Data Based Filter Design
for RASTA-like Channel Normalization in ASR

Carlos Avendano, Sarel van Vuuren and Hynek Hermansky

Oregon Graduate Institute of Science & Technology
P.O. Box 91000, Portland, OR 97291 USA

ABSTRACT

RASTA processing has proven to be a successful technique for channel normalization in automatic speech recognition (ASR). We present two approaches to the design of RASTA-like filters from training data. One consists of finding the solution to a constrained optimization problem on the feature time trajectories while the other uses Linear Discriminant Analysis (LDA). Whereas LDA is often applied to one or a few frames of the feature vectors we apply LDA to feature time trajectories. Both approaches result in similar filters which are consistent with the ad hoc designed RASTA filter.

1. Introduction

Relatively unstructured data-driven systems are the mainstream in today's ASR. These systems acquire their structure from the large amounts of training data and are susceptible to failure when used in conditions that assume a different structure than that acquired during the training.

It is our belief that more knowledge-constrained and structured designs will result in simpler and ultimately more reliable systems. However, if we are to hardwire any constraints into the system, it is crucial that these constraints are based on well tested, reliable and relevant knowledge.

Some reasonable constraints may be implied by properties of the human hearing process and we have been relatively successful when incorporating them into ASR [1, 2]. On the other hand, it is hard to deny the power of real speech data. Thus, we support using the speech data, as long as they are used in a way to provide permanent and reusable knowledge.

Along this line, we came to realize that since speech developed to optimally use the properties of human auditory perception, any relevant auditory knowledge may have its counterpart in the structure of the acoustic speech signal, and the constraints derived from the data may either correct or support the knowledge-based constraints.

1.1. RASTA Processing

For the past several years we have been working on the incorporation of temporal auditory masking into speech processing [2]. As an engineering simulation of this powerful auditory constraint we pro-
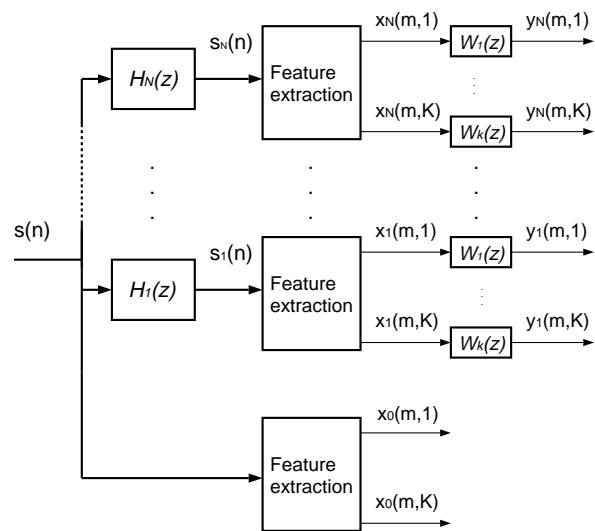

Figure 1: *Problem setup block diagram*

posed to filter out slow and fast changes in the trajectories of an critical logarithmic short-time spectrum of speech. The initial ad hoc form of the RASTA filters was optimized on a relatively small series of ASR experiments with noisy telephone digits. The experiments yielded a filter with a spectral zero at zero modulation frequency and a pass-band approximately from 1 to 12 Hz. The same filter was used for all frequency components.

Optimizations using ASR experiments are costly and there is no guarantee that the solutions obtained will not be specific to a given ASR problem. Any data-based optimization which would avoid using a specific ASR paradigm is desirable.

In the current work we use two optimization techniques to design a set of RASTA filters from realistic noisy data. The first is a constrained optimization technique which uses the natural time-varying structure of the continuous speech data. The second technique is based on a Linear Discriminant Analysis (LDA) [7] which uses phoneme-labeled speech data. The data used in both cases consisted of parallel speech recordings corrupted by different channels.

2. Filter Design by Constrained Optimization

In this section we find a set of filters for channel normalization by constructing a constrained optimization program. The design criteria is the distance minimization between the processed features when they are obtained from speech corrupted by different communication channels.

The idea can be explained with the aid of Fig. 1. A speech signal $s(n)$ is corrupted by N different channels $H_j(z)$. After an appropriate feature extraction procedure we have a set of NK corrupted time trajectories $x_j(m, k)$, where $k = 1, ..., K$ is the feature index, m is the decimated time index and $j=1,2 ...,N$. Notice that we also show a set of trajectories of clean speech which will be used to set the constraints described in the next section.

