A COMPANDING FRONT END FOR NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION

Jethran Guinness[¥], Bhiksha Raj[¥], Bent Schmidt-Nielsen[¥], Lorenzo Turicchia[$], Rahul Sarpeshkar[$]

[¥]Mitsubishi Electric Research Laboratories, Cambridge, MA
[$]Massachusetts Institute of Technology, Cambridge, MA

ABSTRACT

Feature computation modules for automatic speech recognition (ASR) systems have long been modeled on the human auditory system. Most current ASR systems model the critical band response and equal loudness characteristics of the auditory system. It has been postulated that more detailed models of the human auditory system can lead to more noise-robust speech recognition. An auditory phenomenon that is of particular relevance to robustness is simultaneous masking, whereby dominant frequencies suppress adjacent weaker frequencies. In this paper we present a companding-based model that mimics simultaneous masking in the front end of a speech recognizer. In an automotive digits recognition task, the front end improves word error rate by 4.0% (25% relative to Mel cepstra) at –5 dB SNR at the cost of a 1.7% increase at 15 dB SNR.

1. INTRODUCTION

Human beings are able to recognize speech amazingly well in high levels of background noise. On the other hand, the performance of automatic speech recognition (ASR) systems degrades dramatically with increasing noise. Part of the reason for this difference lies in the fact that the auditory system incorporates several features that make it more robust to noise. Most contemporary ASR systems attempt to incorporate some of these features. For instance, the most common feature representation of speech signals in ASR systems, the MEL spectrum, incorporates a simplified model of Critical band analysis, which resolves the speech signal into overlapping frequency bands of increasing width, with center frequencies spaced like the human auditory system. Another popular feature representation, the perceptual linear prediction or PLP spectrum, incorporates models of the equal loudness characteristic as well as the intensity-loudness relationship present in the auditory system [1].

Several other characteristics contribute to the noise robustness of the auditory system, e.g. the Zwicker effect [2], whereby the auditory system adapts to loud signals and filters them out, and masking. Masking, as defined by the American Standards Association (ASA), is the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound [3]. Temporal masking is the phenomenon whereby loud sounds mask other sounds for a short time before and after their occurrence. Of particular relevance to this paper is simultaneous masking, whereby high-energy frequency components mask adjacent frequencies with lower energy.

These phenomena allow the auditory system to concentrate on high-energy speech components in the received signal and suppress persistent and transient noise phenomena that might obscure comprehension of the speech. It can be expected that signal processing schemes for speech recognition systems that mimic the above phenomena would similarly improve the ability of ASR systems to recognize noise-corrupted speech. Over the years several signal-processing models that attempt to duplicate the processing of the peripheral auditory system (e.g. [3, 4, 5]) have been proposed for ASR. The majority of these proposals attempt to model all the individual steps in the processing of the speech signal *in extenso*, modeling such steps as the response of the basilar membrane and rectification implicit in the response of sensory hair cells in rigorous detail. While such schemes have been observed to improve recognition somewhat, they have fallen short of the performance that their computational detail and complexity might promise.

In this paper we develop a simple signal-processing mechanism that attempts to model the effects, but not the procedures, of some of the mechanisms in the human auditory system. In particular, the signal processing scheme, based on a model presented by Turicchia and Sarpeshkar [6], effectively implements frequency masking by the dynamic adjustments of gains through a companding scheme applied to a two-level filter bank. Although the model is not anthropomorphic in its detail, it exhibits the frequency masking effects, and the corresponding enhancement of spectral peaks noted in the auditory system. The original model proposed by Turicchia and Sarpeshkar was primarily intended for cochlear implants [7]. The model demonstrated that companding-based signal processing is effective at improving the spectral quality of tones in noise, a vowel in noise, and a word in noise [6]. In this paper we extend and modify the model to compute features that are most effective for computer speech recognition. Experiments on the CU-Move database (an extensive database of speech recorded in moving cars) [8] reveal that the proposed signal processing scheme is able to enhance spectral peaks as expected and to improve ASR performance significantly at very low SNRs.

