

Research Article

An FFT-Based Companding Front End for Noise-Robust Automatic Speech Recognition

Bhiksha Raj,¹ Lorenzo Turicchia,² Bent Schmidt-Nielsen,¹ and Rahul Sarpeshkar²

¹Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139-4307, USA

²Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Received 29 November 2006; Revised 14 March 2007; Accepted 23 April 2007

Recommended by Stephen Voran

We describe an FFT-based companding algorithm for preprocessing speech before recognition. The algorithm mimics tone-to-tone suppression and masking in the auditory system to improve automatic speech recognition performance in noise. Moreover, it is also very computationally efficient and suited to digital implementations due to its use of the FFT. In an automotive digits recognition task with the CU-Move database recorded in real environmental noise, the algorithm improves the relative word error by 12.5% at -5 dB signal-to-noise ratio (SNR) and by 6.2% across all SNRs (-5 dB SNR to $+15$ dB SNR). In the Aurora-2 database recorded with artificially added noise in several environments, the algorithm improves the relative word error rate in almost all situations.

Copyright © 2007 Bhiksha Raj et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The performance of humans on speech recognition tasks in noise is extraordinary compared to state-of-the-art automatic speech recognition (ASR) systems [1]. One explanation is that the brain has amazing pattern recognition abilities not well captured by ASR systems. Additionally, the auditory periphery has sophisticated signal representations which are highly robust to noise. While the upper cognitive processes that are brought to bear on speech recognition tasks are not well understood and cannot be emulated, the human peripheral auditory system has been well studied and several of the processes in it are well understood (e.g., [2]), and can be mathematically modeled [3–5]. It may be expected that by simulating some of the processes in the peripheral auditory system within the signal processing schemes employed by a speech recognizer, its robustness to noise may be improved. Following this hypothesis, in this paper, we will focus on the benefits of a front end inspired by the peripheral auditory system for improving the performance of ASR systems in noise.

The procedure by which the peripheral auditory system captures sound pressure waves in a format that can be forwarded to the higher levels of the auditory pathway includes various processes that are analogous to automatic gain control, critical band analysis, equal loudness preemphasis, two-

tone suppression, forward and backward masking, half-wave rectification, envelope detection, and so forth [2].

Several very detailed models of the peripheral auditory system have been proposed in the literature that attempt to mathematically model all the known processes within it in detail, for example, see [3–7]. Some of these models have also been applied to the problem of deriving “feature representations” for automatic speech recognition systems. While these models were found to perform comparably with a speech-recognition system implemented with conventional feature-computation schemes, namely, Mel-filterbank-based cepstral analysis [8], in general the additional gains to be derived from them have not been commensurated with the greatly increased computation required by these models.

The human auditory system incorporates many different phenomena. Some of these specifically aid perception. Others are either simply incidental to the construction and physics of the auditory system, or have other purposes. The more successful trend in anthropomorphic signal processing for speech recognition has been to model specific auditory phenomena that are hypothesized to relate directly to the noise robustness of human perception, rather than the entire auditory process. Davis and Mermelstein [9] demonstrated the effectiveness of modeling critical band response in the computation of cepstral front ends for speech recognition. Critical band response is modeled in the signal processing

schemes employed by almost all current speech recognition. The PLP features proposed by Hermansky [10] also incorporate equal-loudness preemphasis and root compression, and this has been observed to improve noise robustness. Extrapolating from these results, it may be valid to hypothesize that critical band response and equal-loudness compression also contribute to the noise robustness of human perception. Indeed, one may turn the argument around and speculate that improvements in noise robustness of computational models of speech recognition may provide evidence that the modeled perceptual phenomenon contributes to noise robustness in perception.

A well-known psychoacoustic phenomenon that may be related to the noise robustness of human perception is *masking*, an auditory phenomenon whereby high-energy frequencies mask out adjacent lower-energy frequencies. The peripheral auditory system exhibits a variety of masking phenomena. *Temporal* masking is a phenomenon whereby high-energy sounds mask out lower-energy sounds immediately preceding or succeeding them. *Simultaneous* masking is a phenomenon whereby high-energy frequencies mask out adjacent, *concurrent*, lower-energy frequencies.

Computational analogues for *temporal* masking have previously been presented by Strobe and Alwan [11] and Holmberg et al. [12], among others. Tchorz and Kollmeier [13] and Hermansky and Morgan [14] compress and filter the effective envelope of the output of a critical-band filterbank, a procedure that also has the incidental effect that high-energy sounds partially mask adjacent (in time) to low-energy acoustic phenomena. These methods have all been observed to improve noise robustness of ASR, indicating that the phenomenon of temporal masking aids in noise-robust audition.

In this paper, we present a computational model that achieves *simultaneous* masking by mimicking the phenomenon of *two-tone suppression*. Two-tone suppression is a nonlinear phenomenon observed in the biological cochlea [2], whereby the presence of one tone suppresses the frequency response of another tone that is near to it in frequency. The origin of this effect is likely to involve saturating amplification in the outer hair cells of the cochlea. At the psychoacoustic level, two-tone suppression manifests itself as simultaneous masking, defined by the American Standards Association (ASA) as the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound [15].

In [16], we reported a cochlear model with traveling-wave amplification and distributed gain control that exhibits two-tone suppression. In a follow-up publication [17], we described a bioinspired companding algorithm that mimicked two-tone suppression in a highly programmable filterbank architecture. The companding algorithm filters an incoming signal by a bank of broad filters, compresses their outputs by their estimated instantaneous RMS value, refilters the compressed signals by a bank of narrow filters, and finally expands them again by their instantaneous RMS values. As we will explain in Section 2, this processing has the effect of retaining spectral peaks almost unchanged, whereas

frequencies adjacent to spectral peaks are suppressed, resulting in two-tone suppression. An *emergent* property of the companding algorithm is that it enhances spectral contrast and naturally emphasizes high signal-to-noise-ratio spectral channels while suppressing channels with a lower signal-to-noise ratio. Consequently, we suggested the algorithm's potential benefit for improving ASR in noise in [17]. This algorithm has since also been verified to improve significantly the intelligibility of the processed signal, both in simulations of cochlear implants [18–20], and for real cochlear implant patients [19, 21, 22].

In [23] we showed that significant improvement in recognition accuracy can be obtained, particularly at very low SNRs, using a digital simulation of the analog implementation of the proposed companding algorithm. Between the results of [18–21, 23], it is evident that two-tone suppression is important for noise-robust perception. However, the implementation in [23] models additional details such as an analog filterbank based on critical-band analysis. Such an implementation, while suitable for implementation in low-power analog VLSI (which was the original purpose of the design of the algorithm) is, however, highly inefficient for a real-time recognizer that functions entirely on digitized signals. Additionally, it does not determine whether two-tone suppression by itself is important or if it must go in conjunction with critical-band analysis—the results are insufficient to determine which components of the systems are critical and which are incidental to the implementation. In this paper, we build on this prior work by developing an FFT version of the companding algorithm for implementation in the signal processing front end of an ASR system. The FFT-based algorithm presented here does not mimic the two-tone suppression of [23] in its entirety—rather it is an engineering approximation that retains the specific mechanism, that is, the companding architecture that results in two-tone suppression, while eliminating other characteristics such as auditory filterbanks and time-domain processing. Nevertheless, the algorithm is observed to improve speech recognition performance in most situations, indicating that the mere presence of two-tone suppression by itself is important for noise robustness. Additionally, the greatly improved computational efficiency of the FFT version makes it practical for real-time ASR systems.

