IMPROVING THE NOISE AND SPECTRAL ROBUSTNESS OF AN ISOLATED-WORD RECOGNIZER USING AN AUDITORY-MODEL FRONT END

Martin Hunke[1], Meeran Hyun[1], Steve Love[2] and Thomas Holton[1]

[1]School of Engineering, San Francisco State University, San Francisco, CA, USA
[2]Meridian Speech Technology, Danville, CA, USA

ABSTRACT

In this study, the performance of an auditory-model feature-extraction "front end" was assessed in an isolated-word speech recognition task using a common hidden Markov model (HMM) "back end", and compared with the performance of other feature representation front-end methods including mel-frequency cepstral coefficients (MFCC) and two variants (J- and L-) of the relative spectral amplitude (RASTA) technique. The recognition task was performed in the presence of varying levels and types of additive noise and spectral distortion using standard HMM whole-word models with the Bellcore Digit database as a corpus. While all front ends achieved comparable recognition performance in clean speech, the performance of the auditory-model front end was generally significantly higher than other methods in recognition tasks involving background noise or spectral distortion. Training HMMs with speech processed by the auditory-model or L-RASTA front end in one type of noise also improved the recognition performance with other kinds of noise. This "cross-training" effect did not occur with the MFCC or J-RASTA front end.

1. INTRODUCTION

Most automatic speech recognition systems are based on a spectral-energy approach to feature extraction, such as computation of FFTs, LPCs, or cepstral coefficients. These systems are much more sensitive to additive noise and conditions of spectral distortion than human speech perception [1]; hence, it has been suggested that recognition performance could be improved in these adverse conditions by using feature extraction approaches based on the human auditory system [2].

Several previous studies have applied HMM-based isolated word recognizers to digit databases. Discrete HMMs using a 64-element VQ codebook derived from LPC input coefficients achieved a 92.8% recognition rate [9]. The performance increased to 94.8% in a system employing continuous density HMMs with a diagonal covariance model and five Gaussian mixture components [10]. A similar experiment, conducted on the Bellcore Digit database using J-RASTA and PLP front-end techniques reported on the recognition performance in a variety of adverse conditions, including additive noise and linear spectral distortion [8]. The additive noise (Volvo noise) was recorded over a cellular phone from a moving automobile. Linear filtering of the recorded sounds was used to simulate the changes of frequency response due to switching from an electret microphone to a carbon microphone.

In this paper, the performance of an auditory-model front end [5,6] is compared to several other front ends including MFCC and two variants of RASTA (J- and L-) in a variety of adverse conditions. To make the comparison of the performance of different front ends meaningful, we chose a uniform speaker-independent isolated word recognition task (the Bellcore digits task), and used identical testing and training methodology and an identical pattern-matching system based on standard whole-word HMM models.

2. METHODS

2.1 Isolated digit database

The Bellcore Isolated Digit database comprises eleven isolated digits ('one' to 'nine', 'zero', and 'oh') and two control words ('yes' and 'no') uttered by 200 speakers over dial-up telephone lines resulting in 2600 utterances [4]. The utterances of 151 speakers were used for training and the utterances of the remaining 49 speakers for testing. All utterances were automatically cut using a signal energy criterion that retained leading and trailing silence, though not so much as to necessitate silence models in the HMMs. Files also included artifacts such as the click-like transients that are naturally present in the data. Because both the RASTA and the auditory-model front ends require some filter initialization (e.g. RASTA requires 4 frames to "prime" its IIR filter), feature files were actually computed on extended sounds, that is, sounds that had been padded with an extra 100 ms of sound in front of the previously determined word margins. After feature computation on the extended sounds, the feature files were cut back to the originally determined word margins, discarding the initial section.

