RASTA Processing of Speech

Hynek Hermansky, *Member, IEEE*, and Nelson Morgan, *Senior Member, IEEE*

***Abstract*— Performance of even the best current stochastic recognizers severely degrades in an unexpected communications environment. In some cases, the environmental effect can be modeled by a set of simple transformations and, in particular, by convolution with an environmental impulse response and the addition of some environmental noise. Often, the temporal properties of these environmental effects are quite different from the temporal properties of speech. We have been experimenting with filtering approaches that attempt to exploit these differences to produce robust representations for speech recognition and enhancement and have called this class of representations relative spectra (RASTA). In this paper, we review the theoretical and experimental foundations of the method, discuss the relationship with human auditory perception, and extend the original method to combinations of additive noise and convolutional noise. We discuss the relationship between RASTA features and the nature of the recognition models that are required and the relationship of these features to delta features and to cepstral mean subtraction. Finally, we show an application of the RASTA technique to speech enhancement.**

I. Introduction

SPEECH carries information from many sources. Not all information sources are relevant for a given task. Conventional short-term, spectrum-based speech analysis techniques blindly and faithfully represent most information-carrying components in the signal. Then, data-intensive stochastic techniques are commonly applied to reduce the effects of the irrelevant information. However, the sources of nonspeech components are often deterministic, their effect on the speech signal is predictable, and the application of stochastic techniques appears wasteful; the reduction of irrelevant information in the speech analysis module of the recognizer can increase the efficacy of finite amounts of training data.

Take as an example a change in the frequency characteristics of a communications channel caused, e.g., by switching to a new microphone. The linear microphone characteristics show as a convolutional component in the signal and therefore as an additive component in the logarithmic spectrum of speech. Any metric based on a short-term logarithmic spectrum (or cepstrum) of speech will reflect this microphone change. When the speech from this new microphone is not represented in the training data, the performance of a recognizer that employs the short-term spectral representation of speech will be typically severely impaired.

However, the frequency characteristic of a communications channel is often fixed or only slowly varying in time. If our speech representation was invariant to slow changes in the logarithmic spectrum of speech, the problem would not have arisen. The blind deconvolution of signals [23] is one way of addressing the problem; the long-term average is subtracted from the logarithmic spectrum of speech to make it insensitive to changes in the long-term average. A variant of this approach was used for speech recognition by Ney [20] and more recently under the name of cepstral mean subtraction [22]. This method does require a long-term average, which may be difficult to obtain in real-time implementations.

In another case, assume that speech is corrupted by additive noise. If the noise is uncorrelated with the original speech, the noise component is additive in the power spectrum of the signal. Again, most conventional speech representations will be affected by such additive noise. If the noise is changing slower than speech, one accepted way of dealing with the noise is spectral subtraction [1], in which the estimate of noise power spectrum (obtained in nonspeech intervals of the signal) is subtracted from the power spectrum of the signal. At least two problems arise: 1) A speech detector is required to determine intervals from which a reliable noise estimate can be obtained. 2) The subtraction process can result in negative power spectral values. This is typically handled in some *ad hoc* manner (e.g., by setting the negative values to zero or a small positive constant).

In automatic recognition of speech (ASR), the task is to decode the linguistic message in speech. This linguistic message is coded into movements of the vocal tract. The speech signal reflects these movements. The rate of change of nonlinguistic components in speech often lies outside the typical rate of change of the vocal tract shape. The relative spectral (RASTA) technique presented in this paper takes advantage of this fact. It suppresses the spectral components that change more slowly or quickly than the typical range of change of speech. We demonstrate that RASTA processing improves the performance of a recognizer in presence of convolutional and additive noise. Finally, we discuss an application of RASTA processing for enhancement of noisy speech.

II. Human Auditory Perception

The fact that human perception tends to react to the *relative* value of an input (rather than to its absolute values) is quite obvious in vision (how else you could see the *black*-and-white movie on a *white* screen?), but the literature on perception of very slowly varying auditory stimuli seems to be rather

Manuscript received November 16, 1993; revised April 13, 1994. This work was supported by the International Computer Science Institute, by U S WEST Technologies, and by the Joint Services Electronics Program Contract no. F49620-93-C-0014 to UC Berkeley.

H. Hermansky is with the Oregon Graduate Institute, Portland, OR 97291 USA.

N. Morgan is with the International Computer Science Institute, Berkeley, CA 94704 USA.

IEEE Log Number 9403969.

scarce. Some circumstantial evidence indicates that there is a preference for sounds with a certain rate of change. Green [5] cites early experiments of Riesz [21] that were later confirmed by Zwicker [26] and Green [5], which indicate a greater sensitivity of human hearing to modulation frequencies around 4 Hz than to lower (or higher) modulation frequencies.

More convincing evidence comes from some speech experiments. Some years ago, a simple but striking[1] experiment was carried out [2]: A whole spoken sentence was processed by a filter that approximated the inverse of the short-term spectral envelope of the center of one of the vowels in the sentence. Thus, the spectrum of the given vowel became roughly white. In spite of this, the sentence remained perfectly intelligible, and *the given vowel was still perceived*, in spite of its lack of any formant structure. More formal experiments supporting this notion were done by Summerfield and his colleagues [24] who showed that a perception of speech-like sounds is dependent on the preceding sound, namely, that it depends on the spectral *difference* between the current sound and the preceding sound.

III. Principle of the RASTA Method

The relative insensitivity of human hearing to slowly varying stimuli may partially explain why human listeners do not seem to pay much attention to a slow change in the frequency characteristics of the communication environment or why steady background noise does not severely impair human speech communication.

However, even when the experimental evidence from human perception may give us only limited support, the suppression of slowly varying components in the speech signal makes good engineering sense. Thus, to make speech analysis less sensitive to the slowly changing or steady-state factors in speech, we have replaced a conventional critical-band short-term spectrum in PLP speech analysis [6] with a spectral estimate in which each frequency channel is band-pass filtered by a filter with a 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 [7], [11].

The steps of RASTA-PLP are as follows (see [6] for a comparison to the conventional PLP method). For each analysis frame, do the following:

1) Compute the critical-band power spectrum (as in PLP).
2) Transform spectral amplitude through a compressing static nonlinear transformation.
3) Filter the time trajectory of each transformed spectral component.
4) Transform the filtered speech representation through expanding static nonlinear transformation.
5) As in conventional PLP, multiply by the equal loudness curve and raise to the power 0.33 to simulate the power law of hearing.
6) Compute an all-pole model of the resulting spectrum, following the conventional PLP technique.

The key idea here is to suppress constant factors in each spectral component of the short-term auditory-like spectrum prior to the estimation of the all-pole model.

