DATA-DRIVEN RASTA FILTERS IN REVERBERATION

Michael L. Shire, Barry Y. Chen

University of California at Berkeley
International Computer Science Institute
Berkeley, California
{shire,byc}@icsi.berkeley.edu

ABSTRACT

In this work we test the performance of RASTA-style modulation filters derived under reverberant conditions. The modulation filters are constructed through linear discriminant analysis of log critical band energies in a manner described by van Vuuren and Hermansky. In previous work we had observed the properties of the resultant filters under a number of acoustic conditions that were artificially applied to the training speech. Here, we present automatic speech recognition results that compare the performance of these filters under some training and testing reverberant conditions. We also test the effectiveness and robustness of a multi-stream combination using probability streams trained under different reverberant environment. The experiments reveal some performance improvement in severe reverberation.

1. INTRODUCTION

Robustness to reverberant acoustic conditions is a challenging problem in automatic speech recognition (ASR). The effects of reverberation and temporal smearing have been studied by researchers, for example in [14, 6], and efforts to mitigate their effects have been pursued, for example in [1, 13]. One potential impediment may lie in using a single preprocessing algorithm to handle the various acoustic conditions. As finding a single preprocessing method that is robust to all acoustic conditions is a daunting task, we instead sought to augment the preprocessing by deriving filters optimized in reverberant conditions. Such optimized filters may reduce some of the variability caused by reverberation and lead to increased robustness. Modifications to existing preprocessing and development of new techniques have progressed with some success, for example RASTA-PLP [9, 11] and Modulation-Filtered Spectrogram [8, 12]. Here we explore using RASTA-style filters derived through linear discriminant analysis (LDA) for robustness to reverberation.

Previous work described the properties of modulation filters derived using LDA when the speech was artificially subjected to different acoustic conditions [15, 17]. We had noted a tendency for these filters to prefer different frequency ranges, such as frequencies up to 13Hz commensurate with phonetic rates when derived in a clean environment and lower frequencies around 5 Hz (syllabic rates) when derived in reverberant conditions. In this paper we continue the experiments of obtaining filters from both clean and reverberant speech. We follow this with recognition experiments using phonetic targets. Our recognition experiments show that the derived filters improve performance under highly reverberant testing conditions and therefore may be useful as a supplement to other preprocessing methods. We subsequently ran tests for combining two probability streams, one trained in clean and the other in a reverberant condition.

2. EXPERIMENTAL SETUP

2.1. Modulation filter derivation

The modulation filters were calculated in a manner identical to that described in [2, 17]. To summarize, speech was analyzed into power spectral energies that were spaced along a bark-scale and followed by a logarithm as done in RASTA-PLP [9]. The trajectories were sampled in windows of approximately 1 second duration, and each window was assigned to the linguistic class present at its center. In this work, the linguistic classes were phones. From the windowed trajectories, the within-class covariance S_W and the between-class covariance S_B were computed [7]. The eigenvectors of $S_W^{-1}S_B$ having the largest eigenvalues were taken as the discriminant filters.

2.2. Training and testing

This approach involves two training corpora, which were quite distinct from one another, to promote generality. The design of the RASTA-style filters was based on one labeled corpus. Then the full speech recognition system was trained and tested on the second corpus, using filters from the first stage in the feature extraction. The training utterances for the filters were from the English portion of the Oregon Graduate Institute (OGI) Multi-Lingual Database [4]. It consisted of 210 continuous and naturally spoken utterances regarding various topics. These were recorded over the telephone. The utterances were approximately 1 minute in duration and were hand-labeled with phonetic units. A subset of the OGI Numbers corpus [5] was used for recognition experiments. This corpus consisted of naturally spoken connected numbers recorded over the telephone and has a small vocabulary size of 32 words. The training set consisted of approximately 3 hours of speech while the development testing set contained about 1 hour of speech.

A hybrid Artificial Neural Network (ANN) and Hidden Markov Model (HMM) [3] speech recognition system was used to evaluate the recognition performance of the different filters used in the preprocessing. Three layer multi-layer perceptrons (MLP), having 800 hidden units and an input context of 9 frames of speech features, were trained with the Numbers training set and hand-labeled phonetic transcriptions to estimate posterior probabilities. A small portion of the training set served as a cross-validation set for the stopping criterion of the MLP training. The decoder, which produced the best word transcription from the probability estimates,