Signal processing schemes often improve recognition performance in "mismatched" conditions, i.e. when the recognizer has been trained on clean speech, but the data to be recognized are noisy, but fail to improve performance when the training data are similar to the test data (a more realistic situation for most applications). One of the features of our study is that it has been conducted with real-world recordings, under matched

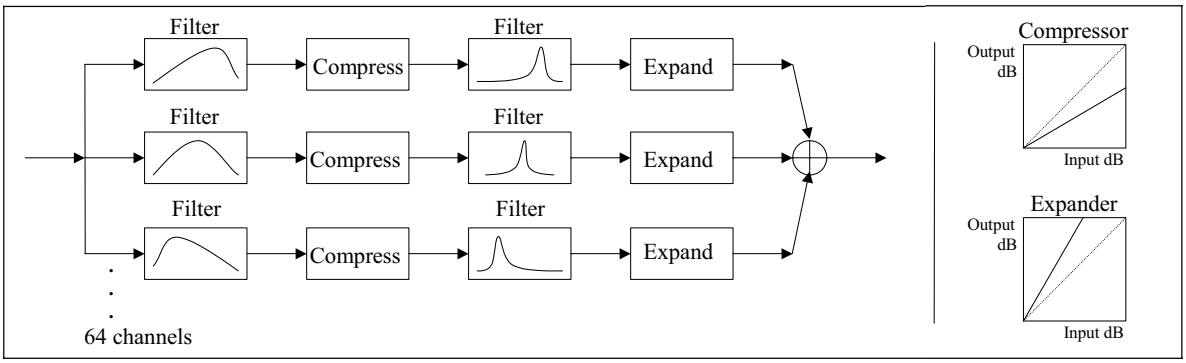

Figure 1: Compander design

conditions. Thus the observed improvements can be expected to carry to real-world scenarios.

2. THE COMPANDING MODEL FOR MASKING

Frequency masking is a phenomenon whereby high-energy frequency bands suppress adjacent low-energy frequencies. This results in a deepening of spectral valleys adjacent to spectral peaks. In order to effect such a phenomenon, the gain applied to any frequency must depend on its relation to its neighboring frequencies. Figure 1 shows the basic signal processing model used to perform frequency masking. The model follows the scheme proposed by Turicchia and Sarpeshkar [6]. There are four stages in each parallel channel of processing: A wideband filter, a compression stage, a narrow-band filter, and an expansion stage. The expanded outputs from all stages are summed to yield the final output.

The first stage consists of a constant-Q filter that we refer to as the "F" filter. The filter was designed by digitizing the analog filter described by

$$F = \left(\frac{\frac{\tau}{q_F} S}{\frac{\tau^2}{q_F{}^2} S + 2\frac{\tau}{q_F} S + 1} \right)^4$$

where q_F is a parameter that controls the Q factor of the filters and τ is the inverse of the resonant frequency (CF) in radians.

Following each F filter is a compressor that compresses the signal by a factor that is proportional to the instantaneous power at the output of the filter. The compressor consists of an envelope detector and a multiplier. The envelope detector used is the digital equivalent of a rectifier followed by a second-order lowpass filter with time constant $\alpha \cdot \tau_F$, where α is a parameter and τ_F is 1/CF of the corresponding F filter. Each compressor compresses the output of its F filter according to

$$output = input \times inputEnvelope^{(n-1)}$$

where n is a parameter that controls the strength of the masking effect of companding.

Each compressor is followed by a second filter, which we call the "G" filter. The G filter has the same design as the F filter, but is controlled by an independent quality-factor parameter q_G. The design requires q_G to be greater than q_F, making the G filters narrower than the F filters.

