RECOGNITION OF SPEECH IN ADDITIVE AND CONVOLUTIONAL NOISE BASED ON RASTA SPECTRAL PROCESSING

Hynek Hermansky‡,†, Nelson Morgan †, and Hans-Gunter Hirsch *,†

† International Computer Science Institute, Berkeley, California
‡ U S WEST Advanced Technologies, Boulder, Colorado
* University of Aachen, Germany

RASTA speech processing was originally developed to reduce the sensitivity of recognizers to frequency characteristics of an operating environment (i.e., to convolutional noise). RASTA does this by band-pass filtering time trajectories of logarithmic parameters of speech (e.g., logarithmic spectral energies or cepstra). In our current paper we study RASTA processing in an alternative spectral domain which is linear-like for small spectral values and logarithmic-like for large spectral values. We show on experiments with a recognizer trained on the clean speech and test data degraded by both convolutional and additive noise that doing RASTA processing in the new domain yields results comparable to results obtained by training the recognizer on known noise.

INTRODUCTION

An operating acoustic environment for a practical recognizer (room, microphone, telecommunication channel, ...) has its own frequency characteristics and may also be noisy. TABLE shows the results of an isolated word recognition experiment in which the recognizer was operating on data which were subject to linear filtering (convolutional noise) and to which the noise was added (additive noise). (Details of the experimental setup and recognition task are described in the Appendix).

Section I of TABLE shows the recognizer accuracy when training is on data with an environment identical to that for the test data. That is, the recognizer was always trained on the data which were subject to the identical distortion as was the test. As long as each operating environment is well represented in the training, the recognizer typically performs well.

Unfortunately, the noise is seldom known in advance. When the data from different environments is used in training and test, the same recognizer typically performs much worse. This situation is illustrated in all remaining sections of TABLE, which show recognition accuracies for the recognizer trained on the clean data and used on the noisy data.

Our goal is to understand and eliminate variance in the speech signal due to the environmental changes and thus ultimately avoid the need for extensive training of the recognizer in different environments. As indicated by the last section VI of the TABLE, our new method is comparable to training on noisy data.

S/N ratio in dB additive noise				**S/N ratio in dB additive and convolutional noise**			
>20	20	10	0	**>20**	**20**	**10**	**0**
all results in % correct							
TRAINED AND TESTED ON DATA WITH IDENTICAL NOISE							
I. PLP							
88.0	87.9	82.8	68.8	**86.0**	**83.2**	**78.5**	**64.6**
TRAINED ON CLEAN, TESTED ON NOISY DATA							
II. PLP							
88.0	88.3	56.6	47.1	**40.3**	**41.0**	**32.5**	**17.4**
III. RASTA-PLP							
87.8	75.0	57.9	32.3	**80.1**	**67.1**	**50.8**	**30.0**
IV. PLP with noise stripping							
90.3	88.3	84.9	69.7	**53.0**	**58.3**	**58.0**	**46.7**
V. RASTA-PLP with noise stripping							
85.7	77.9	67.5	47.9	**75.3**	**68.1**	**60.0**	**42.1**
VI. ADAPTIVE LIN-LOG RASTA-PLP							
88.6	90.3	84.9	73.2	**79.1**	**77.5**	**74.3**	**60.7**

WHO FEELS IT, KNOWS IT

0-7803-0946-4/93 $3.00 © 1993 IEEE

RASTA AND ITS SENSITIVITY TO ADDITIVE NOISE

Our original RASTA (RelAtive SpecTrAl) processing [Hermansky, Morgan, Bayya and Kohn 1991] was designed to alleviate logarithmic spectral components with rates of change outside the typical rate of change of speech spectral components. By operating in the logarithmic spectral domain, RASTA effectively diminishes spectral components that are additive in the logarithmic spectral domain, in particular the fixed or slowly-changing spectral characteristics of the environment (convolutive in the time domain and therefore additive in the log spectral domain). However, uncorrelated additive noise components that are additive in the power spectral domain became signal-dependent after the logarithmic operation on the spectrum and cannot be effectively removed by RASTA band-pass filtering in the logarithmic domain. Thus, as shown in the section III of the TABLE, the original RASTA processing on the logarithmic spectrum or cepstrum is not particularly appropriate for speech with significant additive noise.

PRE-PROCESSING OF NOISY SPEECH

The standard method for dealing with unknown noise is the spectral subtraction technique (see e.g. [Kang]) in which the power or magnitude spectrum of the noise, estimated in "silence" intervals between speech signals, is subtracted from the spectrum of noisy speech. Various adaptive techniques are typically used to update the spectral estimate of the slowly varying noise. Negative power spectral values after subtraction are common and need to be dealt with.