The most important research issues are in steps 2) and 3), i.e.:

- in which domain is the filtering done
- which filter to use.

As for the filter used, we started with an IIR filter with the transfer function[2]

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

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

The high-pass portion of the equivalent band-pass filter is expected to alleviate the effect of convolutional noise introduced in the channel. The low-pass filtering helps to smooth some of the fast frame-to-frame spectral changes present in the short-term spectral estimate due to analysis artifacts. In (1), the low cut-off frequency is 0.26 Hz. The filter slope declines 6 dB/oct from 12.8 Hz with sharp zeros at 28.9 and at 50 Hz.

Note that the RASTA filter has a rather long time constant for the integration (about 500 ms for the filter (1) and 160 ms for more recent implementations). It means that the current analysis result depends on its history (i.e., on previous outputs stored in the memory of the recursive RASTA filter).[3] For example, the analysis result is dependent on where in the signal the analysis starts, i.e., how is the RASTA filter initialized. In our experiments, we typically address this issue by starting analysis as far as possible in silence preceding the speech.

The whole RASTA process is illustrated in Fig. 1.

IV. Experiments on Data from Different Recording Environments

A. Logarithmic RASTA

In the first set of experiments, we were concerned about the effect of convolutive distortions as caused, e.g., by variable frequency characteristics of different communication channels or by using different microphones. Such distortions should appear as an additive constant in the logarithmic spectrum of speech. Thus, we have used logarithmic amplitude transformation as a compressing static nonlinearity in Step 2 of the RASTA-PLP method. The expanding static nonlinearity (Step 4 of the method) was an antilogarithmic (exponential) transformation.

[1] Unfortunately unpublished, to the best of our knowledge.

[2] The pole was modified for later experiments; see Section IV-D.

[3] In one of our experimental runs, we have accidentally included a part of a file header in the analyzed signal. Although this caused only a minor problem for the conventional frame-by-frame PLP analysis (only the first frame was affected, and it was cut off prior to pattern matching), it had a major effect on the RASTA analysis since the effect of the first corrupted frame propagated (due to the memory of the RASTA filter) over a significant part of the analysis output.

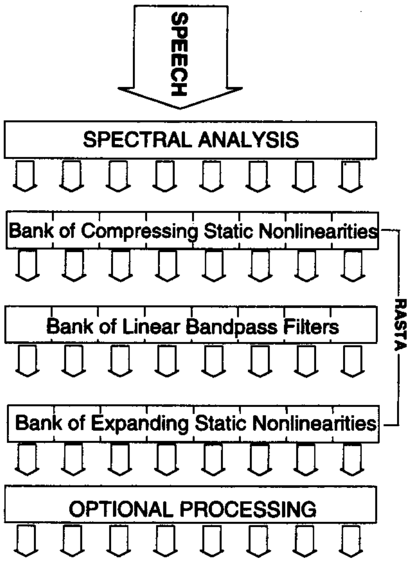

SPEECH
SPECTRAL ANALYSIS
Bank of Compressing Static Nonlinearities
Bank of Linear Bandpass Filters
RASTA
Bank of Expanding Static Nonlinearities
OPTIONAL PROCESSING

Fig. 1. Block diagram of RASTA speech processing technique.

B. Isolated Digit Recognition Experiments

A database was derived from connected digits recorded over dialed-up telephone lines. One hundred and fifty five male and female speakers were used for the recognizer, and data from an additional 56 male and female speakers formed the test. The 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. Additional details of the experiment can be found in the Appendix and are given in [9]. Three experiments were carried out. In all experiments, the system was trained on the training part of the Bellcore 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. The first column of Table I shows the percentage error rates on this test data. 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 (preemphasis by the first-order differentiation of the signal). The recognizer that had been trained on the uncorrupted data was still used. The recognition tests were run on this data set, using the models obtained from Bellcore 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 original Bellcore data. RASTA-PLP 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 (two male and two female) speakers were recorded over the local telephone lines in the US WEST speech laboratory in Colorado. The recognition errors on this set are shown in the third column of Table I.

TABLE I
ISOLATED DIGIT ERROR RATES

method	same environment	controlled modification	different environment
PLP	4.08%	31.35%	31.10%
RASTA-PLP	3.81%	5.0%	7.64%

TABLE II
SPEAKER INDEPENDENT CONTINUOUS SPEECH WORD ERROR RATES

method	unfiltered	filtered
PLP	17.9%	67.5%
RASTA-PLP	18.6%	33.0%

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 moderately.

C. Large Vocabulary Continuous Speech Experiments

As a followup to the work reported above, we applied a simple low-pass filter (a single complex pole pair with a 3-dB point at 2 kHz) to 300 development test sentences from the October 1989 Resource Management speaker independent continuous speech recognition corpus. This filter was chosen to implement an approximation to the effect of muffled speech that we had observed with a small obstacle between the microphone and the talker's mouth. Both eighth-order PLP and eighth-order RASTA-PLP were computed on these data as well as on unfiltered versions of the standard 3990-sentence training set. The recognizer used was a hybrid recognizer with a neural network trained on the 3990 sentences to predict monophones for each frame and then used in recognition to estimate likelihoods for a simple context-independent HMM system.

The word error results, which are shown in Table II, show that RASTA processing causes only slight degradation of performance for the clean data but cuts the error in half for the filtered case.

Informally, we have observed that RASTA-PLP gives a substantial advantage in our on-line 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, RASTA-PLP makes the recognizer much more robust to such factors.

D. Some Optimizations of the RASTA Filter

In the experiments reported above, we used an AR integration constant (real pole at $z = 0.98$) corresponding to a time constant of 500 ms. Later runs with these same data showed that a smaller pole value ($z = 0.94$) corresponding to a 160-ms time constant appeared to be optimal.

Results of the optimization experiment are shown in Fig. 2. The recognition vocabulary was 11 isolated digits plus

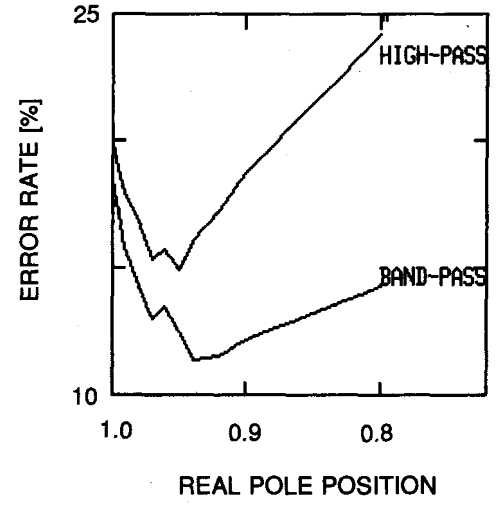


Fig. 2. Results of pilot experiments with telephone-quality digits.

two control words ("yes" and "no") recorded by 30 speakers over dialed-up telephone lines. Digits were hand end-pointed. The recognizer was a DTW-based multitemplate recognizer. Twenty seven speakers out of 30 were used for training of the recognizer in a jack-knife experimental design, thus yielding 52780 recognitions trials per experimental point. (See the Appendix for further a description of this experiment).