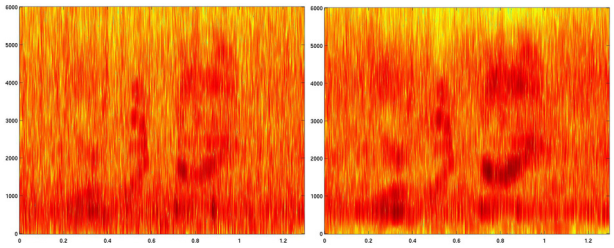
Figure 2: Spectrogram of sample data before companding (left) and after companding (right)

Each G filter is followed by an expander that expands the signal by a factor that is proportional to the instantaneous power at the output of the filter. The expander has the same design as the compressor, but is controlled by an independent parameter β in place of α, and raises its envelope to the power (1-n)/n. The outputs of the expanders are summed back together to obtain a spectrally enhanced signal.

The key to understanding the companding model is to realize that each channel in Figure 1 is intended to primarily process frequency components that lie in a narrow frequency range around its center frequency. All processing aspects are intended to impose appropriate frequency masking features on those frequencies. The initial F filters are wide band, and permit frequencies that lie in the neighborhood of the center frequency to alter the gain of the compressor. The G filter lets through a significantly narrower band of frequencies around the center frequency. Only these frequencies can alter the gain of the expander. Thus, if the center frequency lies in a valley, the compression factor, being related to the energy in adjacent high-energy bands, is greater than the expansion factor that depends mainly on the center frequency. On the other hand, for peaks, the compression and expansion are both related chiefly to the energy in the center frequency itself and effectively preserve the signal by undoing each other's compression and expansion effects. This results in a deepening of the spectral valleys with respect to peaks. A more detailed explanation of the effects of companding can be found in [6].

Figure 2 shows the spectrogram of a speech signal before and after it has been enhanced by the compander. The frequency masking effect and the enhancement of spectral peaks is evident in the figure. The signal is a section of an utterance from the CU-Move database. Recognition without companding results in the deletion of one of the words in the segment, whereas it is recognized in the companded signal.

3. SIGNAL PROCESSING FOR SPEECH RECOGNITION

The signal energies at the output of the expander in each of the channels in Figure 1 could potentially be used to directly obtain an estimate of the power spectrum of the signal. However, empirically, this has been found to be suboptimal for speech recognition. Instead, we compute features from the enhanced signal obtained by adding the channels back together. The enhanced signal is then passed through another filterbank (the "H" bank), the energy at the output of which is sampled to obtain a short-time Fourier transform representation of the signal. Cepstral feature vectors are obtained by taking the DCT of the log of the power spectra. This scheme is shown in Figure 3, along with the Sphinx-3 Mel-frequency cepstra (MFC) front end that we use as a baseline.

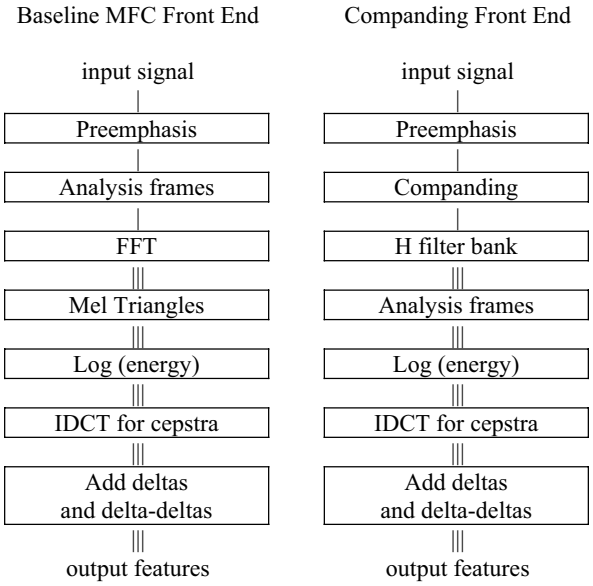

Figure 3: MFC front end versus companding front end

The important difference between the two systems is that the FFT and Mel triangle stages are replaced by the companding and H filter bank stages. As a consequence, the division of the signal into analysis frames is deferred until it is needed.

We note that the actual implementation of the signal processing scheme described in Section 2 and above can be implemented in a conventional ASR feature computation module by appropriate manipulation of the power spectra. However, our system is designed to enable direct implementation in analog hardware. Consequently we explicitly construct the filterbanks shown in Figures 1 and 3.