Similarly to spectral subtraction, RASTA processing on the power spectrum should alleviate the spectral components due to additive noise. [Hirsch, Mayer and Ruehl] reports good results with RASTA-like processing in the power spectral domain. However, since RASTA band-pass filtering attempts to set the mean value of all processed parameters to zero, it makes about half of the power spectral values negative. Therefore, just as in the case of spectral subtraction, the negative power spectrum requires some *ad hoc* post-processing.

An alternative strategy is to clean up the noisy speech prior to sending it to the recognizer. We have had some success with noise suppression using RASTA processing instead of a conventional spectral subtraction in the overlap-add analysis-synthesis of [Kang]. The RASTA processing was done on the cube root of the power spectrum with a subsequent setting of all negative spectral values to a small positive constant. We have observed that such processing is comparable to the conventional adaptive spectral subtraction technique while being conceptually much simpler. Sections IV and V of TABLE show the results of cascading the overlap-add RASTA processing with PLP and RASTA-PLP based recognizers from the previous experiment.

The pre-processing with PLP analysis is quite effective. Performance in the presence of convolutional noise improves slightly. Applying the noise-suppressing pre-processing with RASTA-PLP softens the degradation of the original RASTA-PLP based system in noise. However, considering that on the clean speech, the pre-processing causes about 20% increase in the error rate, our results essentially confirms [Accero and Stern] which reports negative experience with cascading two systems, one dealing with the additive and other with the convolutive noise.

LIN-LOG RASTA

In [Morgan and Hermansky 1992] we have proposed as a substitute for the logarithmic transform in RASTA processing the function

$$y=\ln(1+J\cdot x), \qquad (1)$$

where J is a signal-dependent positive constant. The amplitude-warping transform (1) is linear-like for $J\ll 1$ and logarithmic-like for $J\gg 1$. Its inverse

$$x=(e^{y}-1)/J, \qquad (2)$$

where e is the base of natural logarithm, is not guarantied positive for all y and, like the conventional spectral subtraction, would require some *ad hoc* magic to ensure the positivity of the processed power spectrum. To avoid this we use an approximate inverse transform

$$x=e^{y}/J. \qquad (3)$$

This inverse is equivalent to the sum of the exact inverse and and additive constant $1/J$. It is therefore more inaccurate for small spectral values than for the larger ones.

Isolated Digit Experiment

We repeated the earlier isolated word recognition experiments using the nonlinearities (1) and (3). The results shown in Fig. 1 were generated using a number of different values of J . There is a distinct optimal value of J for each particular noise level. The optima are always better than either PLP or RASTA-PLP result.

With similar results we have also experimented with a transform pair

$$y=J\cdot x/e, \text{ for } J\cdot x<e,$$

$$y=\log(J\cdot x), \qquad \text{for } J\cdot x\le e, \qquad (4)$$

and its approximate inverse

$$x=e^{y}/J. \qquad (5)$$

The inverse of (4) using (5) is exact for $J\cdot x>e$. Results using either transform pair Eqs. (1) and (3), or Eqs. (4) and (5) are generally very similar and, throughout the paper, we only give results using the first transform pair. This comparison indicates that the exact form of the nonlinearity may not be crucial, as long as it is roughly linear for small arguments and logarithmic for large arguments.

