REDUCED CHANNEL DEPENDENCE FOR SPEECH RECOGNITION

Hy Murveit, John Butzberger, and Mitch Weintraub

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

1. ABSTRACT

Speech recognition systems tend to be sensitive to unimportant steady-state variation in speech spectra (i.e. those caused by varying the microphone or channel characteristics). There have been many attempts to solve this problem; however, these techniques are often computationally burdensome, especially for real-time implementation. Recently, Hermansky et al. [1] and Hirsch et al. [2] have suggested a simple technique that removes slow-moving linear channel variation with little adverse effect on speech recognition performance. In this paper we examine this technique, known as RASTA filtering, and evaluate its performance when applied to SRI's DECIPHER™ speech recognition system [3]. We show that RASTA filtering succeeds in reducing DECIPHER™'s dependence on the channel.

2. INTRODUCTION

A number of techniques have been developed to compensate for the effects that varying microphone and channels have on the acoustic signal. Erell and Weintraub [4, 5] have used additive corrections in the filter-bank log energy or cepstral domains based on equalizing the long-term average of the observed filter-bank log energy or cepstral vector to that of the training data. The techniques developed by Rose and Paul [6] and Acero [7] used an iterative technique for estimating the cepstral bias vector that will maximize the likelihood of the input utterance. Nadas et al. [8] used an adaptive linear transformation applied to the input representation, where the adaptation uses the VQ distortion vector with respect to a predefined codebook. VanCompernolle [10] scaled the filter-bank log energies to a specified range using running histograms, and Rohlicek [9] experimented with a number of histogram-based compensation metrics based on equalizing different aspects of the probability distribution.

One important limitation of the above approaches is that they rely on a speech/nonspeech detector. Each of the above approaches computes spectral properties of the input speech sentence and subsequently compensates for the statistical differences with certain properties of the training data. If the input acoustical signal is not segmented by sentence (e.g. open microphone with no push-to-talk button) and there are long periods of silence, the above approaches would not be able to operate without some type of reliable automatic speech-input/sentence-detection mechanism. An automatic sentence-detection mechanism would have considerable difficulty in reliably computing the average speech spectrum if there were many other nonspeech sounds in the environment.

A second class of techniques developed around auditory models (Lyon [11]; Cohen [12]; Seneff [13]; Ghitza [14]). These techniques use various automatic gain control and other auditory-type modeling techniques to output a spectral vector that has been adapted based on the acoustic history. A potential limitation of this approach is that many of these techniques are very computationally intensive.

3. THE RASTA FILTER

RASTA filtering is a high-pass filter applied to a log-spectral representation of speech. It removes slow-moving variations from the log spectrum. The filtering is done on the log-spectral representation so that multiplicative distortions (such as a linear filter) become additive and may be removed with the RASTA filter. A simple RASTA filter may be implemented as follows:

$$y(t) = x(t) - x(t-1) + (C \cdot y(t-1))$$

where x(t), as implemented in DECIPHER™, is a log band-pass energy which is normally used in DECIPHER™ to compute the Mel-cepstral feature vector. Instead, x(t) is replaced by y(t), the high-pass version of x(t), when performing the cepstral transform.

The constant, C, in the above equation defines the time constant of the RASTA filter. It is desirable that C be such

that short-term variations in the log spectra (presumably important parts of the speech signal) are passed by the filter, but slower variations are blocked. We set C = 0.97 so that signals that vary faster than about 1 Hz are passed and those that vary less than once per second tend to be blocked. Figure 1 below plots the characteristic of this filter.

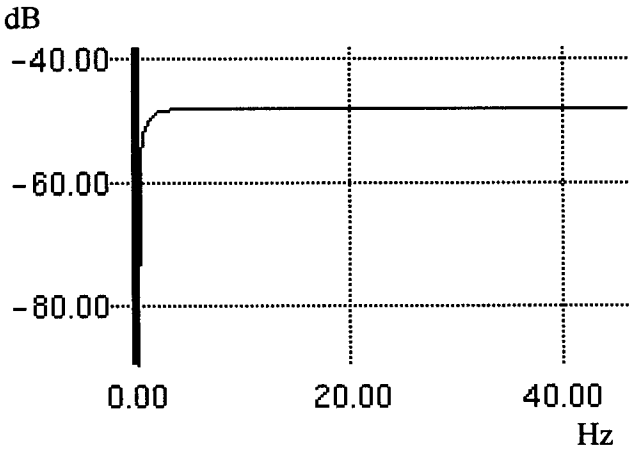

Figure 1: Characteristics of the C = 0.97 RASTA filter

When used in conjunction with SRI's spectral estimation algorithms [4, 5], the high-pass filter is applied to the filter-bank log energies after the spectral estimation operation. The estimates of clean filter-bank energies are highpass filtered and then transformed to obtain the cepstral vector. The cepstral vector is then differenced twice to obtain the delta-cepstral vector and the delta-delta-cepstral vector.

