Study on the Dereverberation of Speech Based on Temporal Envelope Filtering

Carlos Avendano and Hynek Hermansky

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

ABSTRACT

In this paper we explore speech dereverberation techniques whose principle is the recovery of the envelope modulations of the original (anechoic) speech [3],[4]. Based on our previous experience with such kind of processing for additive noise reduction applications, we apply a data designed filterbank technique [2] to the reverberant speech. Comparing our results with other works we discuss the effectiveness and limitations of this type of approaches.

1. Introduction

Recent advances in teleconferencing, multi-media and hands-free telephony have spurred an interest in reducing the effect of room reverberation in speech communications. In the past, the problem has been approached from several different perspectives depending on the particular application. While several viable solutions exist in situations where more than one channel is available [8],[6], single channel systems still pose a formidable challenge.

For single channel systems two main approaches have been taken. One approach consists of estimating some properties of the room from the corrupted data and applying deconvolution techniques to recover the speech. The main drawback in such cases is that some simplifying assumptions about the speech and channel have to be made, thus yielding only partial solutions [11],[13].

In a different approach, an attempt is made to recover the energy envelope of the original (anechoic) speech by applying a theoretically derived inverse modulation transfer function (MTF) [3], or ad hoc high-pass filtering [4]. Such approaches were motivated by studies on the the effect of reverberation on the Modulation Index (MI) of speech and the reduction of intelligibility in reverberant environments [14].

In other works [5], the envelope modification has even been attempted with filters derived from pure observations, which further simplify the reverberation phenomena.

1.1. Motivation

In this study we apply a technique which we originally developed for background noise reduction purposes [2], to the dereverberation problem.

We have achieved a considerable reduction of additive noise by filtering compressed short-time power spectrum trajectories (STPT's) of noisy speech. The filters used were derived from parallel recordings of clean and noisy training data. The magnitude frequency response of such filters showed that, in the presence of additive noise, modulation frequencies characteristic of clean speech (around 3-6 Hz) need to be enhanced and otherwise attenuated, thus supporting [14] in their observation about the noise effects on the MI of speech. This gave us some confidence that the data-derived filters are partially compensating for the deteriorating effect of some disturbances on the MI of speech.

Recently, we have also applied our technique (with some modifications) to the design of optimal channel normalization filters for applications in ASR [1], and found that that the enhancement of dominant speech-related modulation frequencies (in this application in the log spectral domain) is required to alleviate the influence of channel distortions in speech data.

Since the compensation for the effect of reverberation on the MI of speech was the prime motivation behind both Langhan's-Strube's and Hirsch's techniques, it is of interest how would our data-derived filters look and perform on this dereverberation task.

2. Effects of Reverberation on Speech

2.1. Envelope Smearing

As speech is produced inside an enclosure, the finer details of its time-intensity distribution are blurred before reaching the listener. This modification results from the superposition of the reflected sound waves with different delays and intensities to the original (direct path) waveform [14].

In the absence of discrete echoes, the effect of such a super-

position shows as reverberation tails on the energy envelope of the signal. These tails have an approximately exponentially decaying envelope with a time constant determined by the room's dimensions, wall reflectivities and the positions of the source and receiver.

Tails produced by past acoustic events fill in low energy regions between consecutive sounds reducing the modulation depth of the original envelope and thus modifying its MI [14]. The MTF of the reverberant room can be derived from it's impulse response [9] and thus the effect of the room on the MTF of speech is predictable. This motivates the application of inverse MTF's to recover the original modulations present in the original (anechoic) speech. Notice that in this case the phase modifications in the fine structure are not considered.

2.2. Linear System

Reverberation can be formally described in terms of a linear system characterized by the impulse response of the room. Reverberant speech is then modelled as the convolution of speech with this impulse response. The effective length of this impulse response can be very long, in fact, most of the times it is longer than the interval over which the speech signal can be considered as stationary.

Deconvolution techniques take advantage of this model and are more effective when some knowledge about the room can be obtained as side information. In situations where several channels and/or a reference signal are available, an estimate of the room's characteristics can be derived [12], [10] and some kind of inverse filtering can be applied to recover the original speech.

Until recently, the inverse filtering approaches have had as main problem the non-minimum phase nature of the reverberation process [10], however, with the availability of multiple microphones exact inverse filtering has been achieved in [7].