It is worth emphasizing that the companding algorithm simply mimics tone-to-tone suppression and masking in the auditory system; spectral-contrast enhancement emerges as a consequence, and perception in noise is improved. Other work that *explicitly* tries to enhance spectral contrast in the signal has also shown benefits for improving speech perception in noise: Stone and Moore proposed an analog device for spectral contrast enhancement in hearing aids [24]. Later work from members of the same group [25] showed that a digital spectral-contrast-enhancement algorithm yielded a modest but significant improvement of speech perception in noise for hearing-impaired listeners. Similarly, the peak-isolation mechanism of [11], based on raised-sine cepstral liftering [26], enhanced spectral contrast and revealed its benefit for ASR.

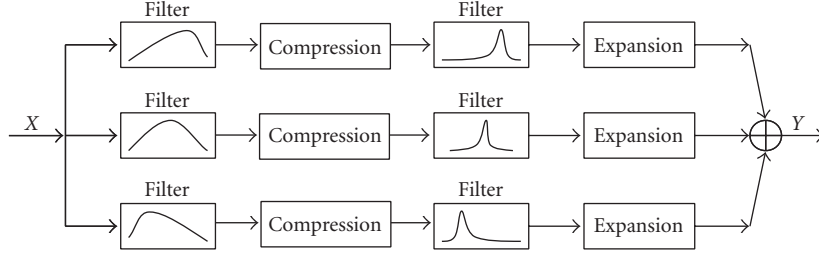


FIGURE 1: Block diagram of our companding strategy.

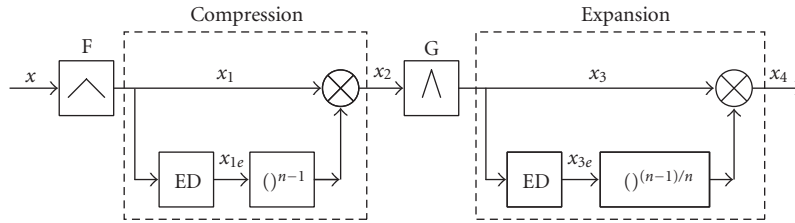


FIGURE 2: A detailed view of a single channel of processing in Figure 1.

In Section 2, we review the companding algorithm as it was first described in [17], as a filterbank implementation. In Section 3, we describe the new FFT-based companding algorithm. In Section 4, we report experimental results from an HMM-based ASR system that uses an FFT-based companding front end.

Signal processing schemes often improve recognition performance in “mismatched” conditions, that is, when the recognizer has been trained on clean speech but the data to be recognized are noisy; yet they may fail to improve performance when the training data are similar to the test data, a more realistic situation for most applications. They also often suffer the drawback that while they may result in significant improvements on speech that has been corrupted by digital addition of noise, they fail to deliver similar improvements on genuine noisy recordings. Further, it is common experience that the recognition performance obtained on noisy speech with systems that have been trained on noisy speech is generally better than that obtained on denoised noisy speech using systems that have been trained on clean speech [27]. The experiments reported in this paper have therefore been conducted both with real-world recordings from the CU-Move database [28], an extensive database of speech digits recorded in moving cars, and on Aurora-2 [29], a smaller database of speech recordings that have been artificially corrupted by digital addition of noises of various types. Experiments have been conducted under both mismatched and matched conditions.

In Section 5, we conclude by summarizing the main findings of our paper. We note that improvements have been obtained in all conditions, for almost all noise types. Thus our observed improvements can be expected over to carry to real-world scenarios.

2. FILTER-BASED COMPANDING

In this section, we review the companding algorithm that mimics two-tone suppression [17]. The strategy uses a non-coupled filterbank and compression-expansion blocks as shown in Figures 1 and 2. Every channel in the companding architecture has a relatively broadband prefilter, followed by a compression block, a relatively narrowband postfilter, and finally an expansion block. The prefilter and postfilter in every channel have the same resonant frequency. The resonant frequencies of the various channels are logarithmically spaced and span the desired spectral range. Finally, the channel outputs of this nonlinear filterbank are summed to generate an output with enhanced spectral peaks. Alternately, they may be used without summation, and features may be directly computed from the expander output.

The broadband prefilter determines the set of frequencies in a channel that are allowed to affect the gain of the compressor. The compressor consists of an envelope detector, a nonlinearity, and a multiplier. The output of the envelope detector x_{1e} , which we denote by $\text{AMP}(x_1)$, represents the amplitude of x_1 , the output of the broadband prefilter. The nonlinearity raises the envelope to a power $(n - 1)$. As a result, the amplitude of x_2 , the output of the multiplier, is approximately $\text{AMP}(x_1)^n$. If n is less than one, this results in a compression of the output of the broadband prefilter.

The narrowband postfilter selects only a narrower subset of the frequencies that are allowed by the prefilter. The expander is similar to the compressor and also consists of an envelope detector, a nonlinearity, and a multiplier. The output of the envelope detector x_{3e} represents the amplitude of x_3 , the output of the postfilter. The nonlinearity raises the envelope to a power $(1 - n)/n$. Consequently, the amplitude of

x_4 , the output of the multiplier, is approximately $\text{AMP}(x_3)^{1/n}$. If n is less than one, this results in an expansion of the output of the narrowband postfilter.

Consider the case where the input to a channel, x , consists chiefly of a tone $a \cos(\omega_1 t)$ at the resonant frequency ω_1 for the channel. The broadband prefilter permits the tone through unchanged, that is, $x_1 = a \cos(\omega_1 t)$ (assuming a unit gain, zero phase filter) and $x_2 = a^n \cos(\omega_1 t)$. The narrowband postfilter, having a resonant frequency identical to the prefilter, also permits the tone. Hence, the amplitude of the output of the postfilter is the same as the amplitude of the output of the compressor, that is, $x_3 = a^n \cos(\omega_1 t)$. The amplitude of the final output of the channel x_4 is $\text{AMP}(x_3)^{1/n} = a$, that is, $x_4 = a \cos(\omega_1 t)$. Thus the channel has no effect on the overall level of an isolated tone at the resonant frequency.

Now, consider the case where the input to the channel is the sum of a tone at the resonant frequency ω_1 of the channel, and a second tone with higher energy at an adjacent frequency ω_2 , such that ω_2 lies within the bandwidth of the broadband prefilter, but outside that of the narrowband postfilter, that is, $x = a \cos(\omega_1 t) + ka \cos(\omega_2 t)$, where the amplitude of the second sinusoid is k times that of the first. Assuming that the broadband filter permits both tones without modification, $x_1 \simeq a \cos(\omega_1 t) + ka \cos(\omega_2 t)$. As an extreme case, we consider $k \gg 1$. The amplitude of x_1 is approximately ka , and $x_2 \simeq k^{(n-1)} a^n \cos(\omega_1 t) + k^n a^n \cos(\omega_2 t)$. The narrowband postfilter does not permit ω_2 , hence $x_3 = k^{(n-1)} a^n \cos(\omega_1 t)$. The expander expands the signal by the amplitude of x_3 , leading to $x_4 = k^{(n-1)/n} a \cos(\omega_1 t)$, that is, the output of the channel is the tone at the resonant frequency, scaled by a factor $k^{(n-1)/n}$. Since $k > 1$ and $n < 1$, $k^{(n-1)/n} < 1$, that is, the companding algorithm results in a suppression of the tone at the center frequency of the channel. The greater the energy of the adjacent tone at ω_2 , that is, the larger the value of k , the greater the suppression of the tone at the center frequency.