3.1. Removal of an Ideal Linear Filter

We first evaluated RASTA filtering by applying a bandpass filter (Figure 2 below) to a speech recognition task—continuous digit recognition performance over telephone lines. The filter was applied to the test set only (no filtering was applied to the training data). We compared the resulting performance with the performance of an unfiltered test set for both standard and RASTA filtering. As Table 1 shows, the RASTA filtering was successful in removing the effects of the bandpass filter, whereas the standard system suffered a significant performance degradation due to the bandpass filter. Compared with our standard signal processing, the RASTA filtering was able to give a slight improvement on the female digit error rate, with no significant change in the male digit error rate. The dramatic decrease in performance that occurs when the telephone speech is bandpass filtered is removed by the RASTA filtering, and the results are comparable to the original speech signal.

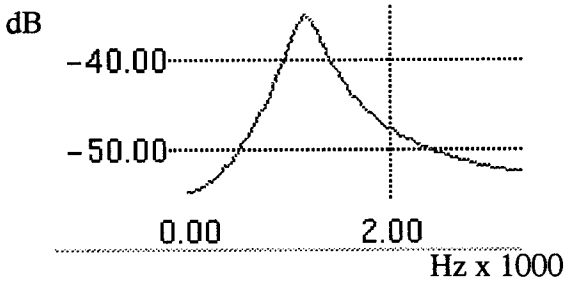

Figure 2: The distorting bandpass filter characteristic.

	Original Speech		Bandpass Speech	
	male	female	male	female
Standard	3.2	3.1	13.9	11.6
RASTA	3.4	2.1	3.0	1.9

Table 1: Word error rates for standard signal processing techniques and RASTA filtering techniques using clean and bandpass-filtered telephone speech.

4. REDUCED MICROPHONE DEPENDENCE

After the encouraging initial study, we tested RASTA filtering in a more realistic manner—measuring the performance improvement, due to RASTA filtering, when dissimilar microphones are used in the test and training data.

To do this, we recorded 50 sentences (352 words) from one talker simultaneously using two different microphones, a Sennheiser flat-response close-talking microphone that was used to train the system, and an Electrovoice 625 handset with a very different frequency characteristic. The user spoke queries for DARPA's ATIS air-travel planning task. Table 2 shows that for this speaker, the error rate was less sensitive to the difference in microphone when RASTA filtering was applied than when it wasn't. Further, there is no evidence from this and the previous study to indicate that RASTA filtering degrades performance when the microphone remains constant.

	Sennheiser	Electro Voice
Standard	13 (3.7%)	31 (8.8%)
RASTA	12 (3.4%)	17 (4.8%)

Table 2: Number and percentage of word errors for a single speaker when test microphone and signal processing were varied.

5. DESKTOP MICROPHONES

RASTA filtering is most effective when differences between training and testing conditions can be modeled as linear filters. However, many distortions do not fit this model. One example is testing with a desktop microphone with models trained with a close-talking microphone. In this scenario, although the microphones characteristics may be approximately related with a linear filter, additive noise picked up by the desktop microphone violates the linear-filter assumption.

To see how important these effects are, we performed recognition experiment on systems trained with sennheiser microphones and tested with a Crown desktop microphone. These test recordings were made at Carnegie Mellon University (CMU) and at the Massachusetts Institute of Technology (MIT). They simultaneously recorded a speaker using both Sennheiser and Crown microphones interacting with an ATIS (air travel planning) system.

The performance of DECIPHER™ on the ATIS recordings is shown in Tables 3 and 4. Table 3 shows the system performance results on MIT's recordings, while Table 4 contains the system performance results on CMU's recordings.

Speaker	Sennheiser	Crown	Crown	Crown	Crown
	Standard	Standard	RASTA	NRFE	NRFE+RASTA
4V	13.0	13.8	22.8	18.7	16.3
4W	1.7	5.1	1.7	4.3	3.4
5E	17.8	26.6	27.8	18.1	14.7
55	18.5	26.6	25.3	23.2	17.6
59	13.7	40.2	41.0	26.6	23.6
Average	12.9	22.5	23.7	18.2	15.1

Table 3: Word error rate for MIT recordings varying microphone and signal processing

Speaker	Sennheiser	Crown	Crown	Crown	Crown
	Standard	Standard	RASTA	NRFE	NRFE+RASTA
IF	20.7	91.8	46.9	46.9	36.7
IH	20.5	93.2	75.7	71.0	35.8
IK	26.2	87.1	62.3	60.3	35.8
Average	22.5	90.7	61.6	59.4	36.1

Table 4: Word error rate for CMU recordings varying microphone and signal processing

For the MIT recordings, note that the best performing system on the Crown microphone data was very close with the performance on the Sennheiser recordings (12.9% vs. 15.1%). The addition of RASTA processing did not help the standard processing on the Crown data (the error rate went up slightly from 22.5% to 23.7%) but it did help the noise-robust estimation processing (18.2% to 15.1%).