To introduce convolutive distortion, test data were filtered by a linear filter simulating the inverse spectral envelope of a sustained vowel /a/. To investigate the effect of low-pass filtering by the MA part of the RASTA filter (1), two different MA polynomials were used, namely

$$M_1(z) = z * (0.5 - 0.5z^{-1})$$

and our original

$$M(z) = z^4 * (0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4})$$

for the two-point MA filter (which is denoted in the figure as "high-pass") and the five-point MA filter (which is denoted in the figure as "band-pass"), respectively. The experiment described above indicates that the most important feature for alleviating the harmful effects of variable environment transfer function seems to be the sharp spectral zero at zero frequency. However, using the five-point MA filter seems to yield a consistent advantage. The position of the spectral pole, which determines the high-pass cut-off frequency (around 0.9 Hz for the pole at $z = 0.94$) exhibits a broad optimum.

Fig. 3 shows the frequency response of our current RASTA filter. This frequency response compares well with the outcome of Green's experiments with detectability of FM signals (see p. 917 of [17]). Fig. 4 shows its impulse response. The filter step response has a single time constant around 160 ms, which compares well with the "nervous integration time" of Liang and Chistovich (see p. 903 of [17]) and is also roughly in the range of estimates of an integration time constant observed in some auditory enhancement experiments [24].

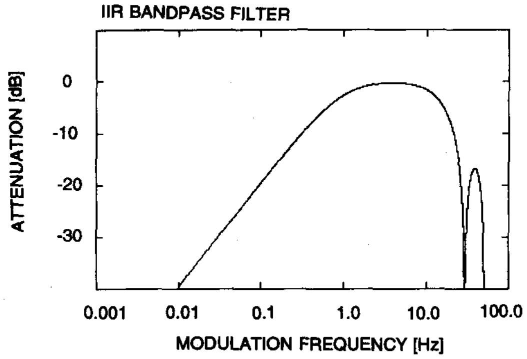


Fig. 3. Frequency response of RASTA band-pass filter.

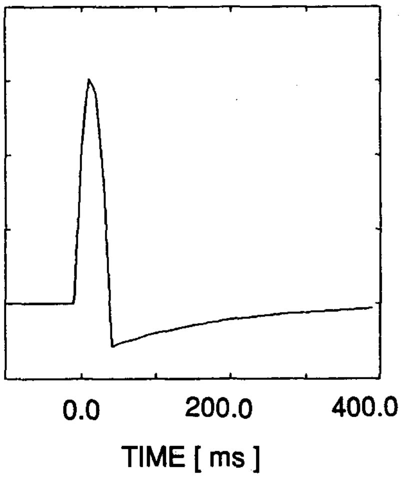


Fig. 4. Impulse responses of RASTA band-pass filter.

V. Experiments with Data Containing both Additive and Convolutional Noise

A. Recognition in Additive Noise

As noted earlier, an operating acoustic environment for a practical recognizer (room, microphone, telecommunication channel, ...) not only has variable frequency characteristics but may also be noisy. Table III shows the results of an isolated word recognition experiment in which the recognizer was operating on data that 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 III shows the recognizer accuracy when training is on data with an environment that is identical to that for the test data, that is, the recognizer was always trained on the data that 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.

TABLE III
ISOLATED DIGIT ERROR RATES USING DTW-BASED SYSTEM. IN THE FIRST ROW, TRAINING AND TESTING WERE BASED ON THE SAME NOISE CONDITION (AS A BEST-CASE COMPARISON). FOR ALL OTHER CASES, THE RECOGNIZER WAS TRAINED ON CLEAN SPEECH. NOISE INDICATES TEST DATA WITH SNR=10 dB, AND NOISE-FILTERED INDICATES TEST DATA WITH THE SNR=10 dB AND WITH AN ADDITIONAL CONVOLUTIONAL LINEAR DISTORTION INTRODUCED BY FILTERING.

	clean	noise	clean-filtered	noise-filtered
PLP same environment	12.0	17.2	14.0	21.5
PLP	12.0	43.4	39.7	67.5
RASTA	12.2	42.1	19.9	49.2

Unfortunately, the noise characteristics are 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 III, 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 reduce the need for extensive training of the recognizer in different environments. In this section, we show that our new method is comparable to training on noisy data.

B. Lin-Log RASTA

When 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 (which are convolutive in the time domain and, therefore, additive in the log spectral or cepstral 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 Table III, the original RASTA processing on the logarithmic spectrum or cepstrum is not particularly appropriate for speech with significant additive noise.

Hirsch *et al.* [13], using a high-pass filtering approach primarily in the power spectral domain, achieved encouraging results in suppressing additive noise on a different set of speech recognition problems. Their experience appeared to confirm the effectiveness of the RASTA class of techniques. Therefore, we decided to study RASTA processing in an alternative spectral domain, which is linear-like for small spectral values and logarithmic-like for large spectral values.

In [18], we have proposed as a substitute for the logarithmic transform in Step 2 of RASTA processing the function

$$y = \ln(1 + Jx) \tag{2}$$

where J is a signal-dependent positive constant. The amplitude-warping transform is linear-like for $J \ll 1$ and logarithmic-like for $J \gg 1$. The exact inverse of (2)

$$x = \frac{e^y - 1}{J} \tag{3}$$

(where e is the base of the natural logarithm) is not guaranteed to be positive for all y and, as in conventional spectral subtraction, would require some *ad hoc* procedure to ensure the positivity of the processed power spectrum. To avoid this, we use an approximate inverse transform as an expanding static nonlinearity in the Step 4 of RASTA processing.

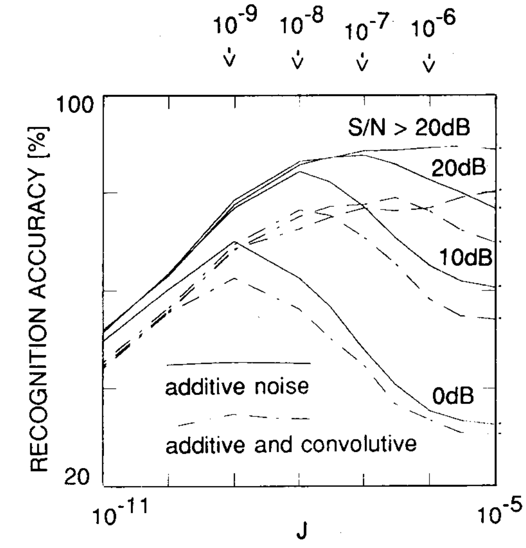


Fig. 5. Digit recognition.

$$x = \frac{e^y}{J}. \tag{4}$$

This inverse is equivalent to the sum of the exact inverse and additive constant $1/J$. It is therefore less accurate for small spectral values than for the larger ones.[4]

C. Isolated Digit Experiment

We repeated the earlier isolated word recognition experiments using the nonlinearities (2) and (4). The results shown in Fig. 5 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 the PLP or RASTA-PLP result.

