SPEECH ENHANCEMENT BASED ON TEMPORAL PROCESSING

Hynek Hermansky, Eric A. Wan, and Carlos Avendano

Oregon Graduate Institute of Science & Technology
Department of Electrical Engineering and Applied Physics
P.O. Box 91000, Portland, OR 97291

ABSTRACT

Finite Impulse Response (FIR) Wiener-like filters are applied to time trajectories of cubic-root compressed short-term power spectrum of noisy speech recorded over cellular telephone communications. Informal listenings indicate that the technique brings a noticeable improvement to the quality of processed noisy speech while not causing any significant degradation to clean speech. Alternative filter structures are being investigated as well as other potential applications in cellular channel compensation and narrowband to wideband speech mapping.

1. INTRODUCTION

The need for enhancement of noisy speech in telecommunications increases with the spread of cellular telephony. Calls may originate from rather noisy environments such as moving cars or crowded public places. The corrupting noise is often relatively stationary, or at least changing rather slowly. This leads us to investigate the use of RelAtive SpecTrAl (RASTA) processing of speech (Hermansky and Morgan, *et. al.*, 1991, 1994, Hirsch *et. al.*, 1991), which was originally designed to alleviate effects of convolutional and additive noise in automatic speech recognition (ASR). RASTA does this by band-pass filtering time trajectories of parametric representations of speech in a domain in which the disturbing noisy components are additive. Recently, RASTA was also applied to direct enhancement of noisy speech (Hermansky, Morgan, and Hirsch, 1993). In that case, RASTA filtering was applied to a magnitude (or cubic-root compressed power) of the short-term spectrum of speech while keeping the phase of the original noisy speech. However, applying rather aggressive fixed ARMA RASTA filters (designed for suppression of convolutional distortions in ASR) yields results similar to spectral subtraction (see *e.g.*, Boll (1979)), *i.e.*, enhanced speech often contains musical noise and the technique typically degrades clean speech.

2. CURRENT TECHNIQUE

RASTA involves nonlinear filtering of the trajectory of the short-term power spectrum (estimated using a 256 point FFT applied with a 64 point overlap Hamming window at a 8kHz sampling rate). Currently, we use linear filters applied to the cubic-root of the estimated power spectrum (Hermansky *et. al.* (1994)). In this study, we have substituted the ad hoc designed and fixed RASTA filters by a bank of non-casual FIR Wiener-like filters. Each filter is designed to optimally map a time window of the noisy speech spectrum of a specific frequency to a single estimate of the short-term magnitude spectrum of clean speech. For a 256 point FFT we require 129 unique filters.

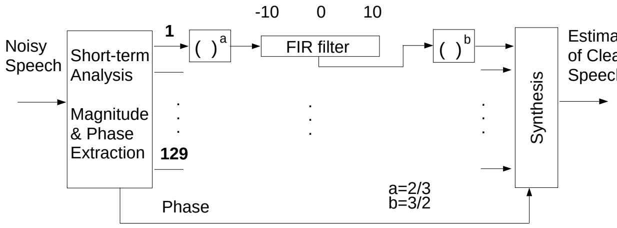

Figure 1: *Block Diagram of System*

Let $p_i^n(k)$ be the cubic-root estimate of the short-term power spectrum of noisy speech in frequency bin i ($i = 1$ to 129 and k corresponds to an 8ms step). The output of each filter is the following:

$$\hat{p}_i(k) = \sum_{j=-M}^{M} w_i(j) p_i^n(k-j), \qquad (1)$$

where $\hat{p}_i(k)$ is the estimate of the clean speech cubic-root spectrum. FIR filter coefficients $w_i(j)$ are found such that $\hat{p}_i$ is the least squares estimate of the clean

This work was sponsored in part by USWEST Advanced Technologies under contract number 9069-111

signal p_i for each frequency bin i. The current design has $M = 10$ corresponding to 21 tap noncausal filters.

As in the spectral subtraction technique, any negative spectral values after RASTA filtering are substituted by zeros and the noisy speech phase is used for the synthesis. The technique is illustrated in Fig. 1.