The performance on CMU's Crown recordings were much lower. CMU's audio recordings for were noticeably noisier; the speaker sounded as if he was much farther from the microphone, and there were other nonstationary sounds in the background. Note that the error rate with the standard signal processing is extremely high (90.7% word error). For the CMU Crown microphone recordings, the addition of RASTA processing helped reduce the error rate for both the standard and noise-robust estimation processing conditions. The NRFE + RASTA processing was able to reduce the error rate by 60% over the no-processing condition on the CMU Crown microphone recordings (90.7% to 36.1%).

SRI's noise-robust spectral estimation algorithms are designed to estimate the filter-bank log energies of the clean speech signal when there is additive colored noise. The estimation algorithms were designed to work independently from any spectral shape introduced by the microphone and channel variations. Therefore, some type of additional spectral normalization is required to compensate for these effects: the combined "NRFE + RASTA" system serves that purpose. The RASTA system (without estimation) can help compensate for the linear microphone effects, but it can help only to a limited degree with the nonlinearities introduced by other sounds.

6. ROBUSTNESS OF REPRESENTATION TO MICROPHONE VARIATION

To understand the benefit that we have obtained using the different processing techniques, we developed a metric for the robustness of the representation that is separate from speech-recognition performance. The DARPA CSR corpus (Doddington [15]) was used for this evaluation since it is contains stereo recordings. By using stereo recordings, we can compare the robustness in the representation that occurs when the microphone is changed. In this CSR corpus, the first channel of these stereo recordings is always a Sennheiser close-talking microphone. The second recording channel uses one of 15 different secondary microphones.

Using this stereo database, we can compute the cepstral feature vector on each microphone channel, and compare the two representations to determine the level of invariance provided by the signal-processing/representation. The metric that we used for determining the robustness of the representation is called relative-distortion and is computed in the following equation.

$$\text{Relative Distortion } (C_i) = \frac{\sqrt{(C_{i(Mic1)} - C_{i(Mic2)})^2}}{\sigma_{C_{i(Mic1)}} \cdot \sigma_{C_{i(Mic2)}}}$$

The relative distortion for cepstral coefficient C_i is computed by comparing the cepstral value of the first microphone with the same cepstral value computed on the secondary microphone. This average squared difference is then normalized by the variance of this cepstral feature on the two microphones. This metric gives an indication of how much variance there is due to the microphone differences relative to the overall variance of the feature due to phonetic variation. This metric is plotted as a function of the cepstral coefficient for different signal processing algorithms in figure 3.

Figure 3 shows that the RASTA processing helps reduce the distortion in the lower order cepstral coefficients. When combined with SRI's noise-robust spectral estimation algorithms, the distortion decreases even further for the lower order cepstral coefficients. Neither of the algorithms help reduce the distortion for the higher cepstral coefficients. This metric indicates that even though the robust signal processing has reduced the recognition error rate due to microphone differences, there is still considerable variation in the cepstral representation when the microphone is changed.

7. SUMMARY

We have shown that high-pass filtering of the filter-bank log energies can be an effective means of reducing the effects of some microphone and channel variations. We have shown that such filtering can be used in conjunction with our previous estimation techniques to deal with both noise and microphone effects. The high-pass filtering operation is a simple technique that is computationally efficient and has been incorporated into our real-time demonstration system.

REFERENCES

1. H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the Effects of the Communication Channel in Auditory-Like Analysis of Speech," *Eurospeech*, Sept. 1991, pp. 1367-1370.

2. H. Hirsch, P. Meyer, and H.W. Ruehl, "Improved Speech Recognition using High-Pass Filtering of Subband Envelopes," *Eurospeech*, Sept. 1991, pp. 413-416.