More generally, the procedure results in the enhancement of spectral peaks at the expense of adjacent frequencies. Any sufficiently intense frequencies outside the narrowband filter range but within the broadband filter range set a conservatively low gain in the compressor, but get filtered out by the narrowband filter and do not affect the expander. In this scenario, the compressor's gain is set by one set of frequencies while the expander's gain is set by another set of frequencies such that there is insufficiently large gain in the expander to completely undo the effect of the compression. The net effect is that there is overall suppression of weak narrowband tones in a channel by strong out-of-band tones. Note that these out-of-band tones in one channel will be the dominant tones in a neighboring channel where they are resonant. Consequently, the output spectrum of the filterbank will have a local winner-take-all characteristic with strong spectral peaks in the input suppressing or masking weaker neighboring ones and high signal-to-noise-ratio channels being emphasized over weaker ones. A more detailed analysis of the potential benefits and operation of the algorithm may be found in [17].

It is worth emphasizing that the *combination of nonlinearity and filtering* in the companding algorithm results in a center-surround-like kernel¹ [30] on the input spectral energies, which naturally enhances spectral contrast. A linear spatial bandpass filter on the input spectral energies does not yield the local winner-take-all behavior, although it does provide some contrast enhancement.

3. FFT-BASED COMPANDING

The companding strategy described above is well suited to low-power analog circuit implementations. On the other hand, the straightforward digital implementation of the architecture is computationally intensive. In this section, we extract a computationally efficient digital implementation of the companding architecture based on the FFT.

Figure 2 shows the details of a single channel of the analog time-domain architecture. We now derive a frequency domain architecture that is equivalent to Figure 2 over a short time frame of fixed duration T_N . Let X represent the FFT of the input signal x over an analysis frame (the upper case always refers to signals in the frequency domain, while lower case denotes signals in the time domain). In our representation X , is a column vector with as many components as the number of unique frequency bins in the FFT. Let F_i be the vector that represents the Fourier spectrum of the filter response of the broadband prefilter in the i th channel. The spectrum of the output signal x_1 of the prefilter is given by $X_{i,1} = F_i \otimes X$, where \otimes represents a Hadamard (componentwise) multiplication. Note that the i in $X_{i,1}$ denotes the i th spectral channel while the 1 denotes that it corresponds to x_1 in that channel.

We assume that the ED (envelope detector) block extracts the RMS value of its input such that $x_{i,1e} = |X_{i,1}|$, where the $|\cdot|$ operator represents the RMS value. We also assume that the output of the ED is constant over the course of the analysis frame (it does change from frame to frame). The output of the envelope detector (a scalar over the course of the frame) is raised to the power $n-1$ and multiplied by $X_{i,1}$. The spectrum of the output of the multiplier is therefore given by $X_{i,2} = |X_{i,1}|^{n-1} X_{i,1}$.

Let G_i represent the FFT of the impulse response of the narrowband postfilter in the i th channel. The spectrum of the output of the postfilter is given by

$$\begin{aligned} X_{i,3} &= G_i \otimes X_{i,2} = |X_{i,1}|^{n-1} G_i \otimes X_{i,1} \\ &= |F_i \otimes X|^{n-1} G_i \otimes F_i \otimes X. \end{aligned} \quad (1)$$

¹ Center-surround filtering refers to the application of a filter kernel whose weights have one sign (all positive or all negative) within a central region, and the opposite sign (all negative or all positive) outside the central region, termed the surround. This type of filtering is known to occur in the processing of visual information at several types of retinal cells that convey retinal information to the cortex.

We define a new filter H_i that is simply the combination of the F_i and G_i filters: $H_i = F_i \otimes G_i = G_i \otimes F_i$. We can now write

$$X_{i,3} = |F_i \otimes X|^{n-1} H_i \otimes X. \quad (2)$$

The second ED block computes the RMS value of $x_{i,3}$, that is,

$$x_{i,3e} = |F_i \otimes X|^{n-1} |H_i \otimes X|. \quad (3)$$

Once again, we assume that the output of the second ED block is constant over the course of the analysis frame. The output of the ED block is raised to the power $(1-n)/n$ and multiplied by $X_{i,3}$. The spectrum of the output of the second multiplier is hence given by

$$\begin{aligned} X_{i,4} &= |X_{i,3e}|^{(1-n)/n} X_{i,3} \\ &= (|F_i \otimes X|^{n-1} |H_i \otimes X|)^{(1-n)/n} |F_i \otimes X|^{n-1} H_i \otimes X \\ &= |F_i \otimes X|^{(n-1)/n} |H_i \otimes X|^{(1-n)/n} H_i \otimes X. \end{aligned} \quad (4)$$

The outputs of all the channels are finally summed. The spectrum of the final summed signal is simply the sum of the spectra from the individual channels. Hence, the spectrum of the companded signal y is given by

$$\begin{aligned} Y &= \sum_i X_{i,4} = \sum_i |F_i \otimes X|^{(n-1)/n} |H_i \otimes X|^{(1-n)/n} H_i \otimes X \\ &= \left(\sum_i |F_i \otimes X|^{(n-1)/n} |H_i \otimes X|^{(1-n)/n} H_i \right) \otimes X. \end{aligned} \quad (5)$$

The above equation is a fairly simple combination of Hadamard multiplications, exponentiation, and summation and can be performed very efficiently.

Note that by introducing a term $J(X)$ such that

$$J(X) = \sum_i |F_i \otimes X|^{(n-1)/n} |H_i \otimes X|^{(1-n)/n} H_i \quad (6)$$

we can write

$$Y = J(X) \otimes X. \quad (7)$$

It is clear from the above equation that the effect of the companding algorithm is to filter the signal x by a filter that is a function of x itself. It is this nonlinear operation that results in the desired enhancement of spectral contrast.

Mel-frequency spectral vectors are finally computed by multiplying Y_{power} , the power spectral vector corresponding

to Y by a matrix of Mel filters M in the usual manner:

$$Y_{\text{mel}} = M Y_{\text{power}}. \quad (8)$$

Note that the only additional computation with respect to conventional computation of Mel-frequency cepstra is that of (7). This is negligible in comparison to the computational requirements of a time-domain-filterbank-based implementation of the compounding algorithm as reported in [17].

The companding algorithm has several parameters that may be tuned to optimize recognition performance, namely, the number of channels in the filterbank, the spacing of the center frequencies of the channels, the design of the broadband prefilters (the F filters) and the narrowband postfilters (the G filters), and the companding factor n .