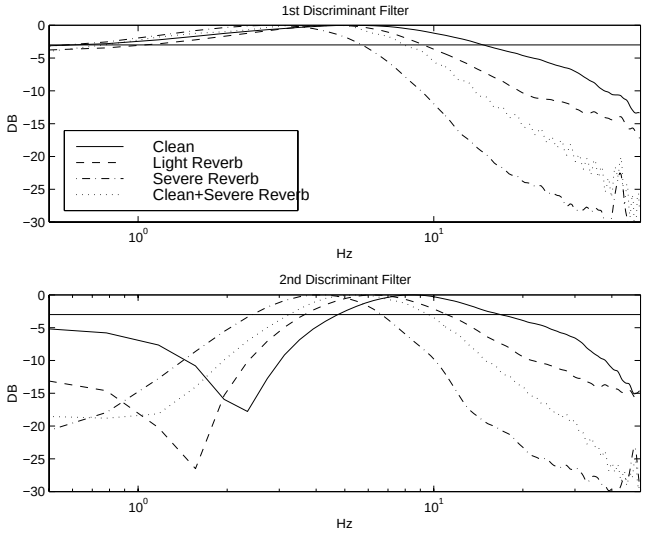

Figure 1: Frequency responses for the two principal discriminant filters derived in different acoustic conditions.

used a bigram grammar and dictionary of word models derived from the Numbers phonetic transcriptions.

2.3. Acoustic conditions

For the experiments here, the training and testing speech utterances were used in their original state (clean condition) and also modified with two examples of reverberation. The first consisted of light reverberation whose impulse response was recorded in a variable echoic chamber. It had the quality of a small office with a reverberation time (T_{60}) of 0.6 seconds and a direct-to-reverberant ratio (DTRR) of -1.9 dB. The second reverberation example consisted of severe reverberation whose impulse was recorded in a concrete basement hallway and with a T_{60} of 1.7 seconds and of DTRR of -16 dB. Each reverberation example was artificially added to the speech via convolution with the reverberation impulse response.

3. MODULATION FILTERS

As observed in previous work, the modulation filters exhibit a "Mexican hat" shape with a bandpass characteristic. Further, for each acoustic condition, they are consistent across the frequency bands. The first principal components from the different bands are virtually identical. The second and third components show a consistent shape across many bands; however, there is a tendency for the lowest few bands to be smoother and more low-pass. For this work, we averaged the filters across all frequency bands, then applied a Hamming window, to generate replacements for the RASTA filter in the RASTA-PLP feature extraction. We did this in turn for each of the three principal discriminant filters.

Figure 1 shows the frequency responses of the two principal discriminant modulation filters trained under four conditions: clean, light reverberation, severe reverberation, and a composite of clean and severe reverberation. A line is placed at the half power level (-3 dB) to assist in observing filter ranges. The first two components together explain between 85% and 95% of the variation. We note the band-pass nature of the filters and that with increasing reverberation, the discriminant filters tend to favor lower frequency ranges. Training the filters with syllabic targets (not presented here) demonstrated trends previously noted in [15] in that the responses were broader and favored lower frequency ranges commensurate with syllabic rates.

The final filter in the figure was derived from both clean data and severely reverberated data. The classes in the clean data had larger variances than the classes under severe reverberation. Since LDA has an implicit assumption of uniform class covariances, the clean condition would dominate and the response would resemble the filters trained solely in that condition. We therefore added a relative weighting when combining the covariance matrices of the two conditions. As we adjust the relative weighting, the derived filters would continuously "morph" from the clean condition filter to the severe condition filter. That is, the shape would approach one of the two extremes and the high frequency cut-off would similarly shift between the ranges bounded by each extreme. We can see in figure 1 that the response for a particular weighting seems to lie between the clean and severe reverberation responses, just as with the light reverberation response.

4. RECOGNITION RESULTS

We replaced the RASTA filter in log-RASTA-PLP with each discriminant filter to compute 8 cepstral feature coefficients. The features from the first, second, and third principal discriminant filters trained in identical conditions were concatenated into a single feature vector for input into the MLP. For comparison, we also ran experiments with standard log-RASTA-PLP with delta and double delta coefficients, yielding the same number of features. As described in [17], the second and third discriminant filters behave like first and second derivatives of the principal filter.

Table 1 shows word error rate (WER) results from word recognition for conditions in which the MLP probability estimator was trained in the clean, light reverberation, and severe reverberation conditions respectively. In the case of the LDA derived filters, we used the filter that corresponded to the same condition with which the MLP probability estimator was trained. Previous tests with the LDA filter trained under dual conditions showed results similar to those using the filter trained in light reverberation and are not shown. Where the difference in WER between the RASTA baseline and the LDA-derived filter was statistically significant (p=0.05), the item is displayed with a superscript plus or minus, for when the derived filters performed respectively better or worse than RASTA. 4673 words comprising 1206 utterances were in the development test set in this Numbers task.