[4] We have observed similar results with a piecewise transform $y = \frac{Jx}{e}$ for $Jx < e$, $y = \log(Jx)$ for $Jx > e$, using the approximate inverse $x = \frac{e^y}{J}$. This inverse is exact for $Jx > e$. Results using either transformation are generally very similar and, throughout the paper, we only give results using the first one (see (3) and (5)). This comparison suggests 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.

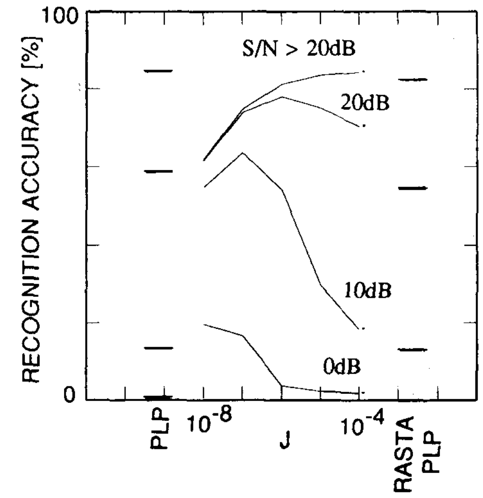


Fig. 6. Large vocabulary continuous speech recognition.

D. Large Vocabulary Continuous Speech Experiment

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 many 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 a further description). Fig. 6 shows the results for additive noise. About 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 SNR. Smaller values of J are preferred for noisy speech.

1) Rationale for the Optimal J: Results shown in Figs. 5 and 6 indicate that there is an value optimal of J for each particular SNR case in the test data. Fig. 7 shows histograms of logarithmic auditory-like spectral energies x for all four SNR's that were used. Spectral values for which $J_{\text{optimal}}x = e$ for all four investigated SNR's are indicated in the figure by arrows. Supporting [25], the histograms are multimodal. 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.[5]

E. 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.

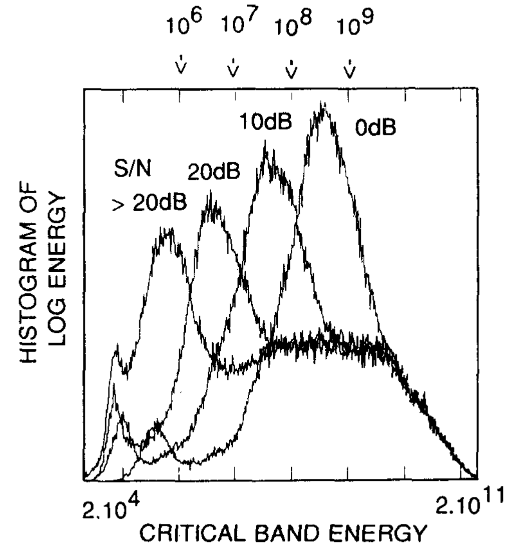


Fig. 7. Histograms of critical-band energies.

TABLE IV
ISOLATED DIGIT ERROR RATES USING DTW-BASED MULTITEMPLATE SYSTEM. THE RECOGNIZER WAS TRAINED ON CLEAN SPEECH. NOISE INDICATES TEST DATA WITH SNR=10 dB, AND NOISE-FILTERED INDICATES TEST DATA WITH THE SNR=10 dB AND WITH AN ADDITIONAL CONVOLUTIONAL LINEAR DISTORTION INTRODUCED BY FILTERING.

	clean	noise	clean-filtered	noise-filtered
Lin-log RASTA	11.4	15.1	20.9	25.7

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. 6, the J depends on the level of noise in the signal that can vary during 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. It would be impractical to reanalyze the training data and retrain the recognizer every time the noise level changes. Therefore, we have followed a different strategy in this work. We have 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.

$$J = \frac{1.0}{CE_{\text{noise}}}$$

In the training of the recognizer, we have used four different sets of templates, where each set is trained with an order-of-magnitude different C_{train}, namely

$$C_{\text{train}} = 3 \times 10^3, 3 \times 10^2, 3 \times 10^1, \text{ and } 3.$$

Thus, compared with the fixed J recognizer (which could be used if there was no change in the S/N conditions during the operation), our recognizer for the operation under variable S/N conditions requires four times as many templates.

[5] The RASTA function of (2) may also be written as

$$y = \ln(J) + \ln\left(\frac{1}{J} + x\right). \tag{5}$$

The first term is constant and is filtered out by the band-pass filter. Therefore, lin-log RASTA may be also viewed as a noise-masking technique in which a fixed amount of additive noise (inversely proportional to an overall noise estimate) is added to every spectral component prior to the RASTA processing in the logarithmic domain.

TABLE V
ISOLATED DIGIT ERROR RATES USING HTK-BASED GAUSSIAN MIXTURE SYSTEM. IN THE FIRST ROW, TRAINING AND TESTING WERE BASED ON THE SAME NOISE CONDITION (AS A BEST-CASE COMPARISON). FOR ALL OTHER CASES, THE RECOGNIZER WAS TRAINED ON CLEAN SPEECH. NOISE INDICATES TEST DATA WITH SNR=10 dB, AND NOISE-FILTERED INDICATES TEST DATA WITH THE SNR=10 dB AND WITH AN ADDITIONAL CONVOLUTIONAL LINEAR DISTORTION INTRODUCED BY FILTERING.

	clean	noise	clean-filtered	noise-filtered
PLP same environment	5.0	10.0	7.2	10.1
PLP	5.0	37.0	24.9	50.4
RASTA	3.3	50.0	3.6	40.4
PLP cepstral mean removal	4.3	42.0	5.0	46.7
Lin-log RASTA	3.7	13.7	5.6	17.1

During the operation of the recognizer, C was fixed at

$$C_{\text{test}} = 3.$$

Results from such an automatically adaptive system are shown in Table IV and indicate a substantial improvement compared with the performance of PLP or RASTA shown in Table III.