In the original companding algorithm presented in [17] and also the work in [23], the center frequencies of the F and G filters were spaced logarithmically, such that each of the F and G filterbanks had constant Q-factor. In the FFT-based implementation described in this paper, however, we have found it more effective and efficient to space the filters linearly. In this implementation, the filterbank has as many filters as the number of frequency bands in the FFT. The frequency response of the broadband prefilters (the F filters) and the narrowband postfilters (the G filters) have both been assumed to be triangular and symmetric in shape. The G filters are much narrower than the F filters. The width of the F filters represents the spectral neighborhood that affects the masking of any frequency. The width of the G filters determines the selectivity of the masking.

The optimal values of the width of the F and G filters and the degree of companding n were determined by experiments conducted on the CU-Move in-vehicle speech corpus [28] (the experimental setup is described in detail in Section 4). The lowest recognition error rates were obtained with F filters that spanned 9 frequency bands of a 512-point FFT of the signal (i.e., the frequency response fell linearly to zero over four frequency bands on either side of the center frequency and was zero elsewhere) and G filters that spanned exactly one frequency band. In the case of the G filters, the optimal support of the ‘‘triangle’’ was thus less than the frequency resolution of the FFT resulting in filters that had nonzero values in only one frequency bin. It is likely that using a higher resolution FFT might result in wider G filters with nonzero values in a larger number of frequency bins. The optimal value of n was determined to be 0.35.

Figure 3 shows the narrowband spectrogram plot for the sentence ‘‘three oh three four nine nine nine two three two’’ in car noise (CU-Move database), illustrating the effect of companding. The energy in any time-frequency component is represented by the darkness of the corresponding pixel in the figure: the darker the pixel, the greater the energy. The upper panel shows the spectrogram of the signal when no companding has been performed. The lower panel shows the spectrogram obtained when the companding algorithm is used to effect simultaneous masking on the signal. It is evident from the lower panel that the companding architecture is able to follow harmonic and formant transitions with

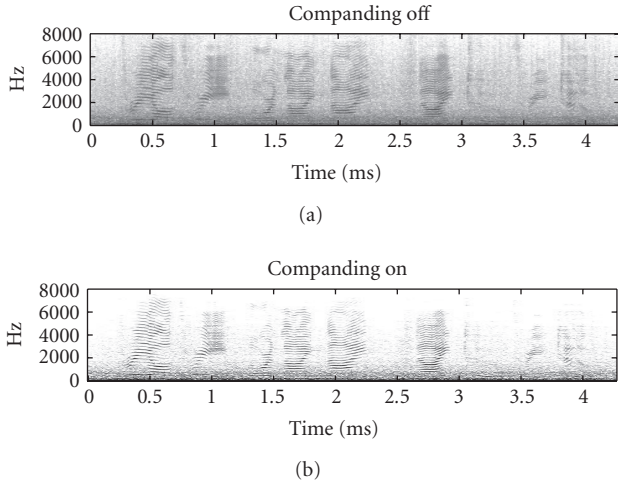


FIGURE 3: Spectrogram plots for the sentence “three oh three four nine nine nine two three two” in car noise (CU-Move database) illustrating the effect of companding. In the top figure, the companding strategy is disabled and in the lower figure the companding strategy is enabled.

clarity and suppress the surrounding clutter. In contrast, the top panel shows that, in the absence of companding, the formant transitions are less clear, especially at low frequencies where the noise is high.

4. EXPERIMENTS

Experiments were conducted on two different databases—the CU-Move in-vehicle speech corpus [28] and the Aurora-2 corpus [29]—to evaluate the effect of the proposed companding algorithm on speech recognition accuracy. The CU-Move data are sampled at 16 kHz, whereas the Aurora-2 data are sampled at 8 kHz. In order to retain consistency of spectral resolution (for companding) between the experiments on the CU-Move and Aurora-2 databases, the latter was up-sampled to 16 kHz. In all experiments, speech signals were parameterized using an analysis frame size of 25 milliseconds. Adjacent frames overlapped by 15 milliseconds. 13-dimensional Mel-frequency cepstral vectors (MFCs) were computed from the companded spectra for recognition. A total of 30 triangular and symmetric Mel filters were employed for the parameterization in all cases. For the CU-Move data, the 30 Mel filters covered the frequency range of 130–6500 Hz. For the Aurora-2 database, the 30 filters covered the frequency range of 130–3700 Hz. The slopes of the triangular Mel filters were set to $\beta \cdot \gamma$, where γ is the slope that would have been obtained had the lower vertex of each Mel triangle extended to lie exactly under the peak of the adjacent Mel triangle. It is known that setting the β values to less than 1.0 can result in improvement in recognition performance for noisy data [31]. β values of 1.0 and 0.5 were evaluated for the experiments reported in this paper. The overall procedure for the computation of cepstral features is shown in Figure 4. Figure 4 consists of two blocks—an upper companding block and a lower cepstrum-computation

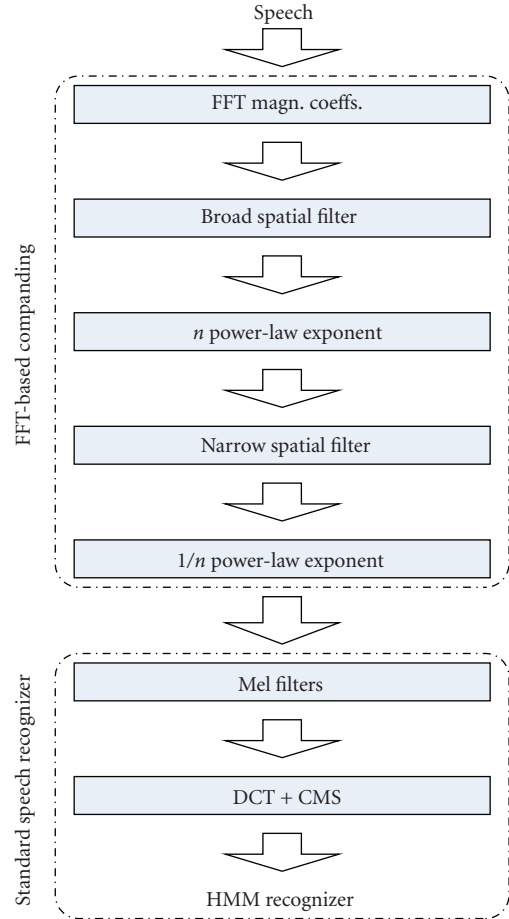


FIGURE 4: Block diagram of FFT-based companding. “DCT” refers to the discrete cosine transform, and “CMS” to cepstral mean subtraction.

block. For experiments evaluating our companding algorithm, both blocks were included in the feature computation scheme. For baseline experiments evaluating regular MFCs derived without companding, the upper companding block was bypassed, that is, the companding was turned off. Cepstral mean subtraction (CMS) was employed in all experiments. The mean-normalized MFCs were augmented with difference and double-difference vectors for all recognition experiments.