3. H. Murveit, J. Butzberger, and M. Weintraub, "Speech Recognition in SRI's Resource Management and ATIS

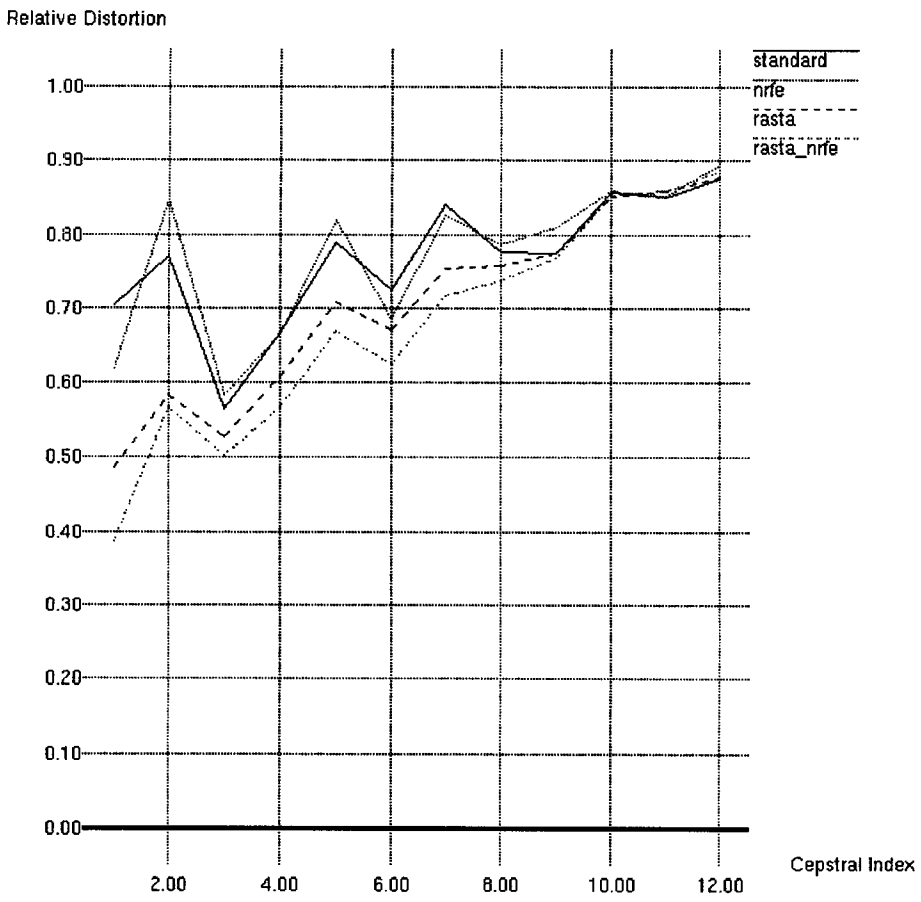

Figure 3: Relative distortion plotted as a function of the cepstral index for different signal processing algorithms (standard, NRFE, RASTA, and RASTA + NRFE).

Systems," DARAP SLS Workshop, February 1991, pp. 94-100.

4. A. Erell, and M. Weintraub, "Spectral Estimation for Noise Robust Speech Recognition," DARPA SLS Workshop October 89, pp. 319-324.

5. A. Erell, and M. Weintraub, "Recognition of Noisy Speech: Using Minimum-Mean Log-Spectral Distance Estimation," DARPA SLS Workshop, June 1990, pp. 341-345.

6. R. Rose and D. Paul, "A Hidden Markov Model Based Keyword Recognition System," *IEEE ICASSP* 1990, pp. 129-132.

7. A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph.D. Thesis Carnegie-Mellon University, September 1990

8. A. Nadas, D. Nahamoo, M. Picheny, "Adaptive Labeling: Normalization of Speech by Adaptive Transformations based on Vector Quantization" *IEEE ICASSP* 1988, pp. 521-524.

9. R. Rohlicek, W. Russell, S. Roukos, H. Gish, "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," *IEEE ICASSP* 1989, pp. 627-630

10. D. VanCompernolle, "Increased Noise Immunity in Large Vocabulary Speech Recognition with the Aid of Spectral Subtraction," *IEEE ICASSP* 1987, pp 1143-1146.

11. R. Lyon, "Analog Implementations of Auditory Models," DARPA SLS Workshop, Feb. 1991 pp. 212-216.

12. J. Cohen, "Application of an Auditory Model to Speech Recognition," *Journ. Acoust. Soc. Amer.*, 1989, 85(6) pp. 2623-2629.

13. S. Seneff, "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing," *Jour. Phonetics*, January 1988

14. O. Ghitza, "Auditory Neural Feedback as a Basis for Speech Processing," 1988 *IEEE ICASSP*, pp. 91-94.

15. Doddington, G., "CSR Corpus Development," DARPA SLS Workshop, Feb 1992.