5. PROBABILITY STREAM COMBINATION RESULTS

Since recognition rates are at their best when trained and tested under like conditions, we predicted that having two streams (each trained under a different acoustic condition) could improve performance over a wider range of conditions. We note that in the single stream cases, the original RASTA filter performs best in the clean condition while the LDA derived RASTA filters, particularly the filter with a 5 Hz range, performs better in the severely reverberated case. An extra information source which detects the acoustic test condition could conceivably serve as a switch to select which probability stream to use. Since often, such a switch is unavailable we experimented with a frame level combination of the probability streams.

The experiments here suppose that we have two probability estimators, one trained on clean speech and the other on severely

Train Conditions		Test Condition WER (%)		
		Clean	Light Reverb	Severe Reverb
Clean	RASTA	6.6	18.9	55.9
	LDA-FIR	7.4	25.4^{-}	43.0^{+}
Light Reverb	RASTA	32.0	13.3	43.6
	LDA-FIR	40.6^{-}	12.2	34.6^{+}
Severe Reverb	RASTA	79.8	70.3	37.7
	LDA-FIR	77.8^{+}	69.2	33.5^{+}

Table 1: WER scores using RASTA-PLP and LDA derived RASTA filters.

Merge Type	Filter Type	Test Condition WER (%)		
		Clean	Light Reverb	Severe Reverb
Avg	RASTA	10.1	25.0	47.2
	LDA-FIR	9.5	30.6	54.5
Log Avg	RASTA	17.5	24.2	48.7
	LDA-FIR	15.5	27.3	48.0
MLP	RASTA	14.8	26.2	48.2
	LDA-FIR	7.7	18.4	33.5
Oracle	RASTA	5.1	12.5	30.3
	LDA-FIR	5.3	15.8	26.9

Table 2: WER scores for dual probability streams.

reverberant speech. We then test the ability to productively combine these streams when tested with these two conditions as well as the third condition of light reverberation. An optimum combination strategy in this case is still a matter of research; here we try a few basic methods for combining probabilities. Table 2 shows a number of results when combining the probability streams with different methods. The first and second methods use an average of the probabilities and of the log probabilities respectively. We see that the performance lies in between the performance for a matching training and testing condition and a mismatched one when using a single stream. We also tested methods using straight multiplication and an entropy weighting criterion with similar results.

Some researchers have found that MLP mergers provide the best merging results [10, 16]. The next test in table 2 show results when using an MLP as the probability merger. We see that results here approach the best results from using either stream alone; that is, the results were as if one or the other streams was used. An unpleasant effect is that the MLP merger must be trained and we found with many further tests the trained combination did not generalize well. In these cases where the MLP merger did not do well,

Filter	Clean	Light Reverb.	Severe Reverb.
RASTA	83.6%	72.1%	88.2%
LDA-FIR	82%	68.0%	87.1%

Table 3: Overlapping frame classification errors for testing conditions and streams used in table 2

the simpler combination strategies such as multiplication work better. We should note that we had to train the merger MLP with the same training set used to train the probability estimation. Lack of sufficient speech material in the Numbers corpus prevented us from using an independent set which would have been more ideal.

The last experiment set of table 2 show results using an oracle to choose between the streams on a frame by frame basis; the oracle picks the stream with the best correct phone probability. They suggest a practical upper limit to the improvement obtainable using a combination of the two streams. The oracle significantly improves the WER compared to the WER of either stream taken singly. Moreover the oracle merging in the mismatched test of light reverberation approaches the performance of the matched training and testing case in table 1.

6. DISCUSSION

The difference between the original RASTA filter and the LDA derived ones appears greatest in the severely reverberated cases. The LDA filter exhibits a narrower bandwidth with a high cutoff down to 5 Hz for the severe case. Our tests corroborate views that this modulation rate is better suited to reverberation. A potential problem with deriving LDA-filters in a particular acoustic environment is that it may be useful only in that environment, or ones very similar. This is a general problem even with probability estimation, where the probability estimator is trained on a restricted set of examples that is in general not representative of the testing set. Adaptation and noise reduction have helped to alleviate some of this problem but the challenge remains. For the case of reverberation, merely training the ASR system on reverberant examples will improve performance on reverberated speech tremendously. Adjusting the feature extraction with these LDA-derived filters further improves the performance significantly. Since reverberation is a difficult case, insights gleaned from observations of the derived filter responses prove useful. Unfortunately, with such an ASR system trained to reverberation, performance then suffers when tested with non-matching acoustic conditions.