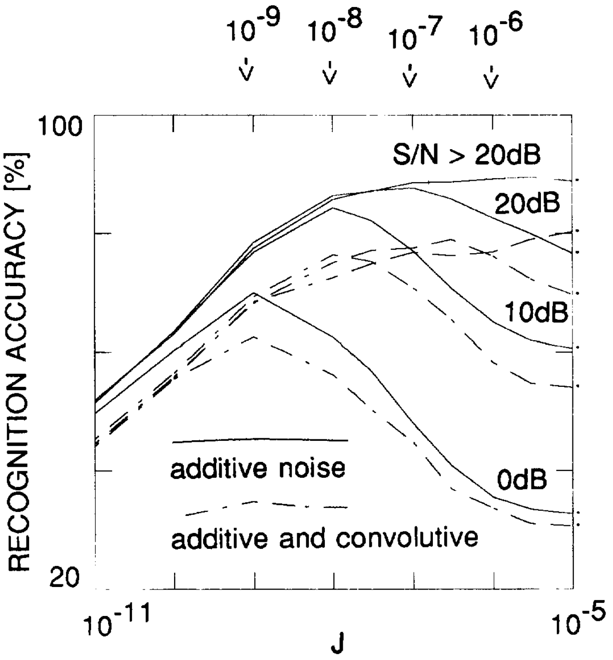

Fig. 1 Digit Recognition

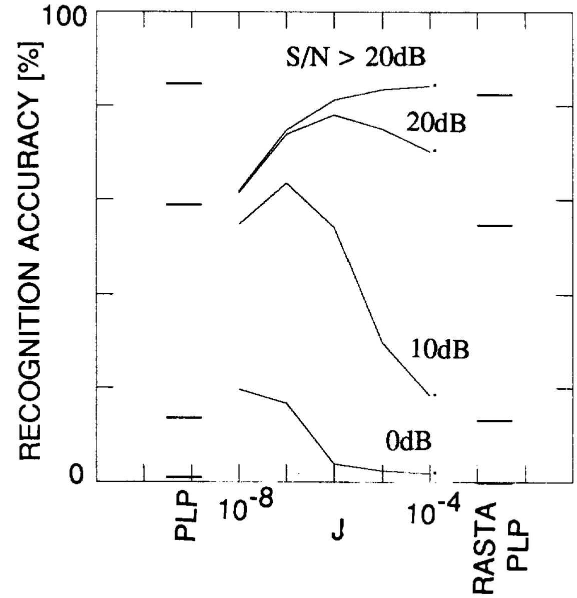

Fig. 2 DARPA Resource Management Recognition

DARPA Task Experiment

All the experiments above were done with a simple DTW recognizer on a small isolated word recognition task. This recognizer and task were chosen for the exploratory research where we had to repeat the recognition experiment manv times. To see whether our approach would scale to large tasks, we used a standard DARPA Resource Management recognition task and hybrid neural network/HMM recognizer (see the Appendix for further description). Fig. 2 shows the results for additive noise. Roughly the same pattern as was observed in the earlier digit experiment can be seen here: there is an optimal value of *J* for each particular S/N ratio. Smaller values of *J* are preferred for noisy speech.

Rationale for the optimal J.

Results shown in Figs. 1 and 2 indicate that there is a particular optimal value of *J* for each particular SNR case in the test data. Fig. 3 shows histograms of logarithmic auditory-like spectral energies *x* for all four SNRs that were used. Spectral values for which $J_{optimal} \cdot x = e$ for all four investigated S/N ratios are indicated in the figure by arrows. Supporting [Van Compernolle], the histograms are multi-modal. Assuming that the strongest mode represents noise, the optimal value of *J* is such that it puts most of the signal into the logarithmic-like part of our nonlinearity and most of the noise into its linear-like part.

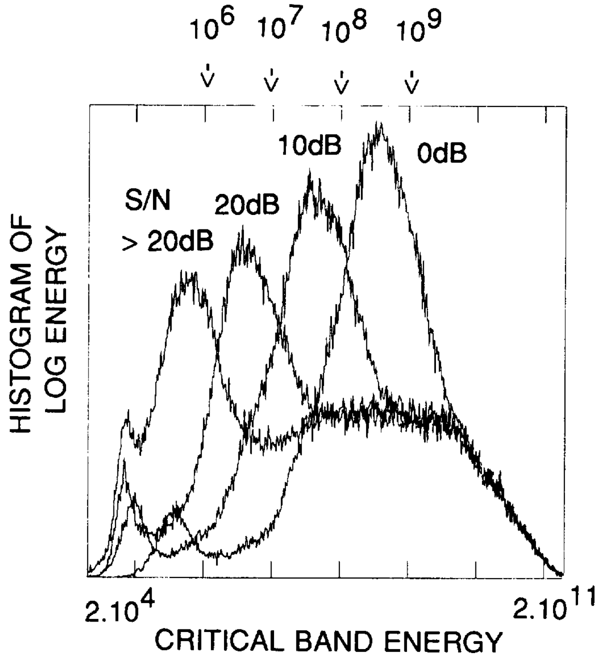

Fig. 3 Histograms of critical-band energies

Adaptive adjustment of the optimal J.

In the experiments described above, the same value of J was used in both the training and the operation 'of the recognizer. This would then require that the analysis for both the training and the operation of the recognizer would change depending on the noise level during the operation. As indicated in Fig. 3, the optimal J seems to depend on the noise level. We have pursued this and measured the mean critical band energy in the first 125 ms of the utterance (there was no speech in this part of utterance in our data). Then, we made J inversely dependent on such measured mean noise energy E_{noise}, i.e.

$$J=1.0/(C{\cdot}E_{noise}).$$

Since the particular value of J influences the shape of the resulting all-pole model spectrum, it would be desirable from the model-matching perspective to use identical J on both the training and the test data. However, as shown in Fig.1, the J depends on the level of noise in the signal which is generally different for the training and the test data. We have approached this problem by using four different sets of templates in the recognizer, each set trained with an order-of-magnitude different C_{train}. namely

$$C_{train}=3{\cdot}10^3,\ 3{\cdot}10^2,\ 3{\cdot}10^1,\ \textit{and}\ 3.$$

Thus, comparing to our standard recognizer, 4 times as many templates are used. Constant C during the test of the recognizer was fixed at

$$C_{test}=3.$$

Results from such an automatically adaptive system are shown in section VI of the TABLE .

