DATA-DRIVEN DESIGN OF RASTA-LIKE FILTERS

Sarel van Vuuren[1] *and Hynek Hermansky*[2]

[1,2] Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA
[2] International Computer Science Institute, Berkeley, California, USA
email: sarelv@ee.ogi.edu, hynek@ee.ogi.edu

ABSTRACT

We describe use of Linear Discriminant Analysis (LDA) for data-driven automatic design of RASTA-like filters. The LDA applied to rather long segments of time trajectories of critical-band energies yields FIR filters to be applied to these time trajectories in the feature extraction module. Frequency responses of the first three discriminant vectors are in principle consistent with the ad hoc designed RASTA, delta and double-delta filters. On a connected digit task the new features outperform the original RASTA processing.

1. INTRODUCTION

A typical automatic speech recognition (ASR) system contains a feature extraction module followed by a stochastic classifier. While the classifier is typically trained on training data, the feature extraction module is most often based on knowledge and beliefs. The knowledge applied in the feature extraction module has a critical role in the ASR process. Any information lost during the feature extraction is lost for the recognition process. On the other hand, the knowledge hardwired into the feature extraction module is the knowledge which does not have to be re-acquired from the data every time the recognizer is used for a new task.

In the late seventies, Hunt [1] proposed the use of Linear Discriminant Analysis (LDA) for deriving improved features for ASR. The LDA is applied to training data which contain sources of non-linguistic variability and the resulting transformation matrix is then a part of the feature extraction module which thus becomes more robust to the source of the particular non-linguistic variability.

The current paper presents a technique which applies LDA to rather long segments of a single time trajectory of critical band energy. Then, the LDA yields FIR filters to be applied to this time trajectory.

1.1. Temporal Domain and RASTA Technique

Acoustic feature vectors typically represent short-term characteristics of the speech signal. Standard HMM-based systems do classification over this short time span under the assumption of independence of the short-term acoustic vectors.

The peripheral human auditory system appears to be able to effectively integrate rather large time-spans (around 200 msec) of the audio signal [2]. Several emergent techniques employ short-term feature vectors from medium-span segments of speech. Among them, the RASTA technique [3] does band-pass filtering of time trajectories of speech features. To alleviate harmful effects of convolutive distortions, frequency components of time trajectories of logarithmic critical-band spectral energies below 1 Hz and above 12 Hz are attenuated. Such processing was found optimal by ASR experiments.

1.2. Toward a data-driven design

The initial ad hoc form of the RASTA filters was optimized on a relatively small series of ASR experiments with noisy telephone digits. The optimizations using these ASR experiments are costly and there is no guarantee that the solutions obtained will not be specific to a given ASR problem. Therefore, data-based optimization which would avoid using a specific ASR paradigm is desirable.

The linear discriminant analysis (LDA) is a stochastic technique which optimizes linear discriminability between classes (see e.g. [1] for examples of LDA in ASR). We examine the use of LDA for data-driven design of RASTA-like filters.

2. TECHNIQUE

Figure 1 shows the Linear Discriminant Analysis technique.

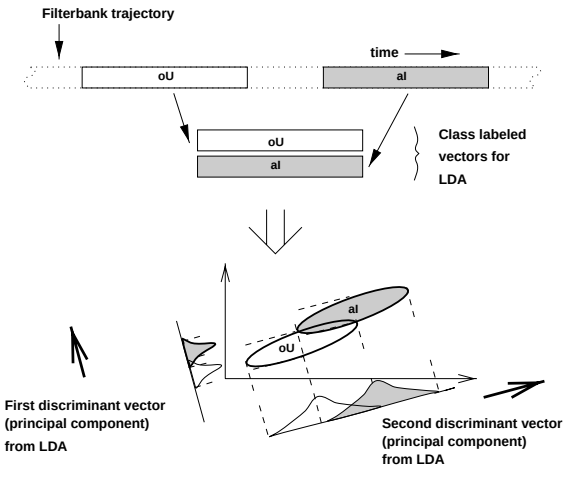

Figure 1. Linear discriminant analysis on segments of the time trajectory of a single logarithmic critical-band energy.

