RASTA-PLP SPEECH ANALYSIS

Hynek Hermansky [*]
Nelson Morgan [†]
Aruna Bayya [*]
Phil Kohn [†]

TR-91-069

December 1991

Abstract

Most speech parameter estimation techniques are easily influenced by the frequency response of the communication channel. We have developed a technique that is more robust to such steady-state spectral factors in speech. The approach is conceptually simple and computationally efficient. The new method is described, and experimental results are reported, showing a significant advantage for the proposed method.

[*] US West Advanced Technologies, 4001 Discovery Drive, Boulder, CO 80303

[†] International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704

1 INTRODUCTION

The Perceptual Linear Predictive (PLP) speech analysis technique [1] is based on the short-term spectrum of speech. Even though the short-term spectrum of speech is subsequently modified by several psychophysically based spectral transformations, the PLP technique (just like most other short-term spectrum based techniques), is vulnerable when the short-term spectral values are modified by the frequency response of the communication channel. Human speech perception seems to be less sensitive to such steady-state spectral factors [2]. We have developed the RelAtive SpecTrAl (RASTA) methodology [3][4] which makes PLP (and possibly also some other short-term spectrum based techniques) more robust to linear spectral distortions. Experimental results using telephone-quality isolated digits and high-quality continuous speech show significant improvements in error rate.

2 APPROACH

We have replaced a common short-term absolute spectrum by a spectral estimate in which each frequency channel is band-pass filtered by a filter with sharp spectral zero at the zero frequency. Since any constant or slowly-varying component in each frequency channel is suppressed by this operation, the new spectral estimate is less sensitive to slow variations in the short-term spectrum. When the filtering is done in the logarithmic spectral domain, the suppressed constant spectral component reflect the effect of the convolutive factors in the input speech signal, introduced by frequency characteristics of the communication media.

The steps of RASTA-PLP are as follows (see [1] for comparison to the conventional PLP method):

For each analysis frame:

1) Compute the critical-band spectrum (as in the PLP) and take its logarithm.

2) Estimate the temporal derivative of the log critical-band spectrum using regression line through five consecutive spectral values [5].

3) Nonlinear processing (such as applying threshold or median filtering) can be done in this domain. Currently, we do nothing here.

4) Re-integrate the log critical-band temporal derivative using a first order IIR system. The pole position of this system can be adjusted to set the effective window size. Currently, we set this value to 0.98, providing an exponential integration window with a 3-dB point after 34 frames.

5) In accord with the conventional PLP, add the equal loudness curve and multiply by 0.33 to simulate the power law of hearing.

6) Take the inverse logarithm (exponential function) of this relative log spectrum, yielding a relative auditory spectrum.

6) Compute an all-pole model of this spectrum, following the conventional PLP technique.

It can be shown that if the derivative of step (2) is estimated by a simple first difference, and if the full integration in step (4) is done (pole at z = 1.0), then all intermediate terms cancel and the technique is equivalent to subtraction of the log spectrum of the first analysis frame from each new frame. In this special case, the RASTA technique resembles the spectral subtraction or blind deconvolution techniques.

However, in the general case presented here, the whole derivative-reintegration process is equivalent to a bandpass filtering of each frequency channel through an IIR filter with the transfer function

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})}. \tag{1}$$

The low cut-off frequency of the filter determines the fastest spectral change of the log spectrum which is ignored in the output, while the high cut-off frequency determines the fastest spectral change which is preserved.

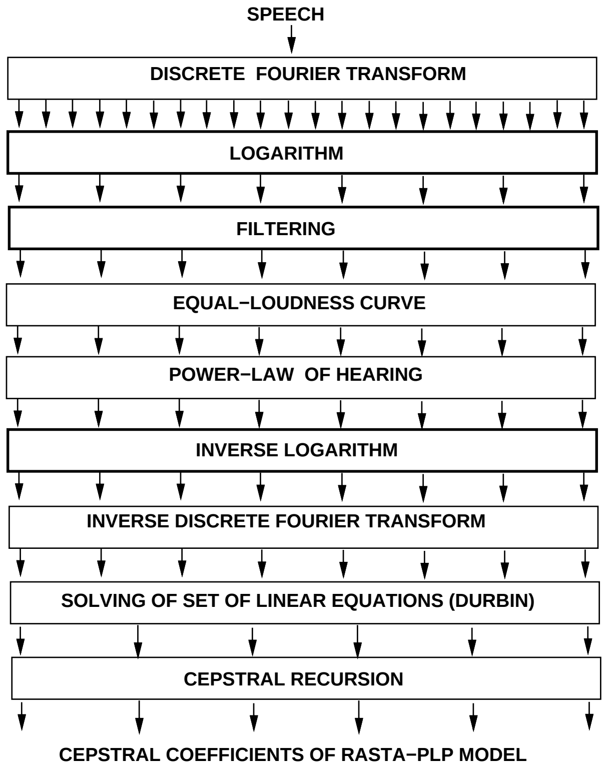
SPEECH
DISCRETE FOURIER TRANSFORM
LOGARITHM
FILTERING
EQUAL−LOUDNESS CURVE
POWER−LAW OF HEARING
INVERSE LOGARITHM
INVERSE DISCRETE FOURIER TRANSFORM
SOLVING OF SET OF LINEAR EQUATIONS (DURBIN)
CEPSTRAL RECURSION
CEPSTRAL COEFFICIENTS OF RASTA−PLP MODEL