4. PARAMETER SETTINGS

The final values for the various parameters of the companding front end were determined experimentally, using the CU-Move data set and experimental setup described in Section 5. The final values are as follows:

1. Analysis frame size: 25 ms wide with a shift of 10 ms.
2. Number of channels: 64, with center frequencies ranging from 130 to 6500 Hz, spaced equally along the Mel frequency scale (i.e. approximately log-linearly on the linear frequency axis.)
3. F quality factor parameter q_F: 2
4. Compressor envelope detector time constant α: 5
5. G quality factor parameter q_G: 4
6. Expander envelope detector time constant β: 20
7. Companding factor 'n': 0.15, but it was found to be better to apply companding only at lower frequencies. Consequently the parameter n is rolled off via a sinusoid function from the value 0.15 at 2.45 kHz and below, to 1.00 (no companding) at 3.45 kHz and above.
8. H filter equal to cascade of F and G with q_F: 4 and q_G: 8
9. Number of cepstral coefficients retained: 13

The parameters of the baseline MFC front end were made analogous to the parameters of the companding front end. E.g., we used 64 Mel triangles across the frequency range 130 Hz - 6500 Hz. We did not however try to make the slopes of the triangles similar to the slopes of the H filters.

5. EXPERIMENTAL EVALUATION

We evaluated the companding front end on the digits component of the CU-Move in-vehicle speech corpus [8]. CU-Move consists of speech recorded in a car driving around various locations of the continental United States, under varying traffic and noise conditions. We estimated the SNRs of utterances by aligning the utterances to their transcriptions with Sphinx-3, identifying silence regions, and deriving SNR estimates from the energy in the silence regions. We used only 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 all of our experiments, therefore, we have trained our recognizer on the entire training set, although the test data were segregated by SNR. The baseline performances obtained with Mel frequency cepstra by our system were therefore found to be comparable to or better than that obtained on the same test set with several commercial recognizers.

The Sphinx-3 speech recognition system was used in all experiments. In all experiments continuous density HMMs with 500 tied states, each modeled by a mixture of 8 Gaussians, were used. A simple "flat" unigram language model was used in all experiments.

The results of our evaluations are shown in Figures 4 and 5. For the plots, the test utterances were grouped by SNR into 5 subsets, with SNRs in the ranges <-2.5db, -2.5db to 2.5db, 2.5db to 7.5db, 7.5db to 12.5db, and >12.5dB respectively. The X axis of the figures shows the centre of the SNR range of each bin.