4.1. CU-Move database

We evaluated the companding front end on the digits component of the CU-Move database. CU-Move consists of speech recorded in a car driving around various locations of the continental United States, under varying traffic and noise conditions. Since the data are inherently noisy (i.e., the noise is not digitally added), the SNR of the various utterances is not known and must be estimated. We estimated the SNRs of the utterances by aligning the speech signals to their transcriptions using the Sphinx-3 speech recognition system, identifying nonspeech regions, and deriving SNR estimates

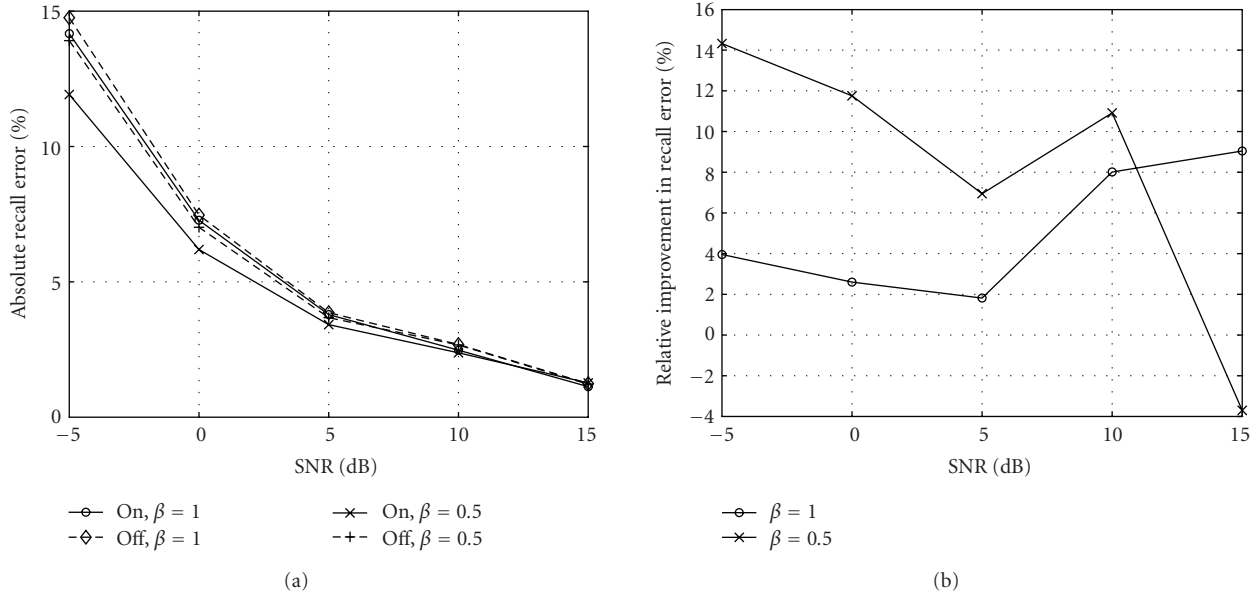


FIGURE 5: Percent recall error by test subset SNR for $\beta = 1$ (standard Mel filterbank) and $\beta = 0.5$ (broad Mel filterbank). In (a), the absolute values are shown and in (b) the relative recognition recall improvement with companding on compared to companding off is shown.

from the energy in these regions. We only used utterances for which we could conveniently get clean transcripts and SNR measurements: a total of 19 839 utterances. The data were partitioned approximately equally into a training set and a test set. A common practice in robust speech recognition research is to report recognition results on systems that have been trained on clean speech. While such results may be informative, they are unrepresentative of most common applications where the recognizer is actually trained on the kind of data that one expects to encounter during recognition. In our experiments on CU-Move, therefore, we have trained our recognizer on the entire training set, although the test data were segregated by SNR.

The Sphinx-3 speech recognition system was used for all experiments on CU-Move data. For the experiments, triphones were modeled by continuous density HMMs with 500 tied states, each in turn modeled by a mixture of 8 Gaussians. A simple “flat” unigram language model was used in all experiments. It was verified that under this setup the baseline performances obtained with regular Mel-frequency cepstra (with $\beta = 1$) by our system were comparable to or better than those obtained on the same test set with several commercial recognizers at all SNRs.

We conducted experiments with two different feature types: conventional MFC features (to establish a baseline), and features produced by the companding front-end. We used two different types of Mel filterbanks: “standard” filterbanks with $\beta = 1$, and broader filters with $\beta = 0.5$.

We report two different measures of performance. The recognition “recall” error is the percentage of all uttered words that were correctly recognized. Recall error is equal to $(D + S)/N * 100$, where N is the total number of labels in the reference transcripts, S is the number of substitution er-

rors, and D is the number of deletion errors. Figure 5 shows both the recall error obtained for the two values of β and the relative improvement in recall error as a percentage of the error obtained with companding turned off.

Recognizers also often insert spurious words that were not spoken. The “total” error of the recognizer is the sum of recall and insertion errors, expressed (as before) as a percentage of all uttered words, and is given by $(D + S + I)/N * 100$, where I is the number of insertion errors. Figure 6 shows the total error obtained for the two values of β as well as the relative improvement in error relative to the performance obtained with companding turned off. We note that spectral-contrast enhancement can result in the enhancement of spurious spectral peaks as well as those from the speech signal. This can result in increased insertion errors. We therefore present the recall and total errors separately so that both effects—the increased recognition of words that were spoken, and any increased insertion errors—are appropriately represented.

The results of our evaluations are shown in Figures 5 and 6. For the plots, the test utterances were grouped by SNR into 5 subsets, with SNRs in the ranges < -2.5 dB, -2.5 dB to 2.5 dB, 2.5 dB to 7.5 dB, 7.5 dB to 12.5 dB, and > 12.5 dB, respectively. The x -axes of the figures show the centre of the SNR range of each bin.

We observe that the recognition performance, measured both in terms of recall error and total error, improves in almost all cases, particularly at low SNRs. Further, while broadening the Mel filters ($\beta = 0.5$) does not produce great improvement in recognition performance when no companding is performed, it is observed to result in significant improvement over recognition with standard Mel filters ($\beta = 1$) when companding is turned on.