Figure 1: RASTA-PLP Method

Linear distortions as caused e.g. by the telecommunication channel or by using a different microphone appear as an additive constant in the log spectrum. The high-pass portion of the equivalent band-pass filter is expected to alleviate the effect of the convolutional noise introduced in the channel. The low-pass filtering is expected to help in smoothing out some of fast frame-to-frame spectral changes present in the short-term spectral estimate due to analysis artifacts. In Eq. (1), the low cut-off frequency is 0.26 Hz. The filter slope declines 6dB/oct from 12.8 Hz with sharp zeros at 28.9 Hz and at $\approx$ 50 Hz. There is no special reason (except historical) for using the particular filter of Eq. (1). Also, the same filter need not be used for all frequency channels. Further, the filtering does not have to be band-pass or even linear.

The result is generally dependent on the starting point of analysis. In our applications we always start analysis well in the silent part which precedes speech.

The whole RASTA-PLP process is illustrated in Fig.1.

3 EXPERIMENTS WITH SMALL VOCABULARY ISOLATED TELEPHONE QUALITY SPEECH

This series of experiments were designed to evaluate the effect of varying telephone network environment. The training data were recorded at the Bellcore facility in Morristown, NJ, and represented channel conditions in the New Jersey area. An isolated-utterance continuous-density HMM recognizer was used in the experiment. A database was formed by manually segmenting digits from connected utterances recorded over dialed-up telephone lines. 155 male and female speakers were used for the training of the recognizer. 5th order autoregressive models were adopted for both the PLP and RASTA-PLP techniques in this experiment. Additional details of the experiment are given in [4].

Three experiments were carried out. In all experiments, the system was trained on the Bellcore training database.

In the first experiment, the test set was a subset of the Bellcore database. Thus, we assume that both the test set and the training set were recorded under similar channel conditions. Data from additional 56 male and female speakers, recorded at Bellcore, formed the test. The first column of Table I shows the percentage error rates on this test data. The RASTA-PLP performs about as well as the standard PLP technique.

In the second experiment, the Bellcore test data set was corrupted by a simulated convolutional noise (pre-emphasis by the first-order differentiation of the signal). The recognizer had been trained on the uncorrupted Bellcore training data. The results are tabulated in the second column of Table I. The standard PLP technique yielded almost an order of magnitude higher error rate than the error rate on the uncorrupted Bellcore data. The new approach can be seen to be far more robust to such simulated channel variation.

To extend the result to an experiment with realistic changes in channel conditions, digit strings spoken by four (2 male and 2 female) speakers were recorded over the local telephone lines in the U S WEST speech laboratory. The recognition results on this set are shown in the third column of Table I. As with the previous experiment, the conventional PLP technique yields a very high error rate. A similar test showed that a standard LPC-based system degraded even further, to a 60.7% error rate. The performance of RASTA-PLP degrades only slightly.

Analysis	Original Speech	Modified Speech	Different Environment
PLP	4.08%	31.35%	31.30%
RASTA-PLP	3.81%	5.00%	7.64%

Table I ISOLATED DIGIT ERROR RATES

4 EXPERIMENTS WITH LARGE VOCABULARY CONTINUOUS HIGH QUALITY SPEECH

We were curious whether our positive results with HMM-based ASR of telephone speech extend to a completely different ASR system and task. The standard large vocabulary continuous speech DARPA Resource Management database was chosen for this test. The recognizer used in the new series of experiments was a hybrid recognizer with a neural network trained on 4000 sentences to predict monophones for each frame, and then used in recognition to estimate likelihoods for a simple context-independent HMM system. 300 development test sentences from the October 1989 Resource Management speaker independent continuous speech recognition corpus were used as the test data. Since the DARPA database has 8 kHz bandwidth (twice the telephone speech bandwidth of the previous experiment), the autoregressive model in both PLP and RASTA-PLP analysis was increased from 5th to 8th order.