F. Compensation for the Variable Static Nonlinearity

As shown in the previous Sections D and E, different SNR's require different static nonlinearities, which in turn result in different all-pole models. To compensate for this deterministic variability in the analysis result, we applied a linear mapping to the RASTA processed auditory-like spectrum based on the multiple regression model

$$\hat{Y}_i = c_{i0}^{(\mathrm{S/N})} + \sum_{k=1}^{N} c_{ik}^{(\mathrm{S/N})} X_k^{(\mathrm{S/N})} \tag{6}$$

where

$\hat{Y}_i$ multiple-regression estimate of the ith element of a RASTA-filtered auditory-like spectrum that would use the static nonlinearity optimal for the noise level in the training data (typically for the clean speech)

$X_k^{(\mathrm{S/N})}$ kth element of the true RASTA-filtered auditory-like spectrum using the static nonlinearity optimal for the given SNR in the test utterance (quantized to seven levels of S/N, i.e., >25 dB, 25 dB, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB)

$c_{ik}^{(\mathrm{S/N})}$ multiple regression coefficients for the given SNR

N number of elements of the auditory-like spectrum.

As in the work described in the preceding section (where we have used multiple templates derived for different values of J), we have quantized J and thus derived three different mapping matrices (the mapping for the low-noise speech is unity) using speech from 10 speakers not used for the training or test sets. In addition, in contrast to the previous experiments, a larger set of 200 speakers was used. Additionally, the HMM Tool Kit (HTK) from Cambridge University with four Gaussian mixtures per state was used instead of a DTW recognizer. Finally, a noise estimate was performed on-line by a histogram-based technique developed by Hirsch [14] that does not require explicit speech/nonspeech detection. Details of the experiment are described in the Appendix.

Results shown in Table V compare this technique trained on clean speech (last row of the table) with the following:

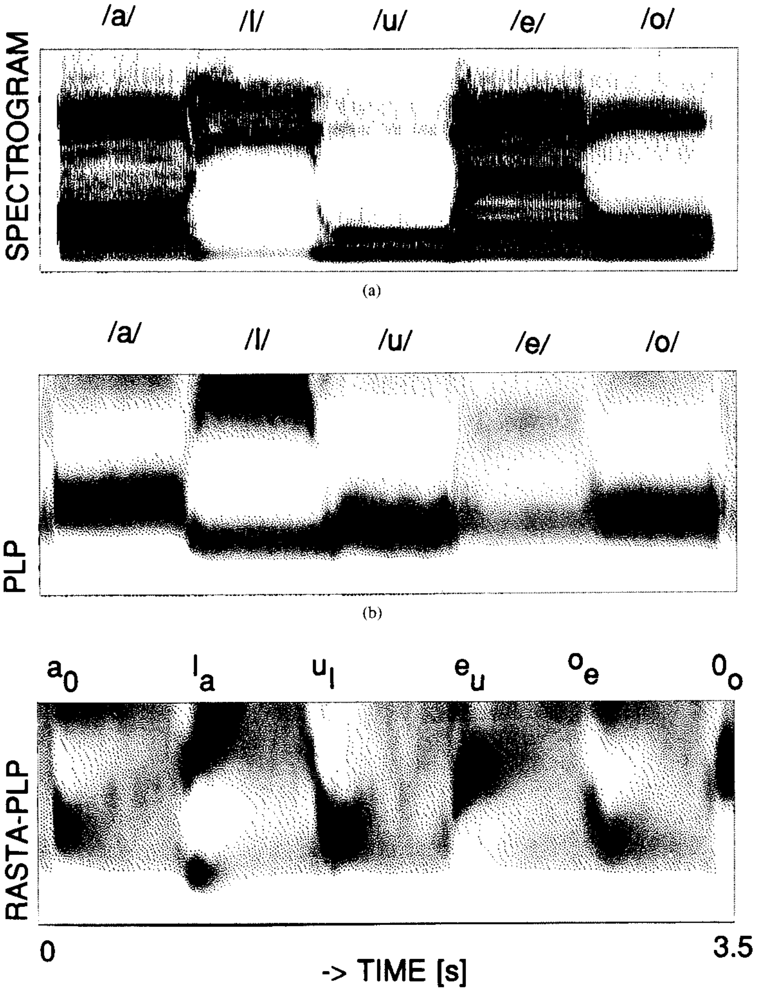
1) conventional PLP trained on the speech from identical environment
2) PLP trained on the clean speech
3) logarithmic RASTA trained on the clean speech
4) PLP with cepstral mean removal [22].

Both logarithmic RASTA and cepstral mean removal help for convolutional noise. However, PLP, logarithmic RASTA, and cepstral mean removal all degrade severely in additive noise. Lin-log RASTA with a linear mapping yielded good robustness over both convolutional and additive noise.

While cepstral mean subtraction performed equivalently or better for a purely convolutional noise, it was not as effective as the lin-log RASTA approach when additive noise was present.

VI. CONSEQUENCES OF RASTA PROCESSING

As already briefly discussed in the Section III, one of the most important differences between conventional frame-by-frame speech analysis and RASTA-based techniques is that the RASTA result depends on its history. Employing some larger part of the signal against which the current analysis frame is compared is a strategy that is also used in in other channel equalization techniques such as the blind deconvolution (cepstral mean removal), and it is the very reason why all such techniques can differentiate between relative steady disturbances and the varying speech components of the signal. However, while cepstral mean subtraction typically compares the current analysis frame against the average of the whole utterance, RASTA uses a relatively short history of the signal on the order of several hundred milliseconds (as implied by the time constant of the RASTA filter). Such a short history employed in RASTA effectively enhances transitions between different speech segments and makes the result dependent on the previous short segment of speech such as phoneme or syllable. This property, which is illustrated in Fig. 8, which shows spectrograms of five sustained Czech vowels produced by a) conventional FFT, b) PLP, and c) RASTA-PLP, makes RASTA less suitable for most current phoneme-based



(a)

(b)

(c)

Fig. 8. Spectrograms of five sustained Czech vowels derived by (a) conventional FFT, (b) PLP, and (c) RASTA-PLP, respectively.

recognizers that assume steady or piecewise steady phoneme-sized models.

In a recent experiment, we observed how the choice of recognition model can strongly change the apparent effect of RASTA processing. We compared the effects of model complexity on PLP performance with and without RASTA processing. Specifically, we tested the difference between RASTA PLP and PLP results for a complex (clustered triphone, five mixtures) and a simple (single mixture monophone) model. HTK software simulation of a continuous density HMM recognizer was used on the Credit Card portion of the Switchboard corpus (see [27] and Appendix for more details of the experiment).

Since cepstral mean removal used in both cases (where the mean was computed over long utterances) was apparently sufficient to remove the effect of the variable communication channel in the data, and the additive noise was not a significant problem in this task, the RASTA processing did not yield any advantage. However, the experiment is interesting since it shows how an insufficient model (single-mixture monophone) does not account for the effect of the left context enhanced by the RASTA processing.