This motivated us to test a combination strategy where we trained two streams on two different conditions and tested on these same conditions as well as a third condition. A simple stream selector based on the condition of the test data could in principle give us some robustness by choosing the stream that performs best. Oracle tests suggested that a proper combination of streams could even improve upon the performance of either stream alone as well as increase robustness in mismatched conditions. Tests with our simple merging schemes, unfortunately, were not able to improve upon the obvious selection strategy of passing the stream that better matches the tested acoustic condition. The probable cause for this is that when one stream is performing optimally, the second is at its worst and is therefore of little help. An examination of the confusion matrices reveal that the streams frequently make the

same types of classification errors (comparing the correct phone class with the highest probability estimate) and therefore do not complement one another well. Table 3 lists the ratio of the concurring classification errors of the two streams to the minimum total misclassification between the streams. The large number of overlapping errors is not surprising in retrospect, considering both sets of features contain the exact same processing with only a difference in the preferred frequency range of the modulation filters. A more dramatic difference in processing may be needed for the streams to complement one another in such a way as to produce orthogonal errors. This may partly account for the success reported in [18] using RASTA-PLP and Modulation-Filtered SpectroGram features. Since some replacement RASTA filters demonstrate utility in reverberation its use in combination with another different feature process in a multi-stream setting may produce more favorable results. Further, since the filter derived in severe reverberation operates at syllabic rates, they may be useful in a multi-stream setting using syllabic targets in addition to phonetic targets.

7. CONCLUSION

A common problem with current ASR systems that the performance can degrade severely when it is presented with speech that is in a different acoustic environment than that trained. In the work here we derived and tested RASTA-style modulation filters using LDA on reverberant speech. The main augmentation was that the preferred pass-band frequency range lowered to more syllabic rates in the presence of reverberation. These filters demonstrated improvement in ASR recognition over the original RASTA filter in the case of severe reverberation, where the optimal LDA filter was most different. However, performance of this filter on the clean speech became abysmal. We ran experiments using a multi-stream setting where we combined the probability estimates from clean- and severe reverberation-trained filters and probability estimators using a number of combination schemes. A trained MLP merger was able to approach the performance of the best of the constituent streams taken singly. Simpler merging schemes that did not require training fared worse. Poor performance of the single streams on mismatched training conditions made simple combinations less robust. Moreover, as the streams were based upon the same processing strategy, they made many overlapping errors and therefore did not complement one another well. Future work will to apply the LDA filters in alternate multi-stream combinations.

8. ACKNOWLEDGMENTS

We would like to expressly thank Hynek Hermansky and Narendranath Malayath at OGI for discussions and their very valuable insights. We would also like to thank Nelson Morgan and ICSI for supporting this work. This work was funded by NSF, grant number (NSF)-IRI-9712579.

9. REFERENCES

[1] C. Avendano and H. Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. In *ICSLP*, volume 1, pages 889–92, Philadephia, Pennsylvania, October 1996.

[2] C. Avendano, S. van Vuuren, and H. Hermansky. Data based filter design for RASTA-like channel normalization in ASR. In *ICSLP*, volume 3, pages 2087–90, Philadephia, Pennsylvania, October 1996.

[3] H. Bourlard and N. Morgan. *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Press, 1994.

[4] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. OGI multi-lingual corpus, 1994.

[5] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995.

[6] R. Drullman, J. M. Feston, and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustic Society of America*, 95(2):1053–64, February 1994.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[8] S. Greenberg and B. E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, volume 3, pages 1647–50, Munich, Germany, April 1997. IEEE.

[9] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[10] A. Janin. Multi-stream speech recognition: Ready for prime time? In *EUROSPEECH*, pages 591–4, Budapest, Hungary, September 1999. ESCA.

[11] B. E. D. Kingsbury and N. Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP*, volume 2, pages 1259–62, Munich, Germany, April 1997. IEEE.

[12] B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–32, August 1998.

[13] T. Langhans and H. W. Strube. Speech enhancement by nonlinear multiband envelope filtering. In *ICASSP*, pages 156–8, Paris, France, May 1982. IEEE.

[14] H. Miyata and T. Houtgast. Weighted MTF for predicting speech intelligibility in reverberant sound fields. In *EUROSPEECH*, pages 289–2, Genova, Italy, Sept 91. ESCA.

[15] M. L. Shire. Data-driven RASTA filters in reverberation. In *EUROSPEECH*, pages 1123–6, Budapest, Hungary, September 1999. ESCA.

[16] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *ICASSP*, volume 2, pages 1255–8, Munich, Germany, April 1997. IEEE.

[17] S. van Vuuren and H. Hermansky. Data-driven design of RASTA-like filters. In *EUROSPEECH*, volume 1, pages 1607–1610, Rhodes, Greece, September 1997. ESCA.

[18] S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. In *ICSLP*, volume 1, pages 160–3, December 1998.