Initial filters were designed on approximately 2 minutes of speech of one male talker recorded at 8 kHz sampling over a public analog cellular line from a relatively quiet laboratory. The speech was artificially corrupted by additive noise recorded over a second cellular channel from a) car driving on a freeway with a window closed, b) car driving on a freeway with a window open, and c) busy shopping mall.

3. RESULTING OPTIMAL RASTA FILTERS

Filter frequency responses are shown in Fig.2 (darker shades represent larger values). The filters are approximately symmetric (*i.e.*, near zero phase). Filters for different frequency channels differ. The whole frequency band between 0 and 4 kHz appears to be subdivided into several regions, each characterized by its own RASTA processing.

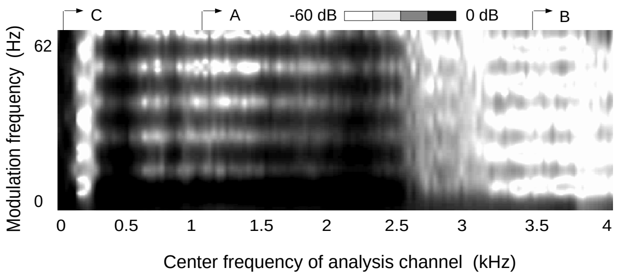

Figure 2: *Frequency Response of RASTA Filters*

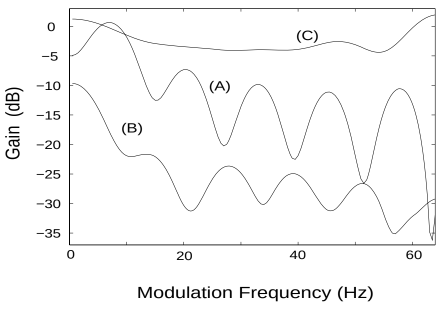

Figure 3: *Freq. Response of Different Filters.*

The highest gain RASTA filters are applied in the frequency bands of about 300-2300 Hz. Frequency response of typical filters in this region (slice A in Fig. 2) are shown in Fig. 3(A). Typically, filters in this frequency band have a band-pass character, emphasizing modulation frequencies around 6-8 Hz. Comparing to our original ad hoc designed RASTA filter (Hermansky, Morgan, and Hirsch, 1993), the low frequency band-stop is much milder, being only at most 10 dB down from the maximum.

Filters for very low frequencies (0-100 Hz) are high-gain filters with a rather flat frequency response (slice C in Fig. 2 and Fig. 3(C)). Filters in the 150-250 Hz and the 2700-4000 Hz regions are low-gain low-pass filters (slice B in Fig.2) with at least 10 dB attenuation for modulation frequencies above 5 Hz. The low frequency pass-band of these filters are typically below the pass-band of the high-gain band-pass filters of Fig. 3(A).

The frequency response of a 129 tap Wiener filter designed on noisy data is shown by the solid line in Fig. 4. For comparison, the *center* taps of the RASTA filters designed on magnitude spectrum (*i.e. a=b=1 in Fig.1*) are shown in the figure by the dashed line. Informal comparisons between the quality of speech processed by the time-domain Wiener filter and the RASTA filters indicates some advantage of the additional filter taps and compressed spectrum used with the new RASTA filters.

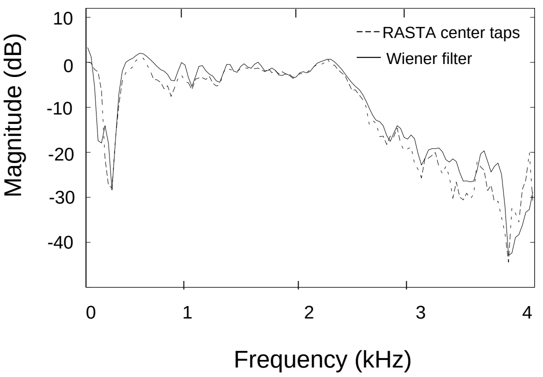

Figure 4: *Wiener filter response and RASTA center taps.*

4. EXPERIMENTS

The RASTA processing using filters described above has been tested on several recordings of actual noisy cellular communications (out of training set). We have informally observed that the processing brings a noticeable improvement in subjective quality of the recordings. Noise seems to be less disturbing while the quality of speech is not impaired. Further, we have also observed that the processing does not seem to impair the quality of clean speech recordings. A visible reduction of the noise after processing is apparent from the spectrograms shown in Fig. 5.