VII. Relation of RASTA Processing to Some Other Channel Equalizing Techniques

A. Comparison of RASTA Processing to Delta Features

Furui [4] correctly observed that convolutional distortions will be also alleviated in the so-called "delta" cepstral features of speech. As a matter of fact, it was this property of delta

TABLE VI
WORD ERROR RATES FOR CONTINUOUS SPONTANEOUS SPEECH RECOGNITION ON CREDIT CARD PORTION OF THE SWITCHBOARD DATABASE, 1102 TRAINING UTTERANCES, 258 TEST UTTERANCES, WORD-PAIR GRAMMAR DERIVED FROM THE TEST SET.

Model	PLP	RASTA-PLP
1-mixture monophone	64.2%	76.2%
5-mixture clustered triphone	41.3%	41.9%

features that originally motivated us to develop RASTA processing, and this influence is apparent in the numerator of our current RASTA filter-taken by itself: It is a transfer function of the delta feature calculation. As Furui also observed, the delta cepstral features do not perform too well by themselves. Therefore, they are typically appended to the vector of static cepstral features, which once again makes the representation vulnerable to convolutional distortions.

The RASTA method differs from the delta feature calculation in using a filter with a broader pass-band. In addition, the general form of RASTA processing does filtering between two static nonlinearities that are not necessarily the inverse of one another. The delta features can be viewed as a special case of temporal RASTA processing in which the RASTA filter is a five-point FIR filter applied to temporal trajectories of cepstral coefficients (linear transformation of logarithmic spectral domain) and in which the second static nonlinearity is absent.

Comparing Fig. 9, which shows frequency characteristics of several FIR filters that have been used in the past for the computation of delta features, with Fig. 3 (showing frequency characteristic of the RASTA IIR filter) shows that the band-pass RASTA IIR filter has a fairly flat frequency response within the 1 to 10 Hz frequency range, thus passing relatively undisturbed those components of the signal that we postulate to be the most relevant for carrying the linguistic information in the speech signal. On the other hand, the delta feature filters have rather selective frequency responses, emphasizing a small range of modulation frequencies and attenuating the rest, therefore modifying the relative importance of various linguistically relevant components in the speech signal. This is true for all delta-feature FIR filters. As seen in the figure, the delta-feature filter with a time constant of 170 ms (comparable with the time constant of our RASTA filter) still has a selective frequency response, in this case, with a maximum at about 2 Hz instead of 10 Hz for the conventional 50-ms FIR delta-feature filter.

B. Comparison of RASTA Processing to Cepstral Mean Subtraction.

As noted earlier, one particular form of blind deconvolution of speech done by subtraction of the mean cepstral vector (which is typically computed over the length of the current utterance) can alleviate the effect of a variable communication environment. This is currently often used for large vocabulary continuous speech ASR [22] and can be also viewed as a particular form of noncausal FIR filter employing a variable length future and past. Frequency responses of several filters implied by cepstral mean subtraction[6] over sentences of different lengths are shown in Fig. 10.

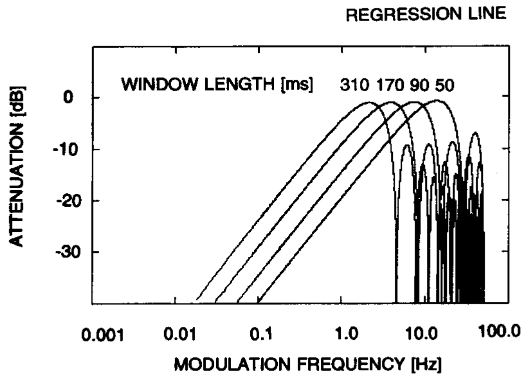


Fig. 9. Frequency response of RASTA filter compared with frequency responses of several forms of delta feature computation.

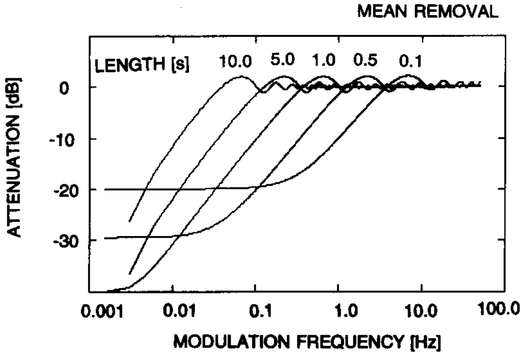


Fig. 10. Frequency responses implied in several forms of cepstral mean subtraction.

Compared with the typical RASTA filter (Fig. 3), the cut-off frequencies of all but the shortest cepstral mean subtraction filters are lower, eliminating only the fixed bias in the cepstral domain. Thus, the main difference between the cepstral mean subtraction and the RASTA processing in the log spectral (cepstral) domain is that the cepstral mean subtraction merely removes the dc component of the short-term log spectrum, whereas the RASTA processing influences the speech spectrum in a more complex way, making the current output dependent on its past and enhancing the spectral transitions, as discussed in the previous section.

As with cepstral mean subtraction, the RASTA technique (when applied in the log spectral domain) attempts to make the long-term average log spectrum identically zero. On a short-term basis, it combines a weighted average of all past log spectra with the current logarithmic spectrum, with the weighting determined by the impulse response of the RASTA filter (see Fig. 4). Fig. 4 illustrates that RASTA processing av-

[6] That is, we show the frequency response of the linear-time-invariant flat-weighted moving average with the same time window as the corresponding mean computation.

erages four log spectra around the current frame and subtracts from this average an exponentially weighted geometric mean of all past analysis frames. It is this logarithmic subtraction of a weighted average of past spectral values that accounts for RASTA environment equalization. Since the time constant of the exponential averaging window is about 160 ms, the running average being subtracted can follow slow changes in the communication environment.

In the course of incremental improvements of the ASR system, one typically tends to shy away from drastic changes of the recognition paradigm. Therefore, many may favor cepstral mean subtraction for dealing with convolutional noise, in spite of its inherent problems with delay when dealing with real continuous input. However, we observed, in our experiments, that cepstral mean subtraction does not appear to handle degradation due to additive noise. We have also observed that in some cases, the combination of cepstral mean subtraction and RASTA can give better results than either of the techniques alone [10].

VIII. Enhancement of Noisy Speech

One of the accepted conventional techniques for noise suppression is spectral subtraction, in which the noise power spectrum is estimated in intervals between speech and subtracted from a power spectrum of the signal. In some implementations, a magnitude spectral domain is preferred. The enhanced signal is then reconstructed by an overlap-add inverse Fourier transform using the modified magnitude and the original noisy phase of the signal spectrum. This technique is critically dependent on a reliable detection of nonspeech intervals of the signal. Typically, in every detected nonspeech interval, the noise estimate is updated with some time constant. This time constant needs to be short enough to allow for slow changes in noise but also needs to be long enough to compensate for possible occasional errors in the noise estimate.