3. Review of Data Designed Filter Bank Technique

In this section we review the technique used to derive a filter bank from clean and reverberant speech data.

The technique used in [2] consists of the following steps: the short-term frequency magnitude of the corrupted speech is computed using the short-time Fourier transform (STFT). After application of a static non-linearity, each time trajectory of this new representation is filtered by a data designed filter (see Fig.1). After filtering, the result is transformed back to the STFT magnitude domain and combined with the original short-time phase in an overlap-add (OLA) synthesis procedure.

Since in the particular case of our study we are after compensation of the short-term power spectrum, we use $a = 2$, i.e. we intent to perform the linear filtering on the Short-Term Power Spectrum Trajectories (STPT's) of speech.

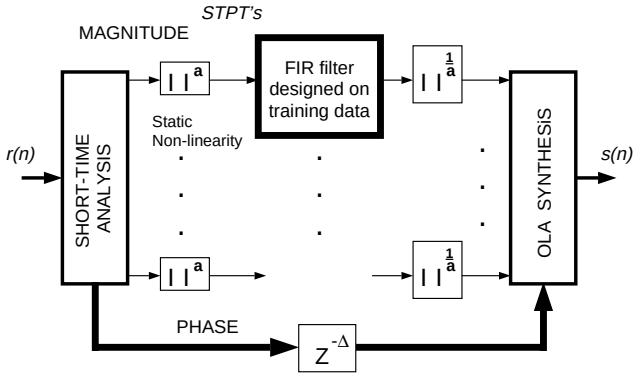

Figure 1: *Block diagram of the technique. r(n) is the corrupted speech and s(n) is the estimate of the clean speech*

3.1. Filter Design

The filters are finite impulse response filters derived by solving the Wiener-Hopf equation for the minimization of the Euclidean distance between the filtered STPT's of the corrupted speech and the corresponding desired trajectories of clean speech. A filter is designed for each frequency channel. For the compensation of the effects of additive noise, filter lengths were typically chosen to be around a syllable length, i.e. about 200 ms.

For the experiments described in this paper, the data used were generated by convolving clean speech (sampled at 8kHz) with artificial room impulse responses. For these impulse responses, reverberation tails were produced using Schroeder's model (i.e. a decaying exponential envelope modulated by a white noise sequence).

4. Experiments

Using the data designed filter bank approach we find the modulation-domain filters and compare them with the theoretical and empirical transfer functions used in [3] and [4].

4.1. High-pass Filtering of STPT's

Hirsch [4] reported an improvement on the recognition of reverberant speech by using high-pass filtering of the STPT's. Improvement of the subjective quality of the reconstructed speech after filtering was also reported and this constitutes the object of our investigation.

We hypothesized that the main effect of the aggressive high-pass filtering was due to the fact that, after filtering, a considerable number of points in the resulting power spectra were eliminated. The filter in [4] was a high-pass filter with a real zero at dc, thus removing the mean of the STPT's and

effectively making half of the spectral power spectra energies negative. A standard technique for handling the negative values in the overlap-add spectral subtraction is to substitute them by zeros, i.e. effectively eliminating them. After filtering, negative values correspond to low energy regions likely to contain reverberation tails. While removal of low spectral energy values reduces the reverberation effects considerably, it may also cause a loss of useful speech information and distortion of the perceived speech signal.

We performed two simple experiments to test our hypothesis. In the first experiment we applied full-wave rather than half-wave rectification and found *no* reduction of the reverberation.

In the second experiment the STPT's of reverberant speech were center clipped below a certain threshold (20% of the maximum value). The result was very similar to that obtained with high-pass filtering.

Although some improvement of the MI is evident after the processing using Hirsch's filter, it appears that the main effect of the high-pass filtering technique is in removing the low-energy spectral values, rather than achieving a restoration of the MI.

4.2. Inverting a Theoretical MTF

Langhans and Strube ([3]) applied a theoretically derived inverse MTF to reduce reverberation. In their method the inverse MTF (IMTF) was applied in critical bands simulated by a weighted sum of the STPT's. The IMTF used was the inverse of a first order low-pass characteristic with a cut-off frequency proportional to the reverberation time (T_{60}) considered (this is analytically derived for artificial reverberation). The modulation frequencies higher than 10 Hz were not enhanced above certain threshold to avoid strong fast fluctuations, and those higher than 40 Hz were further attenuated.

