. **87**, 1738–1752 (1990).

[6] C. P. Chan, P. C. Ching, and T. Lee, "Noisy speech recognition using de-noised multiresolution analysis acoustic features," J. Acoust. Soc. Am. **110**, 2567–2574 (2001).

[7] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**, 2040–2050 (1999).

[8] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," IEEE Trans. Speech Audio Process. **5**, 451–464 (1997).

[9] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination," *International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Salt Lake City, UT, 2001), pp. 273–276.

[10] R. Sinha and S. Umesh, "Non-uniform scaling based speaker normalization," *International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Orlando, FL, 2002), pp. 589–592.

[11] M. D. Skowronski and J. G. Harris, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," *International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Orlando, FL, 2002), pp. 801–804.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process. **28**, 357–366 (1980).

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

[14] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753 (1983).

[15] C. G. M. Fant, "Acoustic description and classification of phonetic units," Ericsson Technics **15**, 1 (1959); reprinted in *Speech Sound and Features*, ISBN 0262060515 (MIT Press, Cambridge, 1973).

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

[17] S. Young, J. Jansen, J. Odell, D. Ollasen, and P. Woodland, *The HTK Book (version 2.0)* (Entropics Cambridge Research Lab, Cambridge, UK, 1995).

[18] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Technical report, Speech Research Unit, Defense Research Agency, Malvern, U.K. (unpublished), http://spib.rice.edu/spib/select_noise.html (2004).

[19] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am. **55**, 1304–1312 (1974).

[20] A. A. Dibazar, J.-S. Liaw, and T. W. Berger, "Automatic gender identification," J. Acoust. Soc. Am. **109**, 2316 (2001).

[21] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. **59**, 640–654 (1976).

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer Verlag, Berlin, 1991).