Sarel van Vuuren and Hynek Hermansky, "Data-Driven Design of RASTA-Like Filters," in Proceedings EUROSPEECH'97, Rhodes, Greece 1997.

The vector space for the LDA is constructed from segments of the time trajectory of a single logarithmic critical-band energy over a relatively long (typically about 1 s) span of time [4]. These segments are overlapped with a one frame spacing.

This approach is different from previous works using LDA [1, 5] since it applies LDA to *rather long time trajectories of features* rather than just to a single feature vector or to a relatively short block of feature vectors. This particular application of LDA results in the principal components (discriminant vectors) forming a set of FIR filters. The LDA on time-shifted segments of the trajectories therefore allows an FIR filtering interpretation of the analysis. This in turn allows us to directly relate the new LDA technique to other processing techniques such as RASTA processing. It should be noted, that up to certain constraints and assumptions[1], the LDA-based FIR filters map most efficiently (with respect to the within-class and the across-class variability) the vector space onto several points of the output space.

2.1. Databases

As LDA tries to optimize class separability in the presence of unwanted variability the result depends crucially on the type of nonlinguistic variability present in the data, as well as on the set of classes in the analysis. In the current paper we examine three different databases and two sets of classes. First, we apply the LDA to a hand-labeled subset of the Switchboard database. This database is labeled according to standard conventions into a set of 56 American English phonemes. Additionally, this database also contained classes of between-word pauses, and utterance beginning and end silences[6]. Second, the Switchboard database is appended by the identical database but with an added simulated convolutional variability. This is achieved by adding a constant approximately representing 2 standard deviations of the data to each time trajectory. Finally, the English portion of the OGI multi-lingual database is used with a representative set of phoneme classes for the analysis. Essentially this set includes prevalent phonemes in the speech and excludes silence. While we obtained the class assignments from a hand-labeled continuous speech corpus, we note that they may as well be obtained using automatic techniques such as forced alignment. Furthermore, as we will show, the LDA-based filters need not be designed and used on the same data. We will show that even when the filters are designed on a database different from the one on which they are eventually used they can still outperform other processing methods.

3. DISCRIMINANT VECTORS AS FILTERS

In previous work we [4] showed that the frequency response of the first discriminant vector agrees well with the frequency response of the ad hoc designed RASTA filter that smooths the feature trajectory. We stress the significance of this result. The discriminant vectors were designed entirely from the data without any intervention whereas the RASTA filter was iteratively optimized for on ASR experiments.

In Figures 2 to 5 we give frequency and impulse responses of the first three discriminant vectors derived on all three above described databases, as well as the frequency and impulse responses of the original RASTA filter and of the RASTA filter combined with the filters approximating the first (delta) and the second (double-delta) derivatives.

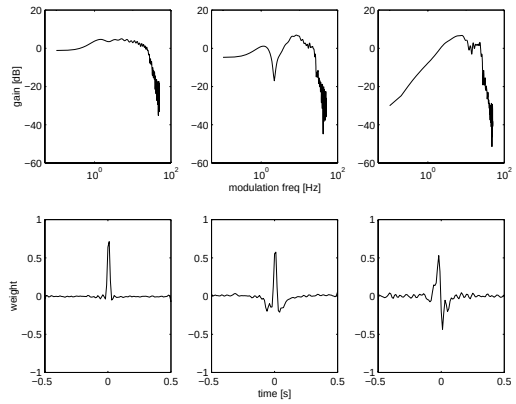

Figure 2. Frequency and impulse responses of the first three discriminant vectors derived on the clean Switchboard database.

