INTEGRATING RASTA-PLP INTO SPEECH RECOGNITION

Joachim Koehler[†], *Nelson Morgan*[†], *Hynek Hermansky*[†],[‡], *H. Guenter Hirsch*[*], *Grace Tong*[†]
ICSI and U. of California, Berkeley, California[†]
Oregon Graduate Institute, Portland, Oregon[‡]
University of Aachen, Germany[*]

ABSTRACT

In previous work, we and others have shown that bandpass filtering of temporal trajectories of simple functions of the critical band spectrum can lead to more robust speech recognizers in the presence of additive and convolutional error. In this study we report results on several mechanisms for incorporating this analysis technique into training, in a way that is consistent with on-line approaches to speech recognition. In particular, we show improved robustness to these forms of degradation for a system that maps the filtered spectral points using a linear regression computed from results of the different transformations.

1. INTRODUCTION

It has been demonstrated [1][2] that bandpass filtering of temporal trajectories of the critical-band spectrum (when it has been processed by a nonlinear transformation) is efficient in alleviating some harmful effects of both additive and convolutional noise. While the technique appeared to be effective, it raised two new problems:

1. The optimal form of the nonlinearity is dependent on the noise level. Thus, the noise power needs to be estimated for the analysis.
2. Since, depending on the estimated noise level, a different compressive nonlinearity may be applied in the analysis, the result is dependent on the noise level. In a sense this is a trade of a deterministic source of variance (the different nonlinearities used) for a stochastic one (the actual additive or convolutive noise).

Previous work [1] simply estimated the noise power from the non-speech part of the signal to address the first problem. The second problem was addressed by using multiple templates derived from the clean speech using a range of nonlinearities corresponding to the range of expected noise levels.

In the current work we estimate noise without requiring explicit speech detection. Further, we investigate three different techniques for compensating for the effect of the variable nonlinearity. The RASTA models derived for recognition need to match the models derived during training. This was always true for the early forms of RASTA in which the nonlinearity was fixed (a logarithm), but is nontrivial for a nonlinearity whose value is dependent on an adaptively determined parameter (noise level).

2. BACKGROUND

The basic idea of RASTA processing is to filter the temporal trajectories of speech parameters (e.g., critical band values) after they have been transformed by a static nonlinearity that (ideally) converts the major sources of environmental interference into an additive component. Over the last year we have been experimenting with a parameterized family of functions

$$Y_i = \log(1 + JX_i) \qquad (1)$$

where i is the critical band number.

For large values of JX_i, this function is close to logarithmic, while for small values it is close to linear. Experiments reported in [1][2] showed that the optimal value for J is dependent on the instantaneous noise power. To estimate this noise power, we use an approach developed by Hirsch[3] which uses the position of the principal mode of the histogram of energy in each frequency band as the noise power estimate for the band. The subband estimates are currently combined for a robust estimate of the total noise power. This noise estimation technique does not require any speech pause detection.

0-7803-1775-0/94 $3.00 © 1994 IEEE

Though the overall processing has been shown to provide some robustness, a drawback remains: the choice of different J values, as required by differing noise conditions, generates different spectral shapes and dynamics of the spectra. This means that the training system must contend with a new source of variability due to the change in processing strategy that is adaptively determined from the data. The rest of this paper is concerned with the solution of this difficulty.

3. APPROACHES TO HANDLING J VARIABILITY

We have been working on three approaches to handling this variability:

1. Multiple recognizers - several systems can be trained using a different J value for each one. Although clean speech is used for each training, the differing J factors provide a range to include the nonlinear function for cases that will be encountered. In the recognition phase, noise estimation is used to select a J value, and the corresponding recognizer is used. As will be shown, this works well, but several recognizers must be trained.
2. Multiple J values for one recognizer - given enough degrees of freedom in the trained system, training data can be processed for training with a range of plausible values for J. This only requires training a single system, but since this technique effectively increases the size of the training set, it requires more computing and possibly also more parameters in the classifier to account for the added variability.
3. Spectral mapping - the noise-level dependent choice of J introduces a deterministic source of variability into the analysis, which one should be in principle be able to compensate for. To this date, however, we have not determined a satisfactory analytic solution to this problem, and therefore we have decided to apply an empirically derived linear mapping which would transform the spectrum obtained from a J value corresponding to noisy speech to a spectrum processed with a J value for clean speech. In other words, we find a mapping between $log(1 + Jx)$ and $log(1 + J_{ref}x)$. For this approach, we have used a linear regression within each critical band. In principle, this solution reduces the variability due to the choice of J, and so minimizes the effect on the training process.

In the next section we describe experiments to test these three methods.

4. EXPERIMENTS AND RESULTS

We tested our approaches with a standard HMM recognizer which was built with the HMM-Toolkit (HTK)[4]. The recognizer used 10-state word-based HMMs, with 8 emitting states and output probability distributions based on N-Gaussian diagonal covariance matrices. The variances were tied across all HMM states of all models (grand variances). The speech was processed using a 25 ms Hamming window, and then parameterized into 9 PLP-cepstral values. The test database consisted of 13 isolated digits spoken by 200 speakers over dialed-up telephone lines. All words were hand end-pointed. To get enough training data to model the HMMs we divided the set of 200 speakers into 150 speakers for training and 50 speakers for testing. A jackknife procedure was used so that all speakers' data could be tested on, resulting in 4 iterations (no overlap of testing). To balance for the number of parameters, we used 4 mixtures per state for all cases but that of 4 recognizers; for this case we used a single mixture per state (a greater number of mixtures actually didn't substantially change performance for an earlier pilot experiment). To simulate additive noise we synthetically added car noise to the clean ($>$ 20 dB SNR) speech to yield a 10 dB SNR level. Convolutional noise was introduced by filtering the speech with a linear filter simulating the spectral ratio between an electret and carbon microphone. The recognition results are presented in Table 1. The first row gives the results when the environment for train and test phases are identical, and is in some sense a best case scenario for non-RASTA processing; often the testing condition is not available during training. In all other rows the training conditions were always "clean", i.e., the additive and convolutional errors were only applied to test data. The second and third row show the results obtained with PLP and log RASTA processing.

Note that log RASTA (called RASTA here) reduces the error rate for the filtered case but is not effective for additive noise. In this task, RASTA also appears to slightly improve the discriminability between the word classes in the clean case, as in fact one-third of the errors were eliminated with a log RASTA front end (with respect to a PLP front end).

The results using multiple recognizers are shown in the fourth row (J-RASTA-mult). This appears to work reasonably well in comparison with PLP or log RASTA, but there is still a noticeable degradation. In addition, there is a significant performance loss for the clean data.

The next row (J-RASTA-uni) uses one recognizer with data processed using different values of J. This is an HMM version of our multi-template approach [1] and appears to work better than the multiple recognizer technique, both for clean and noisy cases. This case only requires a single recognition step, and so is a fairly straightforward way of incorporating J-RASTA into a recognition system. However, it does still require training with multiple processings of the training data, which increases training time.

The final row shows the results from the linear mapping of filtered critical band values. In this case, J-RASTA-filtered critical band outputs from 10 speakers[1] are used to train linear regression models. We have used 2 coefficients for each of 15 critical bands. Thus, we map the J-RASTA-filtered values for small J (high noise) to the corresponding values for a larger J (low noise). In particular, for each of 3 different values of J (10^{-7}, 10^{-8}, and 10^{-9}), we compute a mapping

$$W_{iJ} = c_1 + c_2 Y_{iJ} \qquad (2)$$

where Y_{iJ} is the J-RASTA-filtered output for critical band i, and the coefficients are determined to minimize the mean-squared error between W_{iJ} and $Y_{iJ_{ref}}$.