A. RASTA Approach to Noise Suppression

Explicit estimation of the noise spectrum in nonspeech intervals is cumbersome, error prone, and may not be necessary. In principle, suppressing all slowly varying components in the magnitude spectrum may not significantly harm information bearing components, but it may reduce some slowly varying noise in the observed signal. By the same rationale, a suppression of rapidly varying spectral components may reduce some impulsive noise.

After some experimentation, we band-pass filtered the cubic root of power spectral magnitude using a fifth-order elliptic IIR band-pass filter with cut-off frequencies at 1.0 and 15.0 Hz (thanks to J. Allen of AT&T Bell Labs for the filter).

The enhanced speech was reconstructed from the half-wave rectified filtered power spectrum (its amplitude inverse was exponentiated to the third power to compensate for the cubic root compression prior to the filtering) and the original noisy phase and was delayed by one frame. This one-frame delay was determined empirically and appeared to function as a gross compensation for a phase delay introduced by the RASTA filter. The standard overlap-add technique [16] employing a 31.25-ms Hamming window with a 7.8125-ms analysis step was used for the analysis and resynthesis.

Noisy speech was obtained from a voice-mail message recorded over the cellular public network from a moving automobile. Impulsive noise was artificially introduced into this recording by randomly substituting samples with fixed amplitude for some of the original speech samples.

Informal listening revealed a significant reduction in both the steady background car noise and the impulsive noise. Some colored musical noise was perceived. Another disturbing artifact of the processing is a variable phase-shift-like distortion of the processed speech. Some weaker consonants, which were originally masked by the noise, still appear to be lost. As with conventional spectral subtraction, the processing does not seem to improve speech intelligibility (although at this time we have run no formal tests). Compared with the original noisy speech, however, the processed speech does appear to be more prominent above the background.

We have also explored a processing approach that has more moderate effects. We have tried a less aggressive RASTA filtering by mixing the original and filtered magnitude spectrum prior to the one-way rectification. This appears to be a promising technique for some applications that demand moderate noise suppression while preserving most of the original speech quality.

The RASTA enhancement processing described above is just a first cut at the problem. For this pilot work, we ran no formal perceptual experiments, nor did we explore any significant corpus of noisy data. Consequently, we do not claim that the processing parameters described above are optimal. However, the form of RASTA filter used appears to have a significant influence on the audible results. Based on the outcome of this small exploratory experiment, our interest in RASTA-based speech enhancement was piqued.

IX. Conclusion

The RASTA processing technique presented in this paper employs band-pass filtering of time trajectories of speech feature vectors. In principle, the RASTA processing can be done on time trajectories of any parameters (of course, with different effects). In this work, the processing was done on trajectories of critical-band spectral energies in the context of the previously proposed PLP analysis and applied between two static nonlinearities.

When dealing with a purely convolutional noise, the optimal compressive static nonlinearity appears to be the logarithmic function, and the expansive static nonlinearity is its exact inverse, i.e., the antilogarithm.

However, when dealing with a more realistic situations involving a combination of convolutive and additive noises, the compressive nonlinearity should be dependent on the SNR and be approximately linear for small spectral energies and approximately logarithmic for large ones. The expansive nonlinearity remains the antilogarithm. This creates the situation in which the analysis result is dependent on the particular compressive nonlinearity. Since the optimal compressive nonlinearity depends on the SNR, additional strategies must be used to reduce the effect of this source of variability. We have demonstrated that a recognizer incorporating such a strategy appears to be relatively robust to variations in noise conditions.

RASTA processing can be also used for the enhancement of noisy speech. In such a case, an overlap-add analysis-resynthesis technique is applied to the cubic root of the power spectrum of noisy speech, which has been RASTA filtered and then cubed. The noisy phase is preserved for the signal reconstruction. Results appear to be comparable with the conventional spectral subtraction method while alleviating the need for an explicit determination of nonspeech intervals.

There are several points we have learned in the course of the work described here:

- RASTA processing increases the dependence of the data on its previous context. Therefore, the performance of simple context-independent subword-unit recognizers can be degraded by RASTA processing. We have seen that RASTA processing works well in tasks with whole word models (such as many of the tasks reported in this paper) or in phoneme-based recognizers that used triphones or broad temporal input context (such as the neural net based recognizer used in our large vocabulary experiments).
- RASTA processing in the logarithmic (cepstral) domain does not address the problem of additive noise. Lin-log RASTA processing, which is described in Section V, appears to handle both additive and convolutional noise reasonably well.
- Some users have had difficulty with initial conditions. One needs to be aware that the RASTA filter incorporates a filter with a significant memory. Thus, it is different from the well-established short-term spectral analysis of speech in which each analysis frame is entirely independent of its surroundings.

Frame-by-frame analysis of speech dates from early days of speech analysis-resynthesis. RASTA processing represents a departure from this paradigm. We believe it is a step in the direction of modeling some temporal properties of human auditory processing. It has potential for further improvements as we learn more about the modeling of human auditory perception.

Appendix

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

Eleven isolated digits and two control words ("yes" and "no") were recorded at 8 kHz by 200 talkers over dialed-up telephone lines. All words were hand end pointed. Initial experiments used a subset of 30 speakers. The recognizer was a DTW-based nearest-neighbor multitemplate recognizer. Twenty seven talkers out of 30 were used in for training of the recognizer in a "leave-three-out" experimental design. In the "leave-three-out" design, three templates out of 30 are held for test, and the 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 52 780 recognition trials per experimental point.

The experiments in Section VI were expanded versions of the above in which all 200 speakers were included. In this case, results were the average of four runs, in which 150 speakers were used for training and 50 for test; the test speakers were rotated so that ultimately, all 200 had been tested with systems trained from independent speakers. For these experiments, the HMM tool kit (HTK) was used to train Gaussian mixture HMM's, as opposed to the DTW templates of the earlier experiment. In this case, each of the 13 words were modeled by ten states (including a nonemitting initial and final state), and there were four mixtures per state. Covariance matrices for each mixture were assumed to be diagonal (i.e., only means and variances were computed).

Recognition features were exponentially weighted [8] (exp = 0.6) five cepstral coefficients (zeroth coefficient excluding) of the fifth-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 the windows closed running at 55 mi/hr on a freeway. This noise has some natural slow variations (passing cars), and its long-term spectrum has a peak at about 600 Hz with a following spectral slope of about -12 dB/oct. The noise was added at several signal-to-noise ratios. The SNR's given in the paper represent ratios between the averaged energy over the whole utterance and the averaged energy of the added noise. Note that this will in general be a lower number than one would expect to see from a peak-to-average measure such as is used in the NIST SNR standard.

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). Eight cepstral and eight Δ cepstral coefficients (including the zeroth cepstral coefficient) of the eighth-order PLP or RASTA-PLP all pole model over the nine-frame window [9] 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.

The Credit Card experiment used a 4–hr subset of the 250 hr of the standard Switchboard database (which is available through the Linguistic Consortium). This database contains spontaneous telephone conversations on various subjects. The Credit Card portion consists of conversations about credit cards and uses about 2000 different words. Data were divided into single-speaker turns. Eleven hundred and two (1102) turns were used for the training of the HTK (HMM continuous density) software toolkit recognizer. The word-pair grammar used in the recognizer was derived on the 258 turns of the test set. The speech contains many kinds of spontaneous speech effects such as stutters, hesitations, restarts, interruptions, and poor articulations. Otherwise, it is of reasonable acoustic quality with relatively little additive noise.

ACKNOWLEDGMENT

This work benefited from the input of A. Bayya, H.-G. Hirsch, P. Kohn, J. Koehler, and G. Tong. Thanks to J. Cohen for having piqued our interest in the "inverse vowel spectral envelope" problem, which led to the original RASTA concept. Thanks to C. Wooters, Y. Konig, and S. Renals of ICSI for assistance with the large vocabulary experiments. Last but not least, we thank S. Greenberg and the ICSI speech research team for stimulating discussions.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *Proc. IEEE ASSP-27*, Apr. 1979, pp. 113–120.

[2] J. Cohen, *Personal Communication*, 1990.

[3] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[4] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Tokyo), 1986, pp. 1991–1994.

[5] G. Green, "Temporal aspects of audition," Ph.D. Thesis, Oxford, 1976.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Amer.*, pp. 1738–1752, 1990

[7] ——, "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 Res. Rep., File Folder ST 04-01, Oct. 1990.

[8] H. Hermansky and J. C. Junqua, "Optimization of perceptually based ASR front-end," in *Proc. ICASSP'88* (New York), 1988.

[9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (San Francisco), 1992, pp. I-121–I-124.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," in *Proc. 1993 IEEE Speech Recogn. Workshop* (Snowbird, UT), Dec. 1993.

[11] 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*, (Genova), 1991, pp. 1367-1370.

[12] O. Ghitza and M. M. Sondhi, "Hidden Markov models with templates as nonstationary states, an application to speech recognition," *Comput. Speech Language*, vol. 2, pp. 101–119, 1993.

[13] H. G. Hirsch, P. Meyer, and H. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," *Proc. EUROSPEECH'91*, (Genova), 1991, pp. 413–416.

[14] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR estimation and speech enhancement," Tech. Rep. TR-93-012, ICSI Berkeley, 1993

[15] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, and G. Tong, "Integrating RASTA-PLP into speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech. Signal Processing* (Adelaide, Australia), 1994.

[16] G. S. Kang and L. J. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 37, no. 6, pp. 939–942, June 1989.

[17] R. Kay, "Hearing of modulation in sounds," *Physiolog. Rev.*, vol. 62, no. 3, p. 917, July 1982.

[18] N. Morgan and H. Hermansky, "RASTA extensions, Robustness to additive and convolutional noise," in *Proc. Workshop Speech Processing Adverse Environments*, (Cannes, France), Nov. 1992.

[19] N. Morgan, H. Hermansky, H. Bourlard, P. Kohn, and C. Wooters, "Continuous speech recognition using PLP analysis with multilayer perceptrons," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, (Toronto, Canada), 1991, pp. 49-52.

[20] H. Ney, "Statistical modelling and dynamic programming in speech recognition," *Sprache unt Datenverarbeitung*, vol. 8, no. 1/2, pp. 17–33, 1984.

[21] R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Psycholog. Rev.*, vol. 31, pp. 867–875, 1928.

[22] R. Schwarz *et al.*, "Comparative experiments on large vocabulary speech recognition," *Proc. ARPA Workshop Human Language Technol.*, (Plainsboro, NJ), 1993.

[23] T. Stockham, T. Cannon, and R. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, pp. 678–692, Apr. 1975.

[24] Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *J. Acoust. Soc. Amer.*, vol. 81, no. 3, pp. 700–708, Mar. 1987.

[25] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput. Speech Language*, vol. 3, pp. 151–168, 1987.

[26] E. Zwicker, "Die Grenzen der Hoerbarkeit der Amplitudenmodulation under der Frequenzmodulation eines Tones," *Acustica*, vol. 2 pp. 125–133, 1952.

[27] CD with software, reports, and results from the *Frontiers in Speech Communication 1993 Summer Workshop*; available from the National Institute of Science and Technology.

Hynek Hermansky (M'86) received the Ing. degree from the Technical University Brno, Czech Republic, and the Kogaku Hakase (Doctor of Engineering) degree from the University of Tokyo, Japan, both in electrical engineering.

He is an Associate Professor at the Oregon Graduate Institute, Portland, OR, with a joint appointment at the Department of Electrical Engineering aned Applied Physics and at the Department of Computer Science and Engineering. He is also affilited with the International Computer Science Institute, Berkeley, CA. He has previously been with the faculty of the Technical University Brno, the University of Tokyo, Panasonic Technologies (in the Speech Technology Laboratory, Santa Barbara, CA), and U S WEST Advanced Technologies (Boulder, CO). His current research interest is in speech processing, in particular, in stimulating human processing strategies.

Nelson Morgan (SM'86) received the B.S., M.S., and Ph.D. degrees from the Electrical Engineering and Computer Sciences (EECS) Department at the University of California at Berkeley in 1977, 1979, and 1980, respectively.

From 1980 to 1984, he conducted and directed research in speech analysis, synthesis, and recognition, as well as DSP architectures, at National Semiconductor, Santa Clara, CA. From 1984 to 1988, he worked at the EEG Systems Lab in San Francisco on the analysis of brain waves collected in controlled experiments on cognitive behaviors. In both jobs, he developed connectionist approaches to signal analysis. In 1988, he joined the International Computer Science Institute (ICSI), which is a nonprofit research laboratort closely associated with the EECS Department at Berkeley. His position is that of a research scientist and group leader for the Computer Engineering Department, which is called the Realization Group. His current interests include the design of algorithms, architectures, and systems for parallel signal processing and pattern recognition systems, particularly using connectionist and perceptuall based paradigms. In July 1991, he received an appointment as an Adjunct Professor in Electrical Engineering and Computer Science at UC Berkeley.