2.2 Front ends

The HMM recognition system was trained and tested using feature sets generated by four front ends:

- *MFCC:* Mel-frequency weighted cepstral coefficients were used as a benchmark to evaluate the performance of the recognizer.
- *J-RASTA and L-RASTA:* The RASTA technique estimates the time-varying spectrum based on the filtering of time trajectories of outputs from critical-band filters to achieve robustness against additive noise and spectral filtering [3,4]. The parameters of J-RASTA have been optimized by the method's originators for performance in the presence of additive noise, specifically Volvo noise. The parameters of L-RASTA have been optimized for performance in the presence of spectral distortion. The RASTA code we used was obtained from the International Computer Science Institute at Berkeley (ICSI).
- *Auditory model:* The auditory model front end comprises a bank of 120 discrete-time FIR filter channels whose frequency response was derived from the numerical solution of a three-dimensional cochlear hydrodynamic model representing the velocity of response of the cochlea's basilar membrane at positions logarithmically spanning the range of characteristic frequency (CF) from 250 to 3400 Hz. For

each channel, the instantaneous magnitude and phase of response were computed at an effective sample rate of 16 kHz by inverse Hilbert transformation. In previous studies, we have shown that formants can be detected "asynchronously" using only the instantaneous phase of response[6]. In the present study, our intent was to produce an auditory-model feature set which could be used as a direct replacement for frame-based spectral or MFCC features in conjunction with the HMM. Accordingly, we produced feature vectors based on the instantaneous magnitude of response as follows: at each time point, 1) the logarithm of the instantaneous magnitude of the array of channels was computed; 2) the log magnitude of each channel's output was then normalized by an AGC factor representing adaptation processes in the cochlea. For each channel, this AGC factor was an additive log gain proportional to a broad spatial average of 2.5 critical bands on either side of the channel, of the magnitudes of the output of the channels time averaged with a one-pole filter with a time constant of 200 msecs; 3) the normalized log-magnitude was saturated to a 30 dB dynamic range with a limiter representing the saturating nonlinearity of the average rate of auditory-nerve discharge; 4) the data were downsampled to a 100 Hz rate and the DCT performed so that the auditory-model feature vectors would appear to the HMM as a direct replacement of MFCCs.

2.3 HMM classifier

We employed an isolated word recognizer based on whole-word HMM models using the Entropics HMM toolkit HTK V2.0. The HMMs featured left-right state transition matrices consisting of nine states including non-emitting initial and final states. The HMM output probabilities were modeled with continuous density models comprising Gaussian mixture components with a diagonal covariance matrix. In initial experiments, we determined that full covariance matrices did not enhance recognition results.

For all front ends, feature vectors were generated at a rate of 100 Hz and a DCT was performed to produce cepstral coefficients. Regardless of the front end under test, the HMMs received as input vectors of the first nine of these cepstral coefficients (excluding the zeroth coefficient) plus the nine corresponding delta-coefficients, which were found to enhance the recognition performance significantly.

HMMs were initialized by uniformly dividing the feature data into nine segments with each segment associated with one state of the HMM model. To allow for a variable number of Gaussian mixture components, each state was initialized with one mixture component by computing the mean and standard deviation of all training data corresponding to the segment associated with this state. The subsequent training process employed the Baum-Welch re-estimation procedure. After each training cycle, the mixture component with the largest mixture weight was split by copying the mixture and dividing the weights of both mixtures by two. The means of the copies were modified by plus or minus 0.2 standard deviations. This training procedure and the splitting of the largest mixture component was repeated until the recognition performance achieved a maximum in the clean speech environment, which typically occurred after training 3 to 9 mixture components.

2.4 Training and testing procedure

The impact of environmental effects on recognition performance was assessed in a variety of acoustic environments: a "clean speech" environment; several environments in which additive noise of different types was added at varying amplitudes; and environments in which the speech was subject to spectral filtering at different bandwidths. The clean speech environment consisted of the unaltered Bellcore speech samples recorded in at standard telephone bandwidth in the presence of normally occurring baseline noise. To test the noise immunity of the front ends, white noise, pink noise, and Volvo noise were added to clean speech at various SNRs. The SNR was computed by taking the RMS signal value of the cut speech sample as reference. Pink noise was produced by filtering white noise with a one-pole lowpass filter at 250 Hz; Volvo noise was derived from data files obtained from ICSI. This type of noise was employed in an effort to match the real conditions that might confront a speech recognizer when the user is, for example, on a mobile phone in a car. Spectral distortion comprised a one pole IIR lowpass filter with different cut-off frequencies.