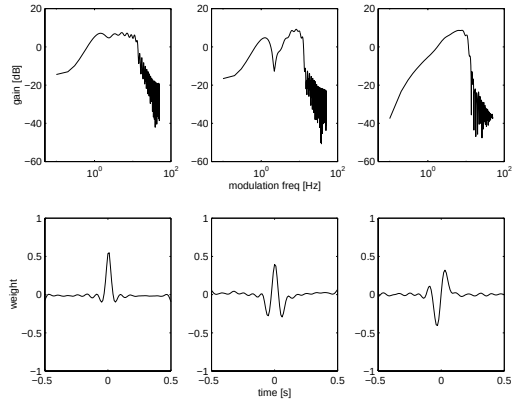

Figure 3. Frequency and impulse responses of the first three discriminant vectors derived on the Switchboard database with additional steady-state variability.

The first thing to notice is that, as expected, filters designed on the Switchboard data with the additional steady-state variability exhibit stronger suppression of low as well as of high frequencies[2]. The stronger suppression of low frequencies is expected because the additional variability is steady-state.

Filters designed on the OGI multilingual database do contain similar general characteristics as the filters derived on the Switchboard data but differ in details (for example the second and the third filters are interchanged). We note that

[1] It is assumed that the data is heteroscedastic.

[2] For the Switchboard experiments, which used only 30 minutes of speech data, to guarantee numerical stability, we enforced a condition number of 500 for the with-in covariance matrix. This conditioning caused a slight suppression for the high frequencies. We did not use this conditioning for the OGI data.

Sarel van Vuuren and Hynek Hermansky, "Data-Driven Design of RASTA-Like Filters," in Proceedings EUROSPEECH'97, Rhodes, Greece 1997.

a reduced class set of only the 20 most common phoneme labels was used with the OGI database.

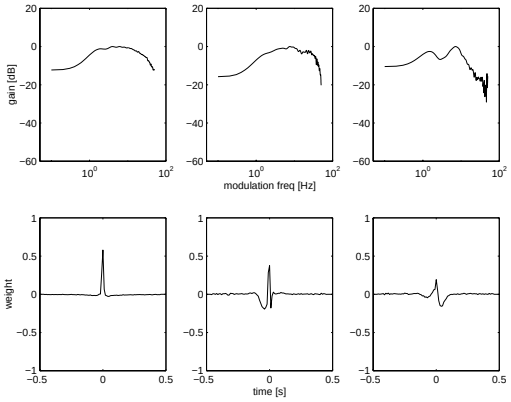

Figure 4. Frequency and impulse responses of the first three discriminant vectors derived on the English portion of OGI multi-lingual database.

The similarity of the first discriminant vectors from all three databases with the original RASTA filter is noticeable. The impulse responses of the first discriminant vector is approximately symmetric, implying close to zero phase and supporting de Veeth and Boves [7].

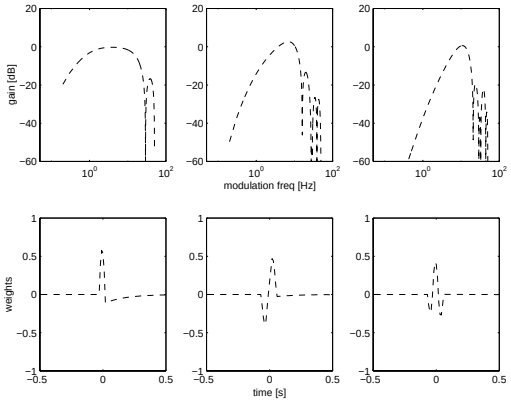

Figure 5. Frequency and impulse responses of the RASTA filter and the RASTA filter combined with the delta and double-delta filters.

The first discriminant vector, while being the most important for discrimination, explains only about 80% of the variability in the data. We therefore decided to investigate the second and third discriminant vectors as well. The frequency characteristic of the second and third discriminant vectors are somewhat comparable to the second (slope) and third (curvature) orthogonal polynomials approximating the time trajectory of the feature within a 9 frame (90 ms) time interval as proposed by Furui [8]. The second peak at around 1 Hz in the two-peak filters can be simulated by adding a small bias to the double-delta orthogonal polynomial.

