DATA-DRIVEN MODULATION FILTER DESIGN UNDER ADVERSE ACOUSTIC CONDITIONS AND USING PHONETIC AND SYLLABIC UNITS

Michael L. Shire

University of California at Berkeley, International Computer Science Institute
1947 Center Street Suite 600, Berkeley, California 94704 USA
email: shire@icsi.berkeley.edu

ABSTRACT

Constructing speech feature extraction methods that are robust to many types of corrupting acoustic environments remains a daunting task and it is instructive to investigate which properties of the speech carry the discriminative information for recognition under a variety of conditions. In this paper we describe results for generating RASTA-style modulation filters under a number of acoustic environments. We utilize Linear Discriminant Analysis in a manner previously described by van Vuuren and Hermansky to automatically generate discriminant filters for speech with artificially added background noise and reverberation. We also generate the filters using both phonetic and syllabic classification targets. Trends in the responses of the discriminant filters lend support to feature extraction design decisions employed by RASTA-PLP and Modulation-filtered Spectrogram features. Further, tests with added reverberation corroborate views on the perceptual stability of syllabic rates.

1. INTRODUCTION

One challenging problem in automatic speech recognition lies in dealing with corrupting acoustic environments such as additive noise and room reverberation. A number of preprocessing strategies have been developed in an effort to improve robustness to environmental conditions. In the case of reverberation, for example, researchers have modified the preprocessing to use longer temporal analysis windows [8] and have even used syllable class targets to better match these longer time windows [12]. Some of this effort has been guided by psychoacoustic evidence or in an empirical manner. In this work, we explore the use of discriminatively derived modulation filters to gain some insight into the modulation frequencies of importance under a variety of conditions. The analysis is performed using artificially added background noise and room reverberation.

Previously, van Vuuren and Hermansky designed filters to to be applied to the log-energy envelopes of subbands based on auditory critical-bands. The filters were derived in an automatic data-driven fashion using Linear Discriminant Analysis (LDA) [10]. Their method provided valuable insights concerning the temporal properties that are useful for discrimination. Results from their work showed similarities to the RASTA [7] bandpass filter and its first and second derivative. The bandpass properties were consistently seen across a number of speech corpora. Additionally, they showed how the filters were modified when artificially channel noise was added. In this work we continue some of these experiments and derive discriminant filters for other corrupting conditions including added car noise, added factory noise, and two examples of room reverberation.

Analysis by Greenberg and others has suggested syllables as a fundamental unit of speech perception and syllabic rates as a reasonable and robust level for analysis [9, 5]. Therefore, in addition to phonetic classes, we also perform tests using syllabic class targets.

2. EXPERIMENTAL CONDITIONS

Our experiments employed LDA on log critical-band trajectories to derive discriminant filters in a manner identical to that described in [10]. The training speech consisted of the English portion of the Oregon Graduate Institute (OGI) multi-lingual database [2]. The portion consisted of 210 naturally spoken utterances, each of which was approximately one minute in duration and recorded over a telephone. The utterances were hand-labeled and segmented into phonetic units.

The hand-labeled phonetic units were used as LDA class targets for the first set of the experiments; syllabic units were used for the second set. The syllabic units were derived from the phonetic units using a grouping algorithm. This algorithm automatically grouped sequences of phonemes into syllables using linguistic and acoustic-phonetic rules. It produced a great number of syllable units, most of which occurred relatively infrequently and hence were not useful for covariance estimation. These experiments therefore used only the 100 most frequent syllables for the syllabic experiments. Since the amount of usable speech data was effectively reduced to less than half, the covariance estimates for the syllabic experiments are less accurate.

We were interested in the effect of various noise conditions on the data-driven discriminant filter design. For this purpose we modified the raw speech with examples of additive and reverberant noise. For the additive noise experiments, we artificially added open-window car (Volvo) noise from NOISEX database [11] at 0 dB SNR. We separately added factory noise also from the NOISEX database at 10 dB SNR. Two examples of reverberation were added for the reverberation experiments. The first consisted of moderate reverberation whose impulse response was recorded in a variable echoic chamber with a reverberation time (T_{60}) of 0.6 seconds and a direct-to-reverberant ratio of -1.9 dB. The second consisted of more severe reverberation whose impulse was recorded in a basement hallway with a T_{60} of 1.7 seconds and a direct-to-reverberant ratio of -16 dB. The reverberation was artificially added to the speech by convolving it with the reverberation impulse response.