In this study, a number of training strategies were investigated. These included training HMMs on features derived from clean speech only and training on features derived from both clean speech and speech with additive white noise at 6 and 0 dB SNR. In all cases, the trained HMMs were tested in 12 environments: with clean speech; with white, pink and Volvo noise at three levels (6, 0 and -6 dB SNR); and with filtered speech at two bandwidths (125 and 250 Hz).

3. RESULTS

3.1 Training in clean speech

Figure 1a-c shows the recognition rates of digit recognizers trained with clean speech using auditory-model, MFCC, J-RASTA, and L-RASTA front ends and tested in white (a), pink (b), and Volvo noise (c) environments. The leftmost data point in each plot (marked with ∞ SNR) corresponds to the recognition rate in clean speech, and is fairly similar for all front ends. While Tong[8] reported a 96.8% recognition rate with the J-RASTA front end in clean speech, in our hands J_RASTA achieved a 98.9% recognition rate, an improvement we ascribe to the delta-features we used, which were not used by Tong.

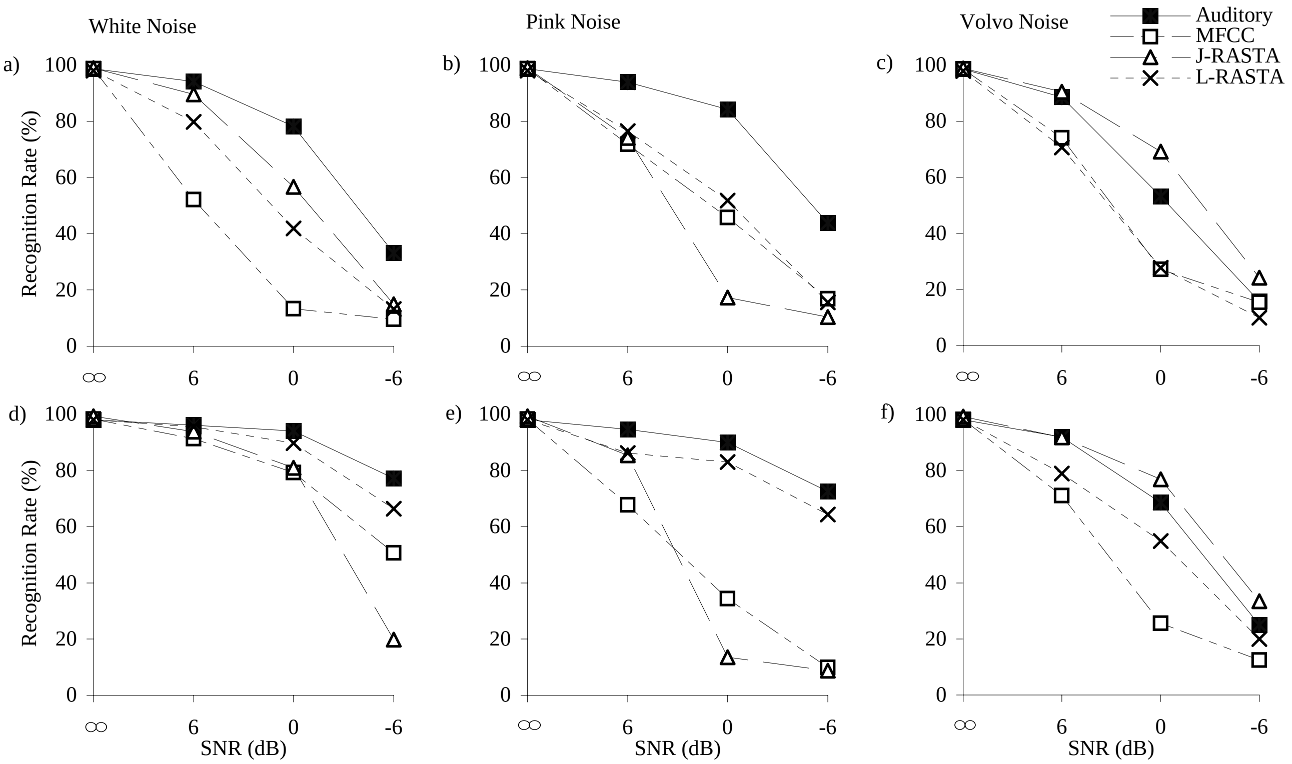

Figure 1: Recognition results for system trained with clean speech only and tested with white (a), pink (b), and Volvo noise (c). Results for a system trained with a mixture of clean speech and white noise and tested with white (d), pink (e), and Volvo noise (f).