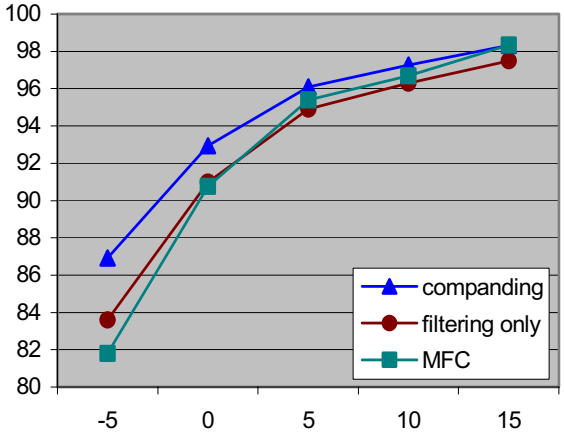

Figure 4: % word accuracy (recall) by test subset SNR

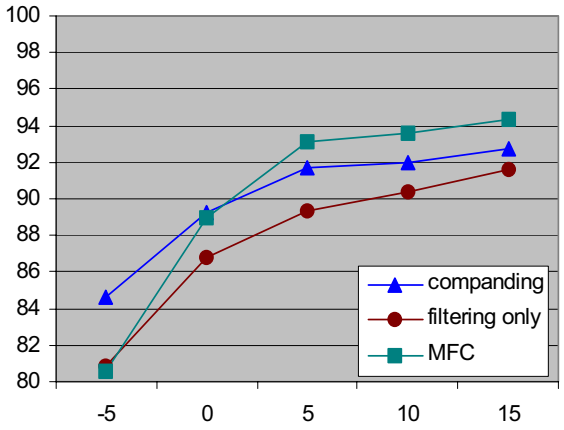

Figure 5: % (1 - word error) by test subset SNR

Experiments were conducted with three different feature types: conventional Mel Frequency Cepstra (MFC), to establish a baseline, features computed using the companding front-end described in Sections 2 and 3 (companding), and finally, a non-companding filterbank based feature, which is implemented by eliminating the companding component of the companding front-end in Figure 3 (filtering only).

We report two different measures of performance. Figure 4 shows the recognition *recall*. This shows the percentage of all words uttered by the speakers that were correctly recognized by the recognizer. Recognizers also often insert spurious words that were not spoken. Figure 5 shows the *accuracy* obtained after adjusting for such insertions.

6. OBSERVATIONS AND DISCUSSION

We observe that the recall obtained with the companding front end is consistently better than that obtained with MFCs. The insertion adjusted accuracy of the companding front end remains below that of MFCs; however at very low SNRs, even this number is significantly better than the baseline.

The companding front-end enhances peaks and the spectral structure of speech sounds, resulting in improved recall performance for the recognizer. However, masking also has the effect of enhancing spurious spectral peaks from the background noise, causing word insertions from these spurious peaks. Our results indicate that while at relatively high SNRs the insertions cancel out the gains from the improved recall, at low SNRs the improvement in recall is significantly greater than the increase in the insertion rate, resulting in improved overall recognition as compared to Mel frequency cepstra. Other experiments not reported here show that these improvements carry over to SNRs as low as -20 dB. The proposed front end is hence useful both in situations where recall is important, and where very low SNRs may be expected. Additionally, it may be expected that adaptive companding, that adjusts the companding level by the SNR of the signal, may result in even better recognition performance.

A second noteworthy point is that both the recall and accuracy obtained with the companding front end is superior to that obtained with the "filtering only" feature. It is important to bring out the relevance of this comparison: Mel cepstral analysis essentially simulates a filter bank. The "filtering-only" feature represents features obtained from an explicitly implemented filterbank. It can hence be expected that with appropriate settings of the filters, the performance obtained with the filtering only front end can be brought to the level of MFCs, and that the performance of the Companding front end can then improve correspondingly.

Finally, we note that the design of the front-end is such that it can be implemented in very low power analog VLSI. Such a front end will enable continuous, robust, low power listening in devices from cellphones to security systems.

7. REFERENCES

1. H. Hermansky. "Perceptual linear predictive analysis of speech". J. Acoust. Soc. Am. 87. pp 1738-1752, 1990.

2. E. Zwicker. "'Negative Afterimage' in Hearing". J. Acoust. Soc. Am. 36, pp. 2413–2415, 1964.

3. Moore, B.C.J. *An Introduction to the Psychology of Hearing. Fourth edition*, Academic Press, New York, 1997.

4. Seneff S. *Pitch and spectral analysis of speech based on an auditory synchrony model* (RLE Technical Report No. 504). MIT Press, Cambridge, 1985.

5. Lyon, R.F.; Mead, C.; "An analog electronic cochlea", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36 , No. 7 , pp. 1119-1134, 1988.

6. L. Turicchia and R Sarpeshkar, "A Bio-Inspired Companding Strategy for Spectral Enhancement". To appear in *IEEE Transactions on Speech and Audio Processing.*

7. L. Turicchia and R Sarpeshkar, "The Silicon Cochlea: From Biology to Bionics", *"The Biophysics of the Cochlea: Molecules to Models"* World Scientific, New Jersey, 2003.

8. University Technology Corporation. "CSLR Speech Corpora." [http://cslr.colorado.edu/beginweb/speechcorpora/corpus.html] Retrieved Aug. 17, 2004.