Ideally we would like to apply a transformation $W_k(z)$ on the time trajectories such that the outputs $y_j(m, k)$ for the kth feature are as similar as possible to each other, that is, we seek channel independence. A trivial solution to this problem is to set $W_k(z) = \bar{0}$ so the need to constrain the solution to some reasonable value is obvious. As will be seen later, the constraints will determine the behaviour of the filters at certain modulation frequencies.

2.1. Technique

Considering that the effect of the channel is multiplicative in the frequency domain we chose our transformation $W_k(z)$ to be a linear finite impulse response (FIR) filter applied to $x(m, k) = \log[a(m, k)]$, where $a(m, k)$ are critical band energy trajectories. Notice that in this case we assume one-dimensional (frequency specific) filters but the methods should still be applicable to a more general case. From now on we will call $x(m, k)$ the critical bands omitting the term logarithmic.

The FIR filter for each critical band was derived in the following way:

Let

$$ J_k = E\{\sum_{j=1}^{N-1} \sum_{i=j+1}^{N} [y_j(m, k) - y_i(m, k)]^2\} \qquad (1) $$

be the objective function for the kth critical band.

From (1) we see that the objective functions are defined as the expected value (with respect to time) of the Euclidean distance between the outputs $y_j(m, k)$ of the $W_k(z)$ filter produced by $x_j(m, k)$ for $j = 1,...,N$. These quadratic functions have a global minimum at $W_k(z)=\bar{0}$. To avoid the trivial solution in which all filter coefficients are set to zero we need to impose a set of constraints.

2.2. Experimental Design

To derive the filters, three parallel speech recordings were used. A sample of clean speech was taken from the TIMIT database (approximately 2 minutes). The other two samples were taken from the corresponding speech of the NTIMIT database (telephone channel) and TIMIT recorded through a cellular telephone channel. Auditory frequency band trajectories for the three recordings were computed by a

weighted sum of their short-term power spectrum as proposed in [1]. The logarithm of these trajectories was taken to produce $x_0(m, k)$, $x_1(m, k)$ and $x_2(m, k)$ respectively.

The same design was applied to each critical band independently. For simplicity we will drop the frequency index k and it should be understood that the following procedure was applied K times.

Using (1) we find the objective function

$$ J = \mathbf{w}^T R_{x_1, x_1} \mathbf{w} + \mathbf{w}^T R_{x_2, x_2} \mathbf{w} - \mathbf{w}^T R_{x_1, x_2}^v \mathbf{w} \qquad (2) $$

In (2), $\mathbf{w} = [w_0, w_1, ... w_{L-1}]^T$ is the vector of filter coefficients, R_{x_j, x_i} refers to the cross-correlation matrix between $x_j(m)$ and $x_i(m)$, and $R^v = R + R^T$. In this case $N = 2$ and we need to set two constraints ($j=1,2$), one for each of the outputs of the filter.

Constraints A reasonable constraint is to restrict the energy at the output of the filter to be a fraction of the energy of the input signal. While this constraint avoids the trivial solution, it imposes no restriction on the characteristics of the output signal and thus it will be less effective if we need to preserve relevant information about speech. Interestingly, by using this constraint we found that the filters had very similar responses (i.e. band-pass with strong dc suppression and narrow pass-band) as the so called delta cepstrum processing [4].

A more reasonable constraint was found by not allowing the distance between the filter outputs $y_j(m, k)$ and the original uncorrupted speech features $x_0(m, k)$ (see bottom of Fig. 1) to be large. This similarity constraint can be more or less restrictive depending on the amount of error allowed, and the resulting filters will have different characteristics depending on this factor. Notice that the availability of clean speech is only necessary for this particular constraint and is not a general requirement of the technique.

The constraints proposed above can be written as:

$$ E\{[\hat{x}_0(m) - \hat{y}_j(m)]^2\} < c_j \qquad (3) $$

In these constraints we removed the dc component from the original clean speech features and from the output of the filter to obtain $\hat{x}_0(m)$ and $\hat{y}_j(m)$ respectively. This normalization is needed in order to make a fair comparison of the clean and corrupted features (since adding or removing a constant in the logarithmic domain corresponds to modifying the power of the signals in the linear domain).

Writing (3) in matrix notation we get:

$$ S_{\hat{x}_0} + \mathbf{w}^T R_{x_j, x_j} \mathbf{w} - 2\mathbf{w}^T \mathbf{p}_{\hat{x}_0, x_j} - \mathbf{w}^T \hat{\mathbf{x}}_j \hat{\mathbf{x}}_j^T \mathbf{w} < c_j \qquad (4) $$

Here $S_{\hat{x}_0}$ is the power of $\hat{x}_0(m)$, $\hat{\mathbf{x}}_j$ is a vector with all elements equal to the mean of $x_j(m)$, and $\mathbf{p}_{\hat{x}_0, x_j}$ is the cross-correlation vector between $\hat{x}_0(m)$ and $x_j(m)$.