DISCUSSION AND SUMMARY

The logarithm or the absolute value are mathematically convenient nonlinearities. There is no particular reason to believe they are optimal for processing natural signals such as speech. We have experimented with RASTA temporal processing in a new spectral domain which is approximately linear for small spectral values and approximately logarithmic for large ones. When such processing was applied in a recognizer trained on clean speech, then, without any explicit knowledge of the noise in the test data, results were comparable with those from the same recognizer trained on noisy data.

ACKNOWLEDGEMENTS

We thank Yochai Konig for help with running DARPA experiments, and Aruna Bayya, Steve Greenberg and the ICSI speech research team for stimulating discussions.

REFERENCES:

A. Accero and R. Stern: Environmental robustness in automatic speech recognition, Proc. ICASSP '90, pp. 849-852, Albuquerque, NM.

R. Duda and P. Hart: Pattern Classification and Scene Analysis, Wiley 1973.

H. Hermansky: Perceptual linear predictive (PLP) analysis for speech. J. Acoust. Soc. Am., pp. 1738-1752, 1990

H. Hermansky and N. Morgan: Towards handling the acoustic environment in spoken language processing. Proc. of Intl. Conf. on Spoken Lang. Processing

H. Hermansky, N. Morgan, A. Bayya, and P. Kohn: Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP) Proc. EUROSPEECH '91, pp. 1367-1370, Genova, 1991.

H. Hermansky and J.C. Junqua: Optimization of perceptually based ASR front-end. Proc. ICASSP '88, pp. New York, 1988, pp.

H.G. Hirsch, P. Meyer, and H. Ruehl: Improved speech recognition using high-pass filtering of subband envelopes. Proc. EUROSPEECH '91, pp. 413-416, Genova, 1991.

G.S. Kang and L.J. Fransen: Quality improvement of LPC-processed noisy speech by using spectral subtraction, Trans. IEEE-ASSP 37, (6), pp. 939-942, June 1989.

N. Morgan and H. Hermansky: RASTA extensions: Robustness to additive and convolutional noise, Proc. Workshop on Speech Processing in Adverse Environments, Cannes, France, November 1992.

N. Morgan, H. Hermansky, H. Bourlard, P. Kohn, and C. Wooters: Continuous Speech Recognition Using PLP Analysis with Multilayer Perceptrons, Proc. ICASSP '91, pp. 49-52, Toronto, Canada

D. Van Compernolle: Noise adaptation in a hidden markov model speech recognition system, personal communications, submitted for publication, 1987.

APPENDIX: EXPERIMENTAL SETUP

The following experimental setup is being used for the isolated digit experiments described in this paper:

11 isolated digits and two control words ("yes" and "no") were recorded at 8 kHz by 30 talkers over dialed-up telephone lines. All words were hand end-pointed. The recognizer was a DTW-based nearest-neighbor multi-template recognizer. 27 talkers out of 30 were used in for training of the recognizer in a "leave-three-out" experimental design [Duda and Hart]. In the "leave-three-out" design, three templates out of 30 are held for test and remaining 27 templates per each utterance are treated as training data. All possible unique choices of 27 templates out of available 30 were used, thus yielding 52780 recognition trials per experimental point.

Recognition features were exponentially-weighted [Hermansky and Junqua 1988] (exp = 0.6) five cepstral coefficients (zeroth coefficient excluding) of the 5th order PLP on RASTA-PLP model, computed from a 25 ms analysis window with a 12.5 ms analysis step.

The data were also degraded by realistic additive noise, recorded over cellular telephone from a 1978 VOLVO 244 with windows closed, running at 55 miles/hour on a freeway. The noise was added at several signal-to-noise (S/N) ratios. The S/N ratios given in the paper represent ratios between the averaged energy over the whole utterance and the averaged energy of the added noise.

To introduce convolutional noise, linear filtering simulating the difference between frequency response of the carbon microphone and the electret microphone in the telephone handset was applied.

In the continuous speech experiment, the noise described above was added to 600 standard test sentences from the February 1989 and October 1989 DARPA Resource Management evaluation sets. The standard 3990 Resource Management training sentences were used to generate a layered neural network to estimate phonetic probabilities for a Hidden Markov Model (HMM), 8 cepstral, 8 δ cepstral and 8 $\delta * \delta$ cepstral coefficients (zeroth cepstral coefficient including) of the 8th-order PLP or RASTA-PLP all pole model over the 9 frame window [Morgan et al. 1991] were used as the features. Both the network and the HMM were somewhat simpler than the ones used for our best recognizer in order to conserve computational resources for our front end experiments.