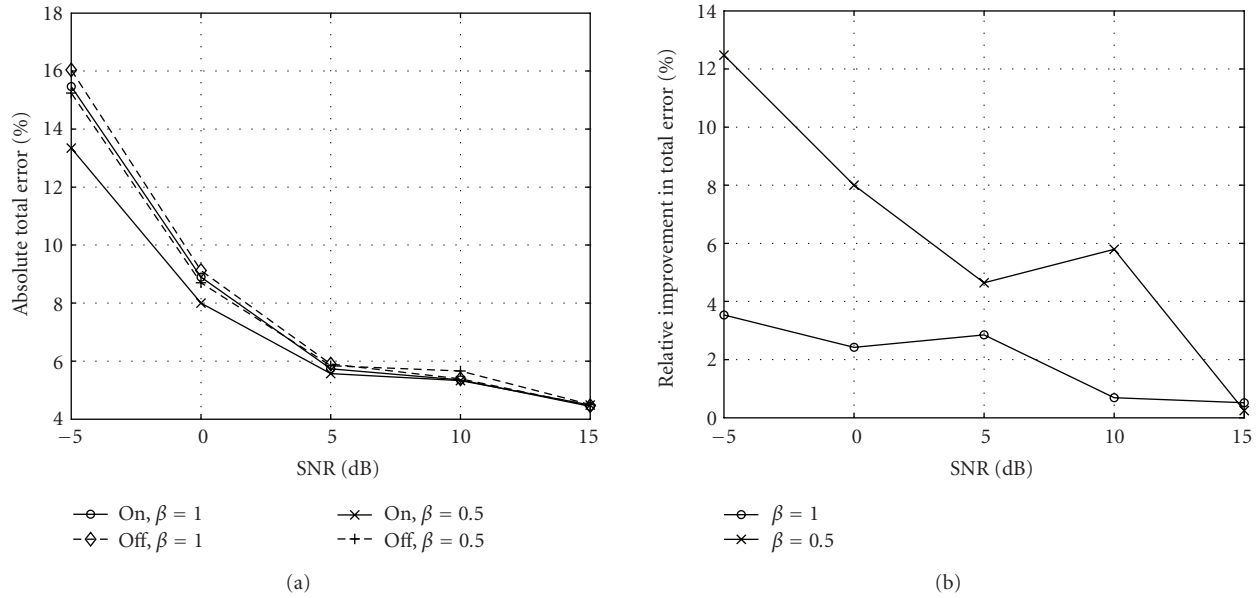


FIGURE 6: Percent total error rate by test subset SNR for $\beta = 1$ (standard Mel filterbank) and $\beta = 0.5$ (broad Mel filterbank). (a), the absolute values are shown and in (b) the relative error rate improvement with companding on to companding off is shown. This figure shows the total error rate including false insertions, substitutions, and deletion, while Figure 5 shows the error rate with substitution and deletion only.

Improvements are observed to increase with decreasing SNR. At -5 dB, a relative improvement of 4.0% in recall error and of 3.5% in total error is obtained with standard Mel filters ($\beta = 1$). With the broader Mel filters ($\beta = 0.5$), a relative improvement of 14.3% in recall error and of 12.5% in total error is obtained. Overall, on average, with standard Mel filters, the relative improvements in recall and total errors are 5.1% and 2.0%, respectively, while with broader Mel filters, the relative improvements in recall and total errors are 8.1% and 6.2%, respectively.

4.2. Aurora-2 database

The effect of two-tone suppression by the companding algorithm was also tested on the Aurora-2 database. Aurora-2 [29] consists of 8 kHz sampled speech derived from the TIDigits database. The training and test utterances are continuous sequences of digits. The database consists of 16 880 recordings designated as training data, which includes both clean recordings and recordings of speech corrupted to a variety of SNRs by digital addition of a variety of noises. The test data include a total of 84 084 recordings partitioned into three sets, each including both clean speech and speech corrupted to several SNRs by a variety of noises.

As mentioned earlier, we up-sampled the database to 16 kHz; however, only frequencies between 130 Hz and 3700 Hz were used to compute MFCs. We employed the HTK recognizer [32] in order to conform to the prescribed experimental setup for the database. Whole-word models were trained for each of the digits. For experiments with Aurora-2, wider Mel-frequency filters ($\beta = 0.5$) were used in all experiments, since these were observed to result in better recog-

nition on the CU-Move database. We conducted two different sets of experiments. In the first, a “clean” recognizer was trained with only the 8440 clean utterances of the Aurora-2 training corpus. For the second set a “multicondition” recognizer was trained using all the available training data, including both clean and noisy recordings.

Figure 7 shows the recall error and the total error for both clean and multicondition recognizers, that has been obtained with companding turned off, as a function of SNR for several noise types. Figure 8 shows the relative improvements obtained due to two-tone suppression by companding for each of these noise types, also as a function of SNR. Figure 9 summarizes these relative improvements and shows the average improvement in each of these metrics.

It is clear from these figures (and particularly from Figure 9) that the companding algorithm is able to improve recognition performance significantly under almost all noise conditions, when the recognizer has been trained on clean speech. On speech corrupted by subway noise, for example, companding results in a relative improvement of 13.5% in recall error and 16.3% in total error. Even for the multicondition recognizer, companding is observed to result in significant improvements in recognition performance for most noise types. For example, for speech corrupted by subway noise, companding reduces the recall error by 10.3% and the total error by 6.9%. The error is not always observed to decrease for the multicondition recognizer, however. On speech corrupted by babble, airport, and train station noises, companding is observed to result in an increase in recognition error. However, even for these conditions, the total error is observed to improve when the recognizer has been trained on clean speech.

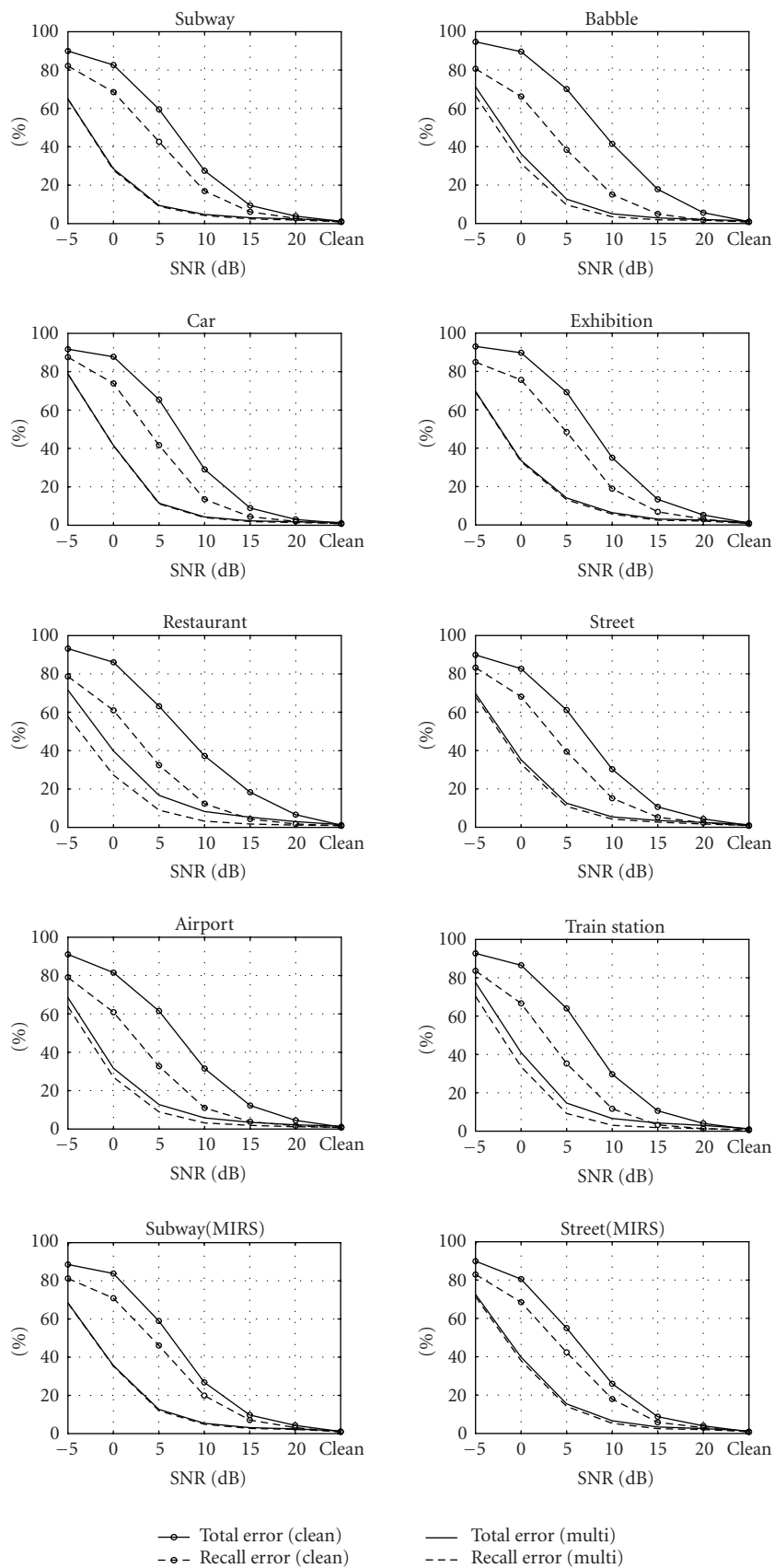


FIGURE 7: Absolute recognition error and recall error by test noise subset with companding turned off. In every noise subset the points correspond to $-5, 0, 5, 10, 15, 20$, and clean, dB SNR from from left to right.