The data-driven filter design procedure involves analyzing relatively long time windows of log critical-band trajectories and is briefly summarized. The speech is transformed into critical-band power spectra (energy computed from a bark-scale spaced filter bank) followed by a logarithm as done in the first few

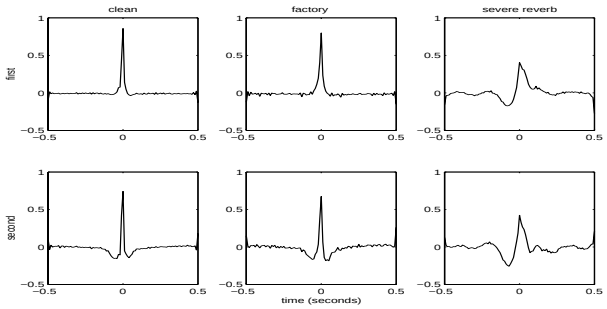

Figure 1: Impulse responses for the first two data-derived discriminant filters (row) for the clean, factory, and severe reverberation conditions (column). Filters were designed using the band centered at 1 kHz using phonetic classes.

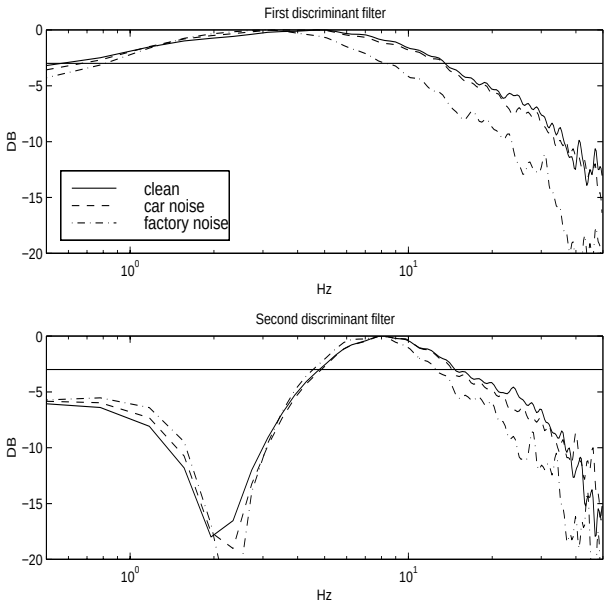

Figure 2: Frequency responses for the first and second discriminant filters derived under clean, car noise, and factory noise conditions using phonetic classes for the band centered at 1 kHz.

processing steps of RASTA-PLP [7] . LDA follows by capturing windows of approximately 1 second's worth of frames separately for each critical- band and subsequently assigning to it the linguistic unit class that falls on the center of this window. From these class bins of windowed trajectories, two quantities are computed: The within-class covariance S_W and the between-class covariance S_B. The principal discriminant filters are then taken as the eigenvectors of $S_W^{-1}S_B$ that have the largest associated eigenvalues [4]. As these eigenvectors are applied to temporal sequences, they are effectively FIR filters.

3. EXPERIMENTS WITH PHONETIC UNITS

In the first set of experiments, we derive discriminant filter using phonetic unit class targets. We do this for each of 5 acoustic conditions: clean (unmodified speech), added car noise, added factory noise, added moderate reverberation, and added severe reverberation. Figure 1 shows example impulse responses for the critical-band centered at 1 kHz and for the clean, added factory noise, and severe reverberation conditions. The first two components together capture over 85% of the variance. The form and frequency response of the filter is typical across bands and re-iterates the results from [10] in that responses are consistent with the RASTA filter and its derivatives. Further inspection across bands reveals that the general "Mexican hat" shape and form of the impulse responses remain with the addition of noise and reverberation. Since other critical-bands show similar results, only the 1 kHz band is included as an example.