Since the results obtained with this technique were reported as negative, we decided to investigate the matter and use our filter design technique to compare the theoretical curve to the magnitude transfer function of the data-derived filters.

4.3. Data-derived Filters

A set of filters was obtained using artificially reverberated speech in the way described in section 3. In this experiment we used critical band energies of corrupted and clean speech. These energies were produced by a weighted sum over the STPT's in one third octave rectangular windows. This smoothing was only necessary to compare our results to those described in [3]. The reverberation time used to obtain the results reported here was $T_{60} = 0.75$ but we got similar results under other conditions. Filter lengths were chosen to be at least as long as the reverberation time considered.

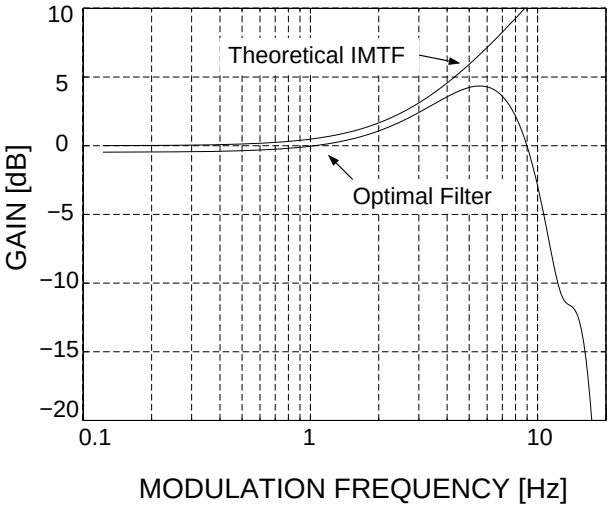

Figure 2: *Magnitude frequency response of a data-derived filter (at 1kHz center frequency band) compared to the theoretical curve.*

Fig. 2 shows the filter for the critical band with center frequency at about 1kHz. The filters for other critical bands have very similar shapes. At low modulation frequencies we can see a close correspondence between the data-derived and theoretical frequency responses. At higher modulation frequencies, however, the filter characteristics differ significantly: the data-derived filters exhibit a strong low-pass character, suppressing modulation frequencies above 10 Hz.

We have observed such suppression of higher modulation frequencies in many of our experiments with noisy [2] and linearly distorted [1] speech. It appears that high modulation frequencies of the short-term spectrum of speech are highly corruptible by many kinds of distortions, and the data-derived filters always tend to alleviate this. (Notice that Langhans and Strube elected at least not to enhance such higher modulation frequencies in their attempt for the dereverberation of speech in spite of the fact that their theoretically derived compensation filter suggested to do so).

4.4. Results

After processing the smoothed STPT's with the data-derived filters we resynthesized the signal by applying the new envelope to the original short-term phase. We observed that a reduction of the reverberation was audible. However, no improvement over using the theoretical IMTF was apparent. A reasonable question to ask in this case is if the filtering really compensated for the reduction of the modulation index of the corrupted speech.

Fig. 3 shows the modulation index of the 1kHz centered bands of original speech, reverberant speech, and recon-

structed speech after the modulation-domain filtering by the data-derived filters. Restoration of the modulation frequencies which were suppressed by dereverberation is significant in the reconstructed speech and was consistently found also at other frequency bands.

Unfortunately, we must conclude that even when the desired restoration of suppressed modulation frequencies is achieved, the recovery of the modulations alone does not guarantee a reverberation free speech. In this scheme the resynthesis procedure makes use of the original corrupted phase which undoubtedly contributes to the perceived artifacts.