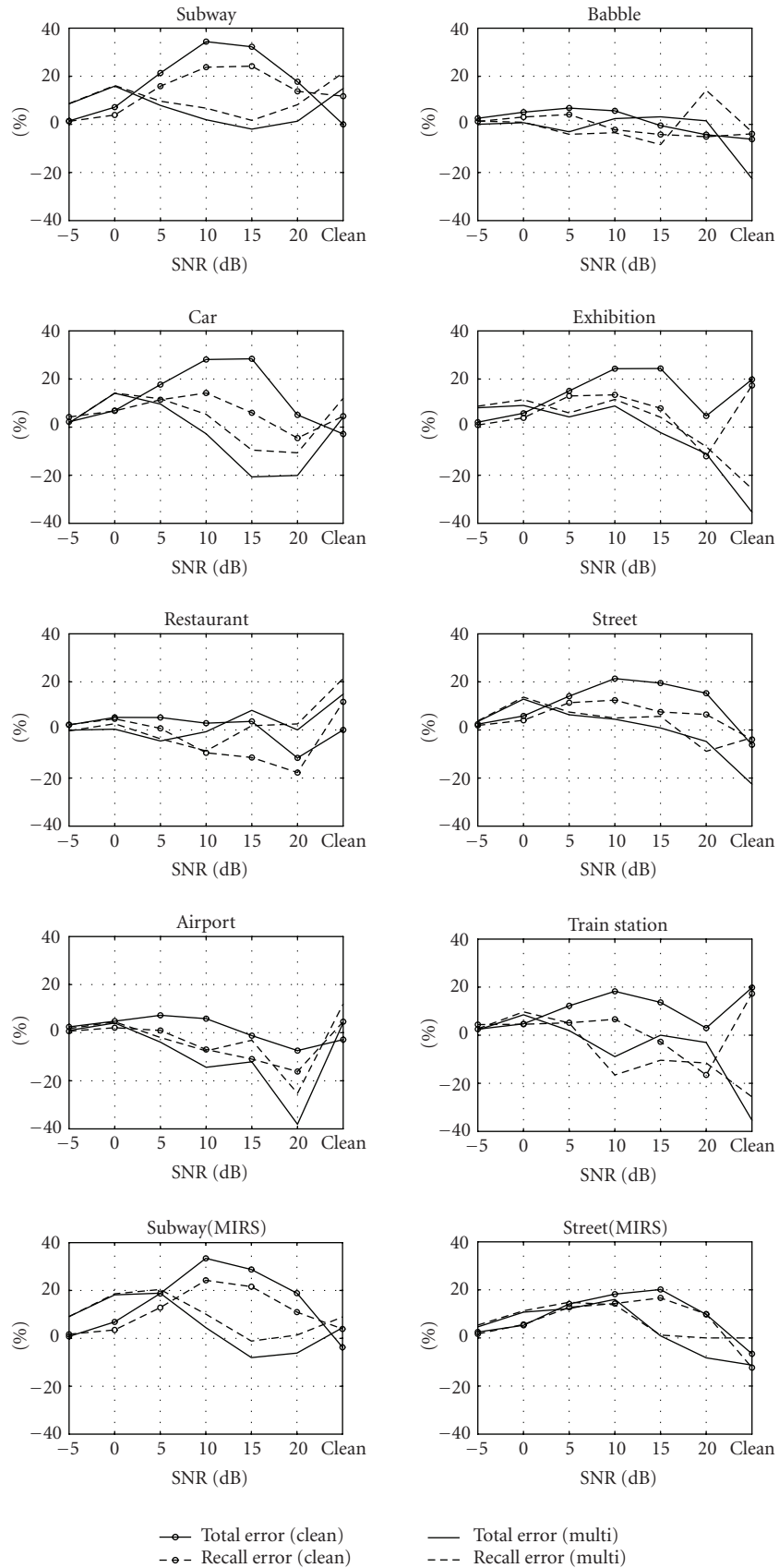


FIGURE 8: Relative improvement in recognition error and recall error by test noise subset with companding on versus companding off. In every noise subset, the points correspond to $-5, 0, 5, 10, 15, 20$, and clean, dB SNR from (a) to (j).

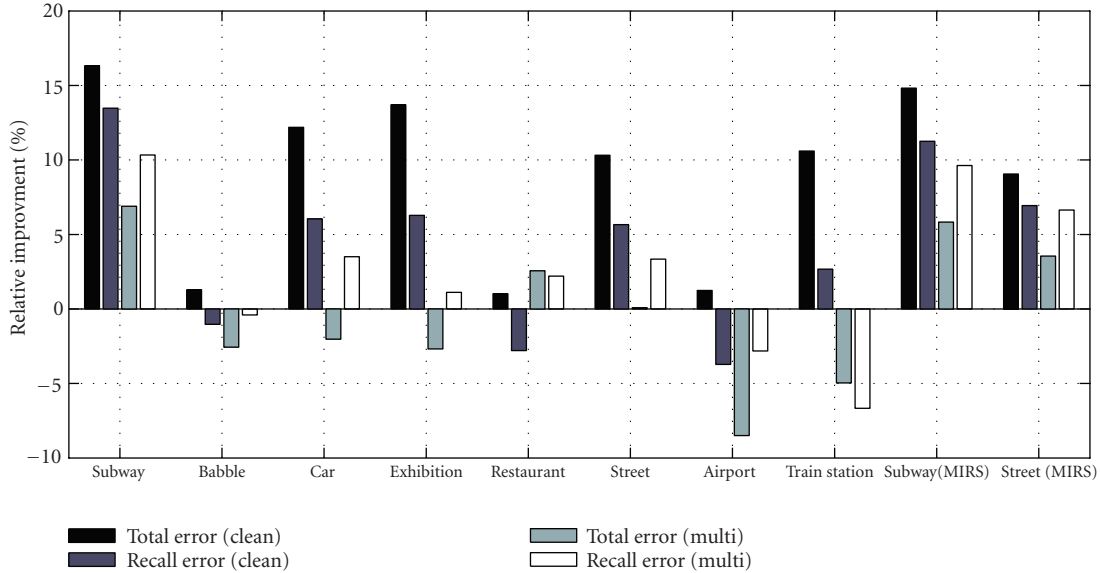


FIGURE 9: Relative recognition error and recall error for clean and multicondition training by test noise subset (averaged over different SNRs). “MIRS” refers to data that have been filtered in order to impose the frequency characteristics of a standard telecommunication channel specified by the ITU [29].