The constants c_j were initially chosen using a -3dB criteria on the power of the inputs to the filter and were varied systematically until

no feasible solution could be found. The last feasible point was chosen to be the new Constraint-OPtimized (COP) filter. The optimization program described by (2) and (4) is non-linear with non-linear constraints and was solved using sequential quadratic programming [5].

3. Filter Design Using Linear Discriminant Analysis

In this section we find a set of filters using LDA [1]. Our approach is different from previous works [3, 8] in that we apply LDA to feature *time trajectories* rather than just to feature frames. We will call this technique DISCriminant Optimized (DISCO) filtering.

3.1. Technique

We start with speech which is labeled into different classes. We wish to design a filter that will maximize discriminability between these classes. The technique is best explained with the aid of Fig. 2. The figure shows an FIR filter acting on the k'th time trajectory of feature $x(m)$ (which in this case is obtained by concatenation of $x_0(m,k)$,$x_1(m,k)$ and $x_2(m,k)$).
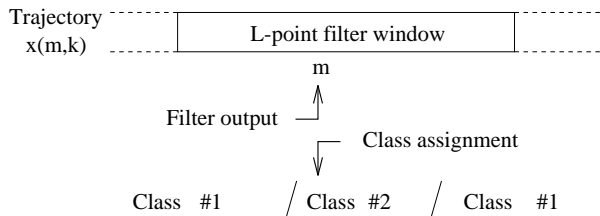

Figure 2: *DISCO filtering technique. Illustrated is an FIR filter centered at time m and acting on the k'th time trajectory of feature x.*

Referring to the figure, view the features on which the filter is acting at time m as a vector $\mathbf{x}_m$. For a filter of length L, this vector lies in an L-dimensional space. For every analysis step m, the output of the filter is a projection of the L-dimensional input vector onto a one-dimensional output space. Assigning a class to each vector $\mathbf{x}_m$, we can use LDA to find the principal directions for which class discriminability is maximum at the output. The first principal component of this analysis is the L-point FIR filter for which linear discriminability is maximized. A separate LDA is carried out for each critical band time trajectory, thus yielding a separate FIR filter for each band.

Fig. 2 also gives an idea of how we assign a class to the input vector $\mathbf{x}_m$. In the current work, we assume a non-causal FIR filtering. Thus, we align the center of the input vector $\mathbf{x}_m$ with the underlying label. This is not the only way to do the class assignment; e.g. had we decided for a causal FIR filter we would have aligned the class label with the first element of the $\mathbf{x}_m$. The next input vector is $\mathbf{x}_{m+1}$, i.e. it is formed by shifting the $\mathbf{x}_m$ by one analysis step.

In our work the vector space for the LDA is constructed from seg-

---

[1] As far as we know LDA was first used in speech by Melvyn Hunt [3].

ments of time trajectories of a single speech feature over a relatively long (typically at least a syllable) span of time. This is different from other LDA-based approaches (eg. [8]) that apply LDA to a feature vector or to a relatively short block of feature vectors.

3.2. Scope

In this paper we find a filter for each feature trajectory $x(m,k)$ separately but it should be understood that our technique is much more general:

- Our technique in principle extends to multivariate filtering or non-linear discriminants [7].

- It could allow for more than just the principal component of the LDA to be used.

- It allows for causal or non-causal filter design.

- It is not limited to any particular domain (e.g. since the aim was to remove linear distortions, the domain was the logarithmic critical band power spectrum while if the aim was to remove additive noise the domain might have been different).

- It applies to different types and sets of classes. In this paper we use broad TIMIT phonetic classes[2].

4. Results

In this section we present the results obtained by the techniques described above. If no distinction is made the reader should assume that the results discussed apply to both approaches.
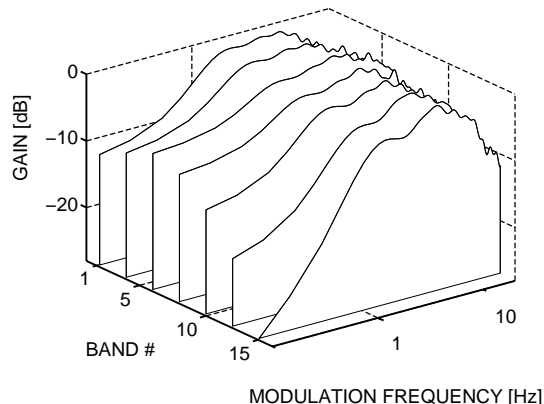