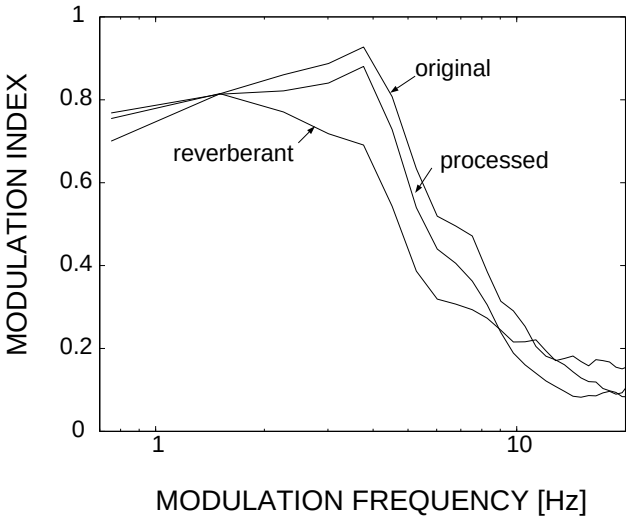

Figure 3: *Modulation index at 1kHz for clean speech, reverberant speech and processed speech.*

5. Conclusions

We have obtained a filter-bank from training data for processing the simulated critical bands of reverberant speech. Results show that these filters approximate some of the characteristics of the theoretical transfer functions used in the past. Listening tests indicate that an audible reduction of reverberation is achieved but artifacts on the processed speech signal are somewhat severe. Although no formal intelligibility tests were carried out it is likely that the restoration of the MI might not improve intelligibility in this case. This result indicates that caution needs to be taken when the MI is used as an indicator of quality or intelligibility of the impaired speech.

6. Acknowledgments

The authors would like to thank USWEST Advanced Technologies (9069-111), DOD (MDA 904-94-C-6196), NSF/ARPA (IRI-9314959), CONACYT, and the member companies of CSLU for their support.

7. REFERENCES

1. Carlos Avendano, Sarel van Vuuren and Hynek Hermansky, "Data Based Filter Design for RASTA-like Channel Normalization in ASR", to appear in *Proc. ICSLP-96*, Philadelphia, PA, October 1996.
2. Hynek Hermansky, Eric A. Wan and Carlos Avendano, "Speech Enhancement Based on Temporal Processing", *Proceedings IEEE ICASSP-95*, pp. 405-408, Detroit, MI, May 1995.
3. Thomas Langhans and Werner Strube, "Speech Enhancement by Nonlinear Multiband Envelope Filtering", *Proceedings IEEE ICASSP-82*, pp. 156-159, 1982.
4. H. G. Hirsch, "Automatic Speech Recognition in Rooms", *Signal Processing IV: Theories and Applications.* Lacourne, Chehikian, Martin and Malbos (editors), Elsevier Science Publishers B.V. (North Holland), EURASIP, pp. 1177-1180, 1988.
5. J. Mourjopoulos and J. K. Hammond, "Modelling and Enhancement of Reverberant Speech Using an Envelope Convolution Method", *Proceedings IEEE ICASSP-83*, pp. 1144-1147, Boston MA, 1983.
6. Hong Wang and Fumitada Itakura, "An Approach of Dereverberation Using Multi-microphone Sub-band Envelope Estimation,", *Proceedings IEEE ICASSP-91*, pp. 953-956, Toronto , Canada 1991.
7. Masato Miyoshi and Yutaka Kaneda, "Inverse Filtering of Room Acoustics,", IEEE Trans. ASSP Vol.36 No. 2, pp. 145-152, February 1991.
8. J. B. Allen, D. A. Berkley, J. Blauert , "Multimicrophone Signal Processing Technique to Remove Room Reverberation of Speech Signals," J. Acoust. Soc. Am., vol.*62*, No.4, pp. 912-915, October 1977.
9. Manfred R. Schroeder, "Modulation Transfer Functions: Definition and Measurement," Acoustica, Vol. 49, pp. 179-182, 1981.
10. Mikio Tohyama, Richard H. Lyon and Tsunehiko Koike, "Pulse Waveform Recovery in a Reverberant Condition," J. Acoust. Soc. Am., vol.*91*, No.5, pp. 2805-2812, May 1992.
11. Mikio Tohyama, Richard H. Lyon and Tsunehiko Koike, "Source Waveform Recovery in a Reverberant Space by Cepstrum Dereverberation," *Proceedings IEEE ICASSP-93*, pp. I 157-160, 1993.
12. Athina P. Petropulu and Suresh Subramaniam, "Cepstrum Based Deconvolution for Speech Dereverberation," *Proceedings IEEE ICASSP-94*, pp. I 9-13, 1994.
13. Alex Stephenne and Benoit Champagne, "Cepstral Prefiltering for Time Delay Estimation in Reverberant Environments," *Proceedings IEEE