5. CONCLUSIONS

In this paper, we have presented a biologically-motivated signal-processing algorithm that affects simultaneous masking of speech spectra via the mechanism of two-tone suppression. Cepstral features derived from spectra enhanced in this manner are observed to result in significantly superior automatic speech recognition performance, compared to conventionally computed Mel-frequency cepstra. In an automotive digits recognition task, the in-car CU-Move database, the algorithm improves the relative word error by 12.5% at -5 dB signal-to-noise ratio (SNR) and by 6.2% across all SNRs (-5 dB SNR to $+15$ dB SNR). In the Aurora-2 database, corrupted by digitally added noise from several environments, the algorithm improves the relative word error rate in most situations when the models are trained on clean speech. The improvements observed are often substantial. Interestingly, for multicondition models (both Aurora-2 and CU-Move), the improvements in recall error were observed to be greater than those in the total error. In particular, on Aurora-2 the total error actually increased in 5 of the 10 conditions as a result of companding. This is in contrast to the CU-Move corpus, where improvements in error were consistently observed.

In the quest for the perfect biologically inspired signal processing scheme for noise-robust speech recognition, it is important to be able to distinguish psychoacoustic phenomena that are relevant to the problem from those that are simply incidental. The algorithm presented here aims to mimic simultaneous masking. Our experiments reveal that incorporating simultaneous masking does improve robustness to noise. It is therefore valid to hypothesize that simul-

taneous masking is a significant component of human noise robustness.

The model presented in this paper is derived from a companding algorithm proposed earlier [17]. As shown in [17], our companding model does in fact model the masking phenomenon of the peripheral auditory system well. Other studies have revealed that the enhancement of speech sounds through the proposed algorithm can improve perception in cochlear implant patients in noise [19, 21]. However, the algorithm presented in this paper is not a direct transliteration of the original algorithm; rather, it is an FFT-based adaptation intended to be more efficient and amenable to incorporation in an automatic speech recognition system than the original algorithm. The most effective FFT-based implementation varies significantly from the original analog design. For instance, the model in [17] incorporates time constants through which past sounds affect the spectrum of current sounds. The FFT-based model, however, is instantaneous within an analysis frame. The F and G filters are simply triangular; however, more biologically-inspired filters would require asymmetric filter shapes that are closer to the typical masking curves measured in humans. In particular, we have found the optimal G filter to be only one FFT bin wide. Since the actual masking is obtained over the frequencies covered by F , but not by G , this is equivalent to restricting each channel to mask out all frequencies passed by the G filters of other channels—in other words, leakage from adjacent channels does not affect the overall masking characteristics (i.e., across all channels). Narrower G filters also result in greater spectral contrast, as shown in [17]. However, the question of what the optimal shape of the G filter would be for FFTs with finer frequency resolution than the number of channels employed for companding remains open. All of these issues represent

avenues that may be explored to derive more optimal computational models for simultaneous masking that might further improve automatic speech recognition performance in noise. These avenues remain to be explored.

ACKNOWLEDGMENTS

The authors thank Keng Hoong Wee and Dr. Rita Singh for useful discussions. The authors thank Dr. Stephen Voran and three anonymous reviewers for their helpful comments on an earlier version of this manuscript.

REFERENCES

- [1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [2] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, London, UK, 1988.
- [3] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, no. 1, pp. 55–76, 1988.
- [4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, part 2, pp. 115–132, 1994.
- [5] A. Van Schaik and R. Meddis, "Analog very large-scale integrated (VLSI) implementation of a model of amplitude-modulation sensitivity in the auditory brainstem," *Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 811–821, 1999.
- [6] J. L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing Research*, vol. 49, no. 1–3, pp. 39–60, 1990.
- [7] R. Meddis, L. P. O'Mard, and E. A. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2852–2861, 2001.
- [8] C. R. Jankowski Jr., H.-D. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 286–293, 1995.
- [9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [11] B. Strobe and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 451–464, 1997.
- [12] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 43–49, 2006.
- [13] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2040–2050, 1999.
- [14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.
- [15] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, New York, NY, USA, 4th edition, 1997.
- [16] L. Turicchia and R. Sarpeshkar, "The silicon cochlea: from biology to bionics," in *Biophysics of the Cochlea: From Molecules to Models*, A. W. Gummer, Ed., pp. 417–423, World Scientific, Singapore, 2003.
- [17] L. Turicchia and R. Sarpeshkar, "A bio-inspired companding strategy for spectral enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 243–253, 2005.
- [18] A. J. Oxenham, A. M. Simonson, L. Turicchia, and R. Sarpeshkar, "Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing," *Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1709–1716, 2007.
- [19] A. Bhattacharya and F.-G. Zeng, "Companding to improve cochlear implants' speech processing in noise," in *Proceedings of Conference on Implantable Auditory Prostheses*, Pacific Grove, Calif, USA, July-August 2005.
- [20] Y. W. Lee, S. Y. Kwon, Y. S. Ji, et al., "Speech enhancement in noise environment using companding strategy," in *Proceedings of the 5th Asia Pacific Symposium on Cochlear Implant and Related Sciences (APSCI '05)*, Hong Kong, November 2005.
- [21] P. C. Loizou, K. Kasturi, L. Turicchia, R. Sarpeshkar, M. Dorman, and T. Spahr, "Evaluation of the companding and other strategies for noise reduction in cochlear implants," in *Proceedings of Conference on Implantable Auditory Prostheses*, Pacific Grove, Calif, USA, July-August 2005.
- [22] L. Turicchia, K. Kasturi, P. C. Loizou, and R. Sarpeshkar, "Evaluation of the companding algorithm for noise reduction in cochlear implants," submitted for publication.
- [23] J. Guinness, B. Raj, B. Schmidt-Nielsen, L. Turicchia, and R. Sarpeshkar, "A companding front end for noise-robust automatic speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 249–252, Philadelphia, Pa, USA, March 2005.
- [24] M. A. Stone and B. C. J. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: effects on speech intelligibility and quality," *Journal of Rehabilitation Research and Development*, vol. 29, no. 2, pp. 39–56, 1992.
- [25] T. Baer, B. C. J. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," *Journal of Rehabilitation Research and Development*, vol. 30, no. 1, pp. 49–72, 1993.
- [26] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 947–954, 1987.
- [27] M. J. Hunt, "Some experience in in-car speech recognition," in *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp. 25–31, Tampere, Finland, May 1999.
- [28] University Technology Corporation, "CSLR Speech Corpora," <http://cslr.colorado.edu/beginweb/speechcorpora/corpus.html>.
- [29] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of*

Automatic Speech Recognition: Challenges for the New Millennium (ISCA ITRW ASR '00), pp. 181–188, Paris, France, September 2000.

- [30] E. R. Kandel, J. H. Schwarz, and T. M. Jessell, *Principles of Neural Science*, McGraw Hill, New York, NY, USA, 2000.
- [31] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, “Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, pp. 273–276, Salt Lake, Utah, USA, May 2001.
- [32] The Hidden Markov Model Toolkit (HTK), University of Cambridge, <http://htk.eng.cam.ac.uk/>.