The front ends showed very different noise-immunity to additive noise. In general, MFCC performed significantly worse than all other front ends in white noise but similar to L-RASTA in pink and Volvo noise. The performance of L-and J-RASTA variants appeared highly dependent on the characteristics of the noise. In pink noise, L-RASTA performed much better than J-RASTA while in Volvo noise the situation was reversed, showing that with RASTA, no single parameter set suffices to produce good performance for all three noise environments. In contrast, the performance of the auditory-model front end with a single parameter set appeared to be relatively good in all environments. Specifically, this front end demonstrated better performance in high-level white and pink noise than the L-RASTA, J-RASTA and MFCC front ends. For example, whereas for RASTA and MFCC the recognition rate in pink noise dropped significantly as the SNR degraded from 6 dB to 0 dB (J-RASTA from 74.1 to 17.3%, L-RASTA from 76.5 to 51.8%, and MFCC from 71.7 to 45.7%), the performance of the auditory-model front end started significantly higher and dropped less (from 93.9 to 84.1%). In white and pink noise environments, the performance advantage of the auditory-model front end is equivalent to an improvement in the SNR of 3 dB to 10 dB. In Volvo noise, J-RASTA performed better, perhaps due to its parameter optimization for this environment.

We determined that better performance for the auditory model front end could be obtained in specific environments if the parameters of the model were adjusted to that environment; however, we did not perform these optimizations since we are interested in producing a front end whose performance would be relatively independent of the environmental distortion imposed on the input signal, whose characteristics may not be known or constant.

3.2 Mixed training conditions

A second experiment involved training the HMMs with “mixed” features derived from both clean-speech and white-noise environments at 6 dB and 0 dB SNR. Figure 1d-f shows performance of systems trained with mixed features and tested with white (d), pink (e), and Volvo noise (f). Training and testing with mixed features had little effect on the recognition rate in clean speech, but markedly improved the recognition performance of all front ends in white noise, compared with the performance of the same front ends trained in clean speech alone (as might be expected). When tested with pink noise, the MFCC front end suffered a substantial performance loss. In contrast, the auditory-model front end demonstrated a “cross-training” benefit; that is, the recognition performance improved even in noise environments of a type *different* from the one used for training. The L-RASTA front end showed a similar cross-training effect while the J-RASTA front end benefited only at high SNR (low noise levels). Notwithstanding the performance effects of training with “mixed” features, the auditory-model still outperformed all other front ends except at the high levels of added Volvo noise.

3.3 Spectral distortion

Figure 2 shows the results of recognition experiments in which HMMs were tested with feature sets generated from speech that had been lowpass filtered at two corner frequencies, 125 and 250 Hz. With HMMs trained only on clean speech (a), the performance of both the auditory-model and the L-RASTA front end was extremely robust, even with speech filtered at a corner frequency of 125 Hz, achieving recognition rates (97.3% and 98.1% respectively) similar to those obtained in the clean speech environment (98.6 and 98.0% respectively). In contrast, the performance of the MFCC and J-RASTA front ends dropped significantly (from 98.6% to 80.9% for MFCC; from 98.9% to 80.1% for J-RASTA). When trained with "mixed" features in clean speech and white noise at 6 and 0 dB, the performance of the auditory-model and L-RASTA front ends appeared unaffected, the MFCC front end suffered a performance loss while the J-RASTA front end showed some cross-training benefit. Again, only one variant of RASTA (this time L-RASTA) showed good recognition performance on this task.