As shown in Fig. 6 which depicts frequency responses of the first discriminant vector at all 15 carrier frequencies (there are 15 critical-band filters covering the telephone-bandwidth), filters at different carrier frequencies are rather similar.

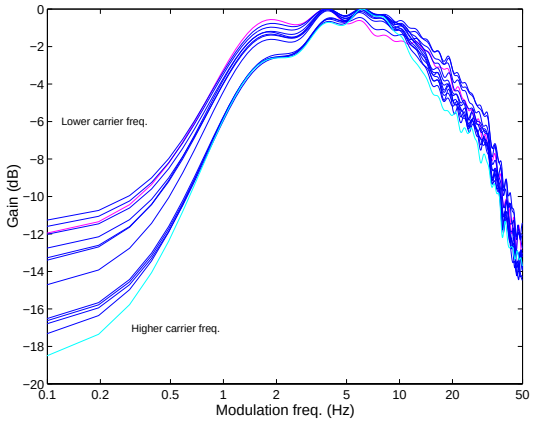

Figure 6. Frequency response of the first discriminant vector at all 15 carrier frequencies derived on the English portion of OGI multi-lingual database.

To further highlight the modulation frequency selective nature of the LDA-based filters Figure 7 shows the frequency response of the resultant first discriminant vector for the case where the log filter bank energies had a disturbance added at modulation frequencies of 5 and 20 Hz with respective amplitudes about 2.5 and 0.5 times the average standard deviation in the log filter bank energies. Such a disturbance can be thought of as a time-varying convolutive disturbance on the speech signal. As expected, the filter attempts to attenuate modulation frequencies at the disturbance.

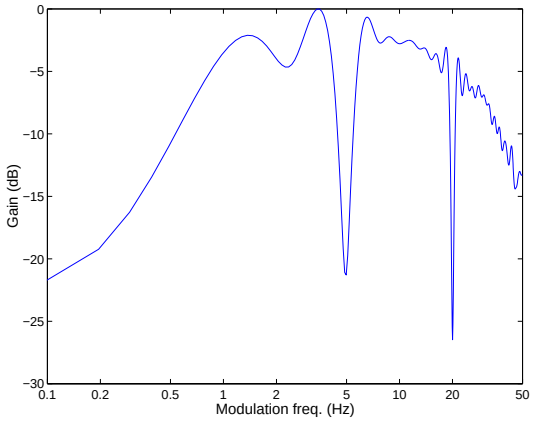

Figure 7. Frequency response of the first discriminant vector for an artificial non-stationary channel disturbance.

4. ASR RESULTS

For the results listed below the filters were derived on the English portion of the OGI multi-lingual database (OGI-TS). This is a database of almost 3 hours' continuous telephone

Sarel van Vuuren and Hynek Hermansky, "Data-Driven Design of RASTA-Like Filters," in Proceedings EUROSPEECH'97, Rhodes, Greece 1997.

speech with both speaker and handset variability. We performed the recognition experiments on 500 connected digit utterances from the OGI-Numbers corpus. This database also has speaker and handset variability. The phoneme classes used for deriving the filters were chosen to match the monophone classes expected in the recognition experiment. For the filter design we used a total of twenty monophones and weighed them according to their natural frequency of occurrence in the OGI-TS database.

The results are competitive with current filtering schemes such as RASTA. Table 1 lists the word level accuracy for the connected digit recognition task. A 5 state left-to-right HMM model was used with 3 mixtures per state. Twenty monophone models were trained and a simple single pronunciation grammar used. The baseline features (base) were critical-band log energies from a PLP analysis. The table lists accuracies for the baseline features processed with RASTA filtering (rasta) and with combinations of the first three LDA-derived filters (lda1, lda2 and lda3). Accuracies for when delta (delta) and double delta (ddelta) features are added are also listed. Features were normalized throughout with a full whitening transform. This normalization was necessary to ensure a fair comparison between the different features for mainly two reasons. a) Decorrelation: The HMM model used diagonal covariances. b) Scale: The HMM model used a numerical floor (1e-4) on the variances parameters.