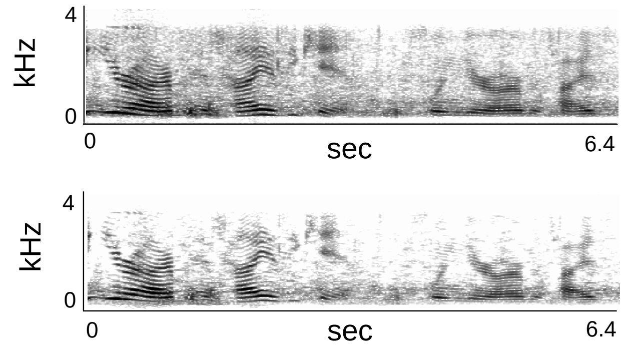

Figure 5: *Spectrogram showing noisy speech followed by the same noisy segment after processing.*

5. WORK IN PROGRESS

5.1. Data Collection

A database that includes the effects of actual ambient noise on cellular speech communications is being collected. A training set of 330 sentences taken from the TIMIT database (complete dr8 set) is played through a portable PC over a public cellular line from different locations. A test set generated with 123 randomly chosen TIMIT sentences (dr1-dr7 only) is recorded under the same conditions for later evaluation of our algorithms.

5.2. Adjacent Channels

Adjacent channel information has been used to train each channel resulting in a set of (Multi-Input-Single-Output) MISO filters. Initial results with 6 adjacent channels (3 at each side of the frequency of interest) show an improvement in the reduction of perceived noise as well as reduction of residual error at each frequency bin. However, the additional improvement obtained with this structure is noticeable only within the training set (original 2 minute source), while the processing generalizes poorly over different speakers and noise types. The effect of training on large data sets and the effect of different architectures need to be investigated before any conclusions can be drawn.

5.3. Cellular to Wideband Speech Recovery

Availability of the original 16kHz sampled speech data has raised the question as to the possibility of recovering a good quality wideband signal from a degraded, 8khz sampled, cellular speech recording. Two steps have been taken towards the investigation of this problem.

5.3.1. Cellular Channel Equalization

The original 16kHz sampled speech is downsampled to 8kHz and used as the desired response in our training scheme. The input data corresponds to the same speech segments recorded over the cellular channel. Additional white Gaussian noise is added to the input in order to avoid excessive gains in the band edges where cellular speech is not present. Several tests done on clean cellular communications have resulted in a noticeable improvement; however, formal comparisons with other techniques have yet to be done.

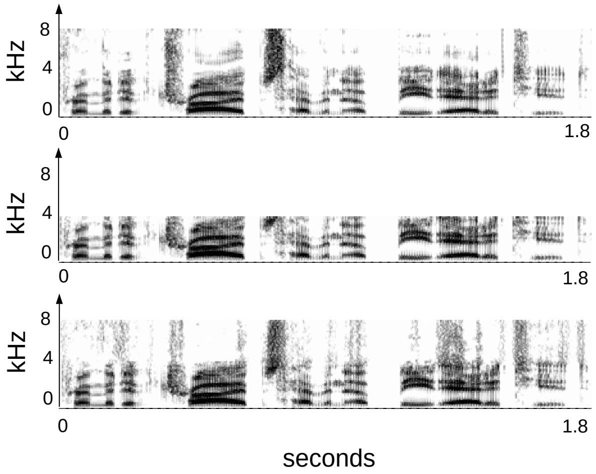

Figure 6: *a) Spectrogram of the original speech, b) spectrogram of downsampled signal and c) recovered spectrogram.*

5.3.2. Wideband Speech Reconstruction

An approximate wideband from narrowband speech map has been produced by extending the adjacent channel concept. Now the 129 trajectories derived from the 8kHz sampled speech analysis are fed to MISO filters whose outputs correspond to frequencies from 4kHz to 8kHz (*i.e.*, bins 130 to 257) and the signal reconstruction is done at twice the original rate. Training was performed with 9 TIMIT sentences (same speaker) with a time window of 40 ms ($M = 2$ in (1)). In Fig.6, we show an out of training set original wideband speech spectrogram, the downsampled version of it, and the recovered spectrogram. Within a given talker the high frequency recovery appears to be viable. However, for time signal reconstruction an estimate of the phase of the recovered high frequencies needs to be performed. This is discussed in the following section.