Figure 3: *DISCO filter bank*

Experimental designs were obtained for every critical band (in this case we used 15 bands) using both approaches and using the same data. For different bands we found no substantial difference in the frequency responses of the filters except for bands where speech had little or no energy and were dominated by noise. We illustrate this fact by showing the magnitude frequency responses for the DISCO

---

[2] We used the following classes: closure, back vowel, mid vowel, front vowel, fricative, stop, flap and affricate, retroflex, and nasal, but did not use silence and non-speech.
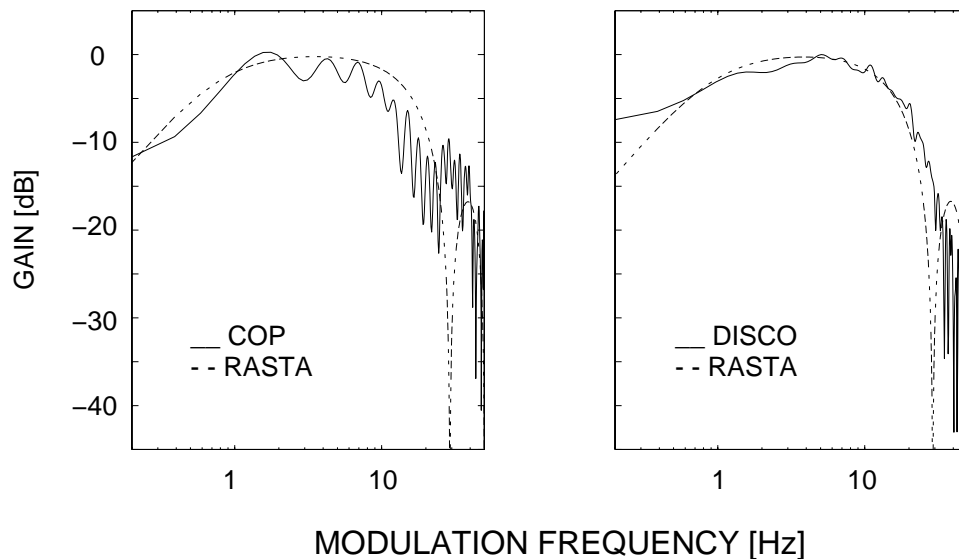
Figure 4: *COP and DISCO filters compared to RASTA*

filters at different critical bands (Fig. 3). We observe that in the last few bands there is a stronger dc suppression.

The general solution shows a strong suppression of dc and slow modulation frequencies. Differences between COP and DISCO filters can be observed in Fig. 4. COP filters have a considerably larger ripple in the pass-band which in this case is narrower. Suppression of high modulation frequencies seems to be stronger with DISCO filtering, while the opposite is true at low frequencies.

The observed frequency characteristics depend on the type of disturbance with which the speech data is corrupted. For instance, if we were to corrupt a sample of speech by convolving it with a periodically time varying linear system, the COP or DISCO filter would show a strong suppression at the corresponding modulation frequency.

The impulse response of the filters shows an almost even symmetry about the center tap (which was chosen to be the origin) thus indicating a nearly linear phase property. It should be mentioned at this point that the length of the filter is a very important parameter since it determines the behaviour at low modulation frequencies. In the results reported here we use 1 second long, 101 point non-causal filters.

It is obvious that there are many common characteristics between the new filters, and with the original RASTA filter. It is interesting to notice the modulation frequency range which is consistently preserved by all techniques and its correspondence with the results obtained in recent perceptual experiments [6].

5. Conclusions

Two new approaches for deriving RASTA-like filters from training data were presented. The results show that there exist common prop-

erties between them, pointing to the intrinsic characteristics of the speech signal and the relation to auditory models used in the past.

6. REFERENCES

1. Hynek Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," J. Acoust. Soc. Am. *87*(4),pp. 1748-1752 April 1990.

2. Hynek Hermansky and Nelson Morgan, "RASTA Processing of Speech," IEEE Trans. Speech and Audio Processing Vol.2 No.4, pp. 578-589, October 1994.

3. Melvyn J. Hunt, "A Statistical Approach to Metrics for Word and Syllable Recognition," J. Acoust. Soc. Am. *66*(S1), S35(A), 1979.

4. Sadaoki Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. ASSP, VOL. ASSP-34. No. 1, February 1986.

5. Roger Fletcher, *Practical Methods of Optimization*, John Wiley and Sons, 2nd ed., March 1991.

6. Takayuki Arai, Hynek Hermansky, Misha Pavel and Carlos Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelope," to appear in *Proceedings ICSLP 96*, Philadelphia, PA 1996.

7. Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 2nd ed., 1990.

8. R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," *Proceedings ICASSP-92*, pp. I 13-16, San Francisco 1992.