In practise we found the whitening transform to give results similar to the DCT transform. To mitigate effects from the language back-end of the system and since it is known that different processing techniques exhibit different insertion and deletion trade-off [9] we report the word level accuracies at the optimum cross-word penalty.

base					83.7
rasta					88.0
lda1					91.8
lda2					85.1
lda3					81.7
base	+	delta			91.2
rasta	+	delta			92.4
lda1	+	delta			93.5
lda1	+	lda2			94.3
base	+	delta	+	ddelta	90.0
rasta	+	delta	+	ddelta	92.9
lda1	+	delta	+	ddelta	94.0
lda1	+	lda2	+	lda3	94.6

Table 1. Percentage word level accuracies for a connected digit recognition task (OGI-Numbers corpus) for the various processing techniques.

From the Table the basic LDA derived feature (lda1) is seen to generally outperform the baseline and RASTA processed features. The differences are significant at the 1% level using Mcnemar's test. These results suggest that while RASTA greatly aids performance on this database, other data-derived filters (here from LDA) may yield even better performance. This observation extends to the case where delta and double delta features are added. Given that the LDA filters were derived from another database and based entirely on the baseline feature and class labels we find the results highly encouraging.

5. CONCLUSION

We propose a new temporal filtering technique to optimize class discriminability. While many aspects of the technique are still subject to further research and optimization we find the performance of the *entirely data-derived* filters on recognition experiments highly promising.

ACKNOWLEDGEMENT

We thank Nagendra Kumar who provided a modified LDA program used in some of our initial experiments on the Switchboard data. We also thank Pieter Vermeulen for helpful discussion and Johan Schalkwyk for help with the HMM software. This work has been supported in part by grants NSF/ARPA (IRI-9314959) and DoD (MDA-904-94-C-6169).

REFERENCES

[1] M. J. Hunt, "A statistical approach to metrics for word and syllable recognition," *J. Acoust. Soc. Am.*, vol. 66(S1), S35(A), 1979.

[2] H. Hermansky, "Exploring temporal domain for robustness in speech recognition," in *Proc. of 15th International Congress on Acoustics*, vol. II, (Trondheim, Norway), pp. 61–64, June 1995.

[3] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[4] C. Avendano, S. van Vuuren, and H. Hermansky, "Optimizing rasta filters on corrupted speech," in *Proc. Intl. Conf. on Spoken Language Processing 96*, (Philadephia,PA), pp. 2087–2090, October 1996.

[5] P. Brown, *The Acoustic-Modeling Problem in Automatic Speech Recognition*. PhD thesis, Computer Science Department, Carnegie Mellon University, 1987.

[6] S. Greenberg, "The switchboard transcription project," in *Proc. 1996 CLSP/JHU Workshop on Innovative Techniques in Continuous Large Vocabulary Speech Recognition*, (CLSP/JHU), November 1996.

[7] J. de Veth and L. Boves, "Comparison of channel normalization techniques for automatic speech recognition over the phone," in *Proc. Intl. Conf. on Spoken Language Processing 96*, (Philadelphia, PA), pp. 2332–2335, October 1996.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[9] P. Bojan, O. Anderson, and P. Dalsgaard, "On the robust automatic segmentation of spontaneous speech," in *Proc. Intl. Conf. on Spoken Language Processing 96*, (Philadelphia, PA), pp. 913–916, October 1996.

Sarel van Vuuren and Hynek Hermansky, "Data-Driven Design of RASTA-Like Filters," in Proceedings EUROSPEECH'97, Rhodes, Greece 1997.