5.4. Phase Reconstruction

Modification of the short-term magnitude spectrum of a signal and synthesis with original phase do not, in general, yield a time signal with the desired magnitude spectrum (see *e.g.*, Allen, J.(1977)). To avoid distortion in the synthesized signal, especially when the power spectrum has been severely modified (see subsection 5.6) , we are investigating an iterative algorithm (see *e.g.*, Griffin and Lim (1984)) in which the mean squared error between the desired spectrum and the spectrum produced by the synthesized signal is minimized. In noise reduction applications, we have used the noisy signal's phase to perform the initial step in the reconstruction. In the case of wideband speech mapping, a linear map from the available low frequency phase components to the higher frequencies is taken as a first approximation.

5.5. Postprocessing with Envelope Enhancement

First proposed for enhancement of ADPCM speech coding (see Ramamoorthy, *et. al.*, (1988)), a prefiltering technique for spectral envelope enhancement has been applied in our system. Spectral domain filtering is done by designing a filter at each 8 ms step with an untilted and smoothed 8th order all-pole model spectrum of the particular frame, giving the effect of enhancing frequencies around formants and suppression elsewhere. Best results have been achieved by prefiltering the compressed power spectrum prior to applying RASTA filters to its trajectories.

5.6. Non-linear Filters

Non-linear RASTA filters have also been implemented with 3 layer artificial neural networks. Preliminary results on small training sets show a more aggressive filtering and greater noise reduction. However, musical-like residual noise as well as greater degradation of clean speech is also observed. The iterative synthesis method described above has been used with some success to alleviate the phase distortion introduced by the large amount of spectral modification. Current efforts are focused on reducing the adverse effect introduced by the networks.

6. CONCLUSIONS

We have presented a new technique for enhancement of noisy speech in cellular communications, which utilizes Wiener-like RASTA filters on cubic-root compressed short-term spectrum of speech. While no formal subjective perceptual tests were carried out, our initial experience indicates a possible advantage of the proposed technique over conventional speech enhancement techniques. Future work will investigate generalization of the technique for additional noise sources and larger training sets. In addition, more formal perceptual evaluations will be undertaken. Related work outlined in section 5 will also be pursued.

7. ACKNOWLEDGEMENTS

We would like to thank Mahesh K. Kumashikar for the neural networks training, and CONACYT (Mexico) for partial support of the third author.

8. REFERENCES

[1] J.B. Allen: Short term spectral analysis, synthesis, and modification by discrete Fourier transform, *IEEE ASSP-25*, pp. 235-238, Jun. 1977.

[2] S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, *IEEE ASSP-27*, pp. 113-120, Apr. 1979.

[3] V. Ramamoorthy, N.S. Jayant, R.V. Cox and M.M. Sondhi: Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback, *IEEE J. Selec. Areas in Comm. Vol. 6, No.2*, pp. 364-382, Feb. 1988.

[4] H. Hermansky: Perceptual predictive (PLP) analysis of speech, *J. Acoust. Soc. Am. 87* pp. 1738-1752 Apr. 1990.

[5] H. G. Hirsch, P. Meyer, and H. Ruehl: Improved speech recognition using high-pass filtering of subband envelopes, *Proc. EUROSPEECH '91*, pp. 413-416, Genova, 1991

[6] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn: Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP) *Proc. EUROSPEECH '91*, pp. 1367-1370, Genova, 1991.

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

[8] H. Hermansky, E. Wan, and C. Avendano: Noise suppression in cellular communications, *Proc. IVTTA 94*, pp. 85-88, Kyoto 1994.