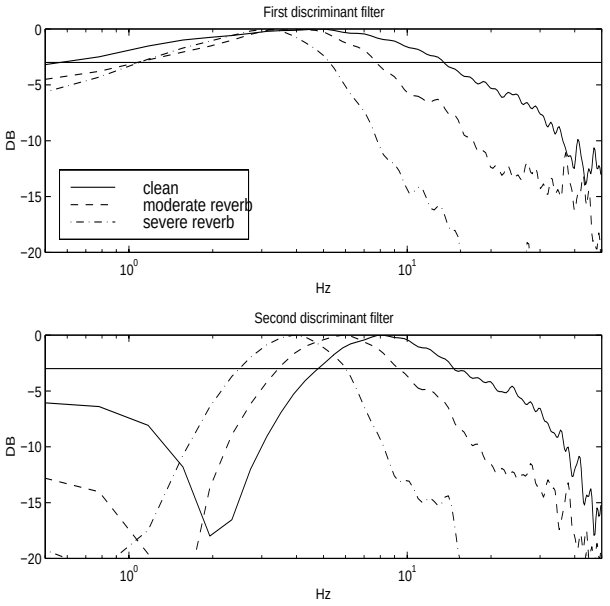

Figure 3: Frequency responses for the first and second discriminant filters derived under clean, moderate reverberation, and severe reverberation conditions using phonetic classes for the band centered at 1 kHz.

Plotted in figure 2 are example frequency responses of the first and second discriminant filters for the critical-band centered at 1 kHz. These frequency responses and all further ones include some frequency smoothing by the application of a hamming window to the impulse response. The hamming window, commonly applied when analyzing short-time segments, suppresses some of the artifacts at the edges of the filter while preserving the "action" of interest in the center. The plot includes filters derived under the clean, added car noise, and added factory noise conditions. Also included is a horizontal line marking the half power level to aide in observing the changes in the response.

We observe that the first discriminant component is roughly bandpass in nature, with some DC suppression and a high cut-off (-3 dB point) of approximately 13 Hz. With added noise the discriminant filters shows little variation but some tendency for a reduced bandwidth. With added car noise, the response almost matches the clean case whereas added factory noise causes some narrowing of the filter's high frequency cut-off to approximately 8Hz. The second component also shows little variability between the responses; only a slight narrowing of the bandwidth.

With the added reverberation, however, we see a more dramatic trend. Figure 3 shows the responses of the first and second principal data-derived filters for the clean, moderately reverberant, and more severely reverberant cases. When comparing the clean case to increasingly reverberant cases, the high frequency cut-off of the first discriminant filter reduces more substantially, going from 13 Hz down to about 5 Hz. Simultaneously there is in-

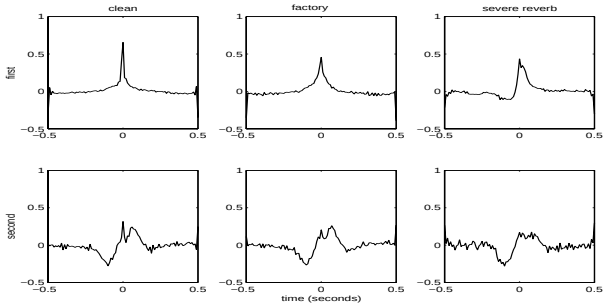

Figure 4: Impulse responses for the first two data-derived discriminant filters (row) for the clean, factory, and severe reverberation conditions (column). Filters were designed using the band centered at 1 kHz using syllabic classes.

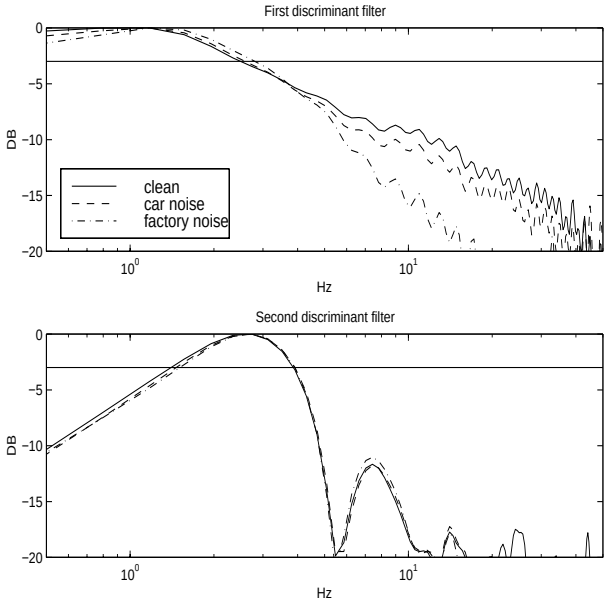

Figure 5: Frequency responses for the first and second discriminant filters derived under clean, car noise, and factory noise conditions using syllabic classes for the band centered at 1 kHz.