To simulate the effect of muffled speech that we had observed with a small obstacle between the microphone and the talker's mouth, a lowpass filter (a single complex pole pair, with a 3dB point at 2 kHz and a 20 dB loss at 8 kHz was applied to degrade the test data.

The word error results, shown in Table 2, indicate that the low-pass filtering significantly degrades the performance of the PLP-based recognizer. The RASTA processing in PLP had almost no effect on performance for the clean data, and kept the recognizer performance insensitive even to the severe low-pass filtering.

Informally we have observed that RASTA-PLP gives a substantial advantage in our live recognition experiments; while the conventional short-term spectrum based front-end is very sensitive to the choice of the microphone or even to the microphone position relative to the mouth, the RASTA-PLP makes the recognizer much more robust to such factors. Further, even the harmful effect of a constant additive noise background, often present in our live recordings, appears to be reduced.

Analysis	Original Speech	Modified Speech
PLP	17.9%	64.7%
RASTA-PLP	18.6%	19.2%

Table II CONTINUOUS SPEECH WORD ERROR RATES

5 DISCUSSION

A major current research concern is the significant degradation of high-performance laboratory systems when used in a real world. We believe that one of reasons for such a degradation is a highly variable frequency characteristics of the realistic recording and communication environments. Previous techniques for dealing with the problem of the convolutional noise introduced by such variable environment (see e.g. [6],[7]) appear to be useful for recognition applications that permit the explicit computation of a communication channel transfer functions. Such applications typically require a separate channel estimation phase. It appears that our simple RASTA-PLP technique is quite efficient in dealing with the convolutional noise. In addition, the RASTA-PLP computes all estimates on-line. That may prove advantageous for applications where the channel conditions are not known a priori or where the conditions might change unpredictably during the use of the recognizer.

Because we have been primarily concerned with convolutional noise in the communication channel, we conducted our corrections in the log spectral domain. RASTA technique could be also used in the magnitude or power spectral domains for additive noise reduction. However, care must be taken to ensure positivity of the enhanced power spectrum, as is also the case for traditional spectral subtraction techniques.

The study reported here made no use of other potential capabilities of the RASTA processing, particularly the ability to apply signal modifiers to the spectral temporal derivative domain. For instance, a threshold imposed on small temporal derivatives could provide a further nonlinear smoothing of the spectral estimates, and nonlinear amplitude modifications could enhance or suppress speech transitions.

Our current band-pass filter may not be optimal. Further, there is no fundamental reason to use the same filter for all spectral channels. Those issues are topics of our current research.

We also note that a German group of researchers, using a highpass filtering approach, primarily in the power spectral domain, has achieved encouraging results in suppressing the additive noise on a different set of speech recognition problems [8]. Their experience appears to confirm the effectiveness of the RASTA class of techniques.

6 SUMMARY

A new technique for estimating a robust time-varying spectrum, RASTA-PLP, based on the filtering of time trajectories of outputs from critical-band filters, has been described. A large test was conducted on a speaker-independent telephone digit recognition task using speech that had been corrupted with convolutional noise. Results from this test show an order-of-magnitude improvement in error rate over conventional spectral estimation techniques such as LPC or PLP. Results from similar tests with large vocabulary continuous speech recognition show that the improvement is consistent across different databases and different recognition techniques.

7 ACKNOWLEDGEMENT

Thanks to Chuck Wooters and Steve Renals for assistance with the large vocabulary experiments.

References

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

[2] Q. Summerfield and P. Assmann: Auditory enhancement and the perception of concurrent vowels, Perception & Psychophysics, 1989, 45 (6), pp. 529-536.

[3] H. Hermansky: "Auditory model for parametrization of speech in real-life environment based on re-integration of temporal derivative of auditory spectrum," U S WEST Advanced Technologies Research Report, File Folder ST 04-01, October 1990.

[4] H. Hermansky, N. Morgan, A. Bayya, P. Kohn: "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," Proc. of Eurospeech '91, pp. 1367-1371, Genova, Italy, 1991.

[5] S. Furui: "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics," Procs. IEEE Intl. Conf. on Acoustic, Speech & Signal Processing, pp. 1991-1994, Tokyo, Japan 1986

[6] A. Accero and R. M. Stern : "Towards Environment-Independent Spoken Language Systems," Proc. Speech and Natural Language Workshop, DARPA, June 1990, pp. 157-162

[7] E. Errel and M. Weintraub: "Recognition of Noisy Speech: Using Minimum-Mean Log-Spectral Distance Estimation," Proc. Speech and Natural Language Workshop, DARPA, June 1990, pp. 341-345

[8] H. Hirsch, P. Meyer, and H. Ruehl: "Improved speech recognition using high-pass filtering of subband envelopes," Proc. of Eurospeech '91, pp. 413-416, Genova, Italy, 1991.