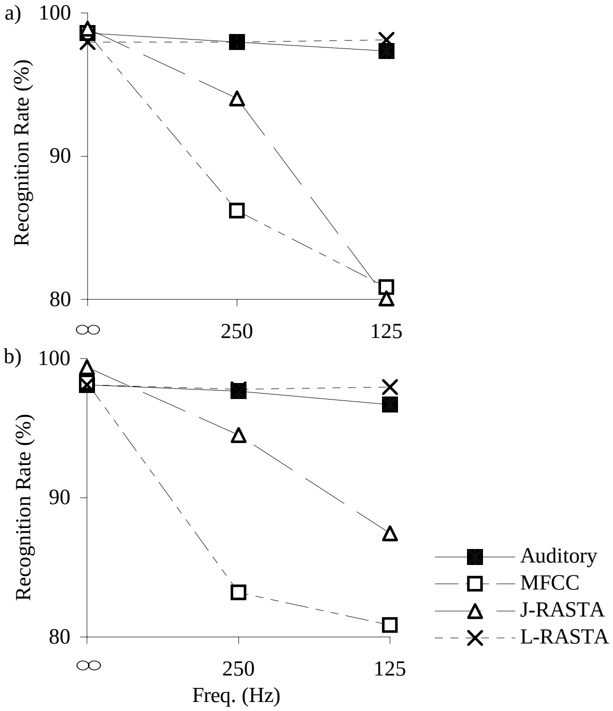

Figure 2: Recognition results for a system trained with clean speech only and tested with lowpass-filtered speech (a). Results for a system trained with a mixture of clean speech and white noise and tested with lowpass-filtered speech (b).

4. CONCLUSION

Summarizing our experiments, the auditory-model front end generally performed better than other front ends in the presence of noise and spectral distortion. Specifically, the auditory-model front end did better in white and pink noise, and better than all but J-RASTA in high levels of Volvo noise. The performance of the auditory model also degraded more gently in the range of 6 dB to 0 dB SNR. The performance of the auditory model in conditions of spectral distortion was better than MFCC and J-RASTA, and comparable to L-RASTA (while the auditory model's performance in noise was simultaneously better than either L-RASTA or MFCC).

Training HMMs with either the auditory-model or L-RASTA front end using "mixed" features derived from both clean speech and white noise yielded significant cross-training benefits when the system was tested with other noises; in contrast, the performance of the MFCC front end in mixed training degraded.

5. ACKNOWLEDGMENTS

We gratefully acknowledge the support of contracts from the U.S. DoD. We thank Doug Nelson for technical advice, Dan Kahn for permission to use the Bellcore digits database, and Brian Kingsbury at ICSI for access to the Volvo noise files. We would especially like to thank Hynek Hermansky at the Oregon Graduate Institute of Science and Technology for having let us use the local computational facilities to run the HMM experiments.

6. REFERENCES

[1] Acero, A., and Stern, R., *Environmental robustness in automatic speech recognition*. Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 849-52, 1990.

[2] Ghitza, O., *Auditory nerve representation as a basis for speech processing*. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, pp. 453-85. Marcel Dekker, New York, 1992.

[3] Hermansky, H., Morgan, N., Bayya, A, and Kohn, P. *Rasta–PLP Speech Analysis*. ICSI Technical Report TR-91-069, Berkeley, California

[4] Hermansky, H., Morgan, N., *RASTA Processing of Speech*, IEEE Transactions on Speech and Audio Processing, 2(4), 1994, pp. 578-89.

[5] Holton, T., and Love, S., *Robust pitch and voicing detection using a model of auditory signal processing*. Proc. of Int. Conf. of Spoken Language Processing, Yokohama, Japan, 1994.

[6] Hunke, M., and Holton, T., *Training Machine Classifiers to Match the Performance of Human Listeners in a Natural Vowel Classification Task*. Proc. IEEE Int. Conf. of Spoken Language Processing, Philadelphia, 1996.

[7] Levinson, S., Rabiner, L., and Sondhi, M., *Speaker independent isolated digit recognition using hidden Markov models*, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1983, pp. 1049-1052.

[8] Tong, G., *Combating Additive Noise and Spectral Distortion in Speech Recognition Systems with JAH-RASTA*, Masters Thesis, University of California at Berkeley, 1994.

[9] Rabiner, L., Levinson, S., and Sondhi, M., *On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated-Word Recognition*, Bell Systems Technical Journal, 62, No. 4 1983, pp. 1075-105.

[10] Rabiner, L., et al., *Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities*, AT&T Technical Journal, 64, No. 6, 1985