creased DC suppression. The second discriminant filter displays not only the reduction in bandwidth but also a shifting towards the lower frequencies.

4. EXPERIMENTS WITH SYLLABIC UNITS

We repeat the analysis in the previous section using syllabic targets. Figure 4 shows an example of the resulting filter impulse responses again using the band centered at 1 kHz and under the clean, added factory noise, and severely reverberated conditions. As with the phonetic case, we again observe the "Mexican hat" response property but with some notable differences: The first component does not exhibit much, if any, DC suppression amounting to more of a low-pass filter or local-averaging structure. We see this change some in the noisy cases. We also see a broader time response; this is expected as the syllable classes usually have longer durations than simple phones. Indeed, the syllable units as constructed will have durations that are dependent on the number of constituent phones. A number of the syllable units consist of single phones, particularly vowels, diphthongs, and some liquids. More common are syllables with a consonant-vowel structure.

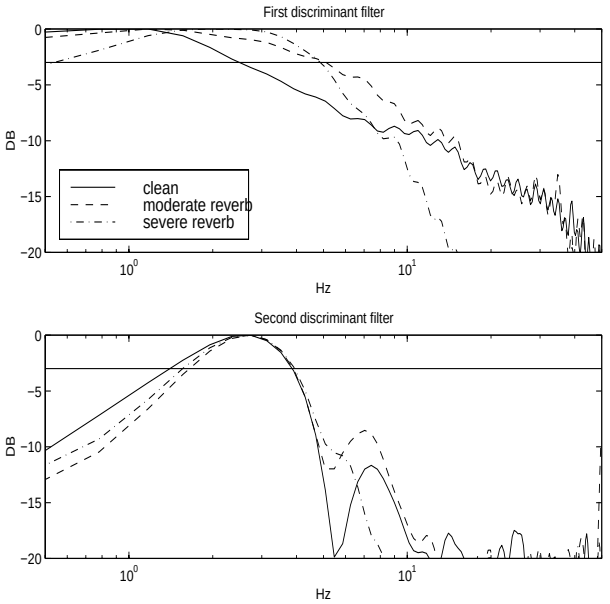

Figure 6: Frequency responses for the first and second discriminant filters derived under clean, moderate reverberation, and severe reverberation conditions using syllabic classes for the band centered at 1 kHz.

Figure 5 plots the frequency responses of the first and second components using the clean, added car noise, and added factory noise cases. We observe frequency responses favoring the lower frequencies, in step with the broader impulse responses. We see almost no change between the frequency responses of the filters derived under these different added noise conditions.

Figure 6 similarly shows the frequency responses of the data-derived filters under reverberant conditions. For the first component, we see that addition of moderate reverberation increases the bandwidth to about 5 Hz. Interestingly, increasing the severity of the reverberation appears to increase the DC suppression but does not further increase the high cut-off frequency. The second component frequency responses show a stability across the reverberant conditions, keeping a range between 1 and 4 Hz.

5. DISCUSSION

When adding simulated channel noise to the speech signal, previous work noted increased DC suppression in the derived filter responses [10, 1]. Since reverberation can also be considered a form of convolutional noise we expected to see similar DC suppression in the responses here. Some DC suppression, thought light, was indeed noticed not only with the reverberant conditions using phonetic classes but also with additive noise and with syllabic classes. Since DC and extremely low frequencies are thought to contain little or no speech information, it appears that suppressing the DC component can prudently eliminate a source of non-discriminant variability. This suppression applies equally to high modulation frequencies. Another observation was a tendency for the data-derived filters to become more band-limited with corrupting environments. Further, there is a tendency for them to gravitate towards the lower frequencies above 1 Hz.

The amount of change appears much more pronounced with reverberation than added noise in this log-spectral domain. With increasingly severe reverberation, the frequency responses using phonetic units changes from favoring frequencies up to 13 Hz to favoring lower frequencies between 1 and 5 Hz. This is also the frequency range that the filters using syllabic targets stabilize to. Since the syllabic targets inherently favor these low frequencies, the filters using syllabic classes do not change very much. This is consistent with analysis on the robustness and invariance of syllables and syllabic rates to this type of acoustic corruption and supports the favorable results seen when using preprocessing that operates within these frequencies and on syllable-length intervals, for example the Modulation-filtered Spectrogram features developed by Kingsbury [6, 8].

With added noise, the syllable-based filters exhibit virtually no change in frequency range, further supporting the stability of syllabic rates. Filters using phonetic targets exhibit some bandwidth narrowing though less pronounced than with reverberant conditions. The responses with phonetic targets still seem to favor frequencies up to 13 Hz in the added noise condition. Added noise and reverberation affect the spectro-temporal realization of the acoustic signal in different ways. Added noise offsets the power spectra on a frame by frame basis relatively independent of the speech. It appears that discarding the frequency content outside the ranges where the speech modulates is a prudent way to reduce the variability caused by the noisy condition. In contrast, reverberation "smears" the spectral energy forward in time and is therefore speech dependent. The discriminant information in this case appears to maintain integrity at the syllabic rates between 1 and 5 Hz. These are consistent with findings by Drullman in [3] where he noted that modulation frequencies below 4 Hz appear vital to speech intelligibility while frequencies above 16 Hz appear marginal.

The differences in the behavior of the discriminant filters under the noise and reverberant conditions suggest that a robust front-end to an automatic speech recognition (ASR) should contain analyses on multiple modulation frequency ranges, for example from 1 to 5 Hz and 1 to 13 Hz. The front-end may alternatively use complementary ranges (e.g. below and above 5Hz) or have some mechanism for adapting the range of modulation frequencies operated on. The relative stability of the syllable-based filters also suggests that employing syllabic units may benefit ASR as well. Experiments with dual preprocessing strategies that each operate on different modulation frequency ranges and additionally include both phonetic and syllabic linguistic units have already attested to the benefit of this strategy, for example in [12].

6. CONCLUSION

Linear Discriminant Analysis has proved to be a powerful technique in statistical data analysis. As applied here, LDA automatically generates an ordered set of discriminant linear filters that operate on the log-spectral energies. By observing the changes in these discriminant filters under a variety of acoustic conditions, we have observed some potentially useful trends in the location of temporally discriminant information in speech. In particular, analysis of syllable length intervals in the frequency range between 1 and 5 Hz appears to harbor robust discriminative information in the presence of reverberation. With additive noise, frequencies of up to 13 Hz may still contain useful discriminant information. Further experimentation with other noise conditions and recognition tests will be conducted to verify this. The experiments here support earlier preprocessing decisions based on psycho-acoustical insights and empirical testing. In the absence of explicit knowledge of the acoustic environment it seems prudent for preprocessing schemes to have the ability to either adjust their frequency range of interest or analyze on multiple modulation frequency ranges.

7. ACKNOWLEDGMENTS

We would like to graciously thank Hynek Hermansky and Narendranath Malayath at the Oregon Graduate Institute for their helpful advice and comments. We would also like to thank Nelson Morgan and ICSI for supporting this work.

8. REFERENCES

[1] 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.

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

[3] 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.

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

[5] S. Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Workshop (ETRW) on The Auditory Basis of Speech Perception*, pages 1–8, Keele, United Kingdom, July 1996. ESCA.

[6] 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.

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

[8] 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.

[9] D. W. Massaro. Perceptual units in speech recognition. *Journal of Experimental Psychology*, 102(2):199–208, 1974.

[10] 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.

[11] A. Varga and H. J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to dtudy the effect of additive noise on speech recognition systems. *Speech Communication*, 12:247–251, 1993.

[12] S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *ICASSP*, volume 2, pages 721–4, Seattle, Washington, May 1998. IEEE.