The recognizer was trained with clean speech processed with $J = 10^{-6}$, and during recognition the optimal value of J was determined by a local estimate of noise level for the isolated digit. Then the J-RASTA-filtered critical band outputs were linearly mapped using the regression coefficients computed earlier. The performance of this method appears to be quite good. In particular, the score for the clean case is essentially the same as for RASTA (in this case actually better than for PLP), while the mapping approach for the degraded cases are better than for the other approaches (roughly equivalent to the J-RASTA-uni approach). Unlike the other approaches, recognition and training are both the same as for log RASTA, as only a simple deterministic mapping is required in the front end.

rec. env.	no conv noise		conv noise	
SNR in dB	clean	10	clean	10
PLP same env	95.0	90.0	92.8	89.9
PLP	95.0	63.0	75.1	49.6
RASTA	96.7	50.0	96.4	59.6
J-RASTA-mult	90.9	76.3	87.5	72.1
J-RASTA-uni	92.2	83.2	91.3	81.1
J-RASTA-map	96.3	84.4	94.3	79.2

Table 1. Recognition Performance in %

5. DISCUSSION

The techniques described here permit incorporation of J-RASTA processing in an HMM-based recognizer, at least for a small vocabulary isolated word recognition task. However, the first two of the three increase training time. The third (linear mapping) approach appears preferable from the data shown here (although the train-with-all J-RASTA-uni approach gives slightly higher performance for one condition). However, we do not have enough experience with this method to know whether the mapping is task or data-dependent.

While these techniques do provide significant robustness to additive and convolutional noise, it is clear that, in comparison to the performance on clean speech, there is a significant increase in error which remains. Aside from the smoothing they provide for fast non-speech events, RASTA techniques only handle the constant (or slowly-varying) components of non-linguistic variation.

We close with some caveats about the use of RASTA. In the 2 years since we first reported some RASTA results on recognition, many sites have experimented with related features. Due to the

[1] Nine of these speakers were independent of the test set; the tenth was one of the 200 speakers in the final testing.

many different conditions under which these tests were done, results varied from wonderful success to dismal failure with many cases falling in between. Fortunately, this variance does not appear to have a random cause; we have learned a few things about the use of RASTA in recognition of speech. Some of these points are:

- RASTA increases the dependence of the data on its previous context. Therefore, simple context-independent subword-unit recognizers can be degraded by RASTA. We have seen that RASTA has worked well in tasks with whole word models (such as the one reported here), or in phoneme-based recognizers that used triphones or broad temporal input context (the latter being used for our neural-network recognizers).
- Log RASTA does not address the problem of additive noise. J-RASTA in one of the forms described here appears to be able to handle both additive and convolutional noise reasonably well.
- Some RASTA users have had difficulty with initial conditions. One needs to be aware that RASTA incorporates a filter with a significant memory, and thus is different from the well-established short-term spectral analysis of speech in which each analysis frame is entirely independent of its surroundings. To illustrate this point, we originally had difficulty in the experiments reported here when some test files started off with a non-audio artifact which itself was cut off prior to pattern matching, but whose effect spread well into the useful part of the speech data due to the RASTA processing, degrading the performance.

6. ACKNOWLEDGEMENTS

The more recent forms of these experiments have been partially sponsored (at UC Berkeley) under the Joint Services Electronics Program by contract number F49620-93-C-0014. We also acknowledge continuing support from the International Computer Science Institute.

REFERENCES

[1] Hermansky, H., Morgan, N., Hirsch, H.G.: " Recognition of speech in additive and convolutional noise based on RASTA spectral processing", *IEEE Proc. ICASSP'93*, pp. 83-86, 1993

[2] Morgan, N., Hermansky, H.: "RASTA extensions: Robustness to additive and convolutional noise ", *Proc. Workshop on Speech Processing in Adverse Conditions*, Cannes, France, November 1992

[3] Hirsch, H.G.: "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement", *Technical Report TR-93-012*, ICSI, 1993

[4] Woodland, P., and Young, S.: "The HTK Tied-State Continuous Speech Recognizer", Eurospeech '93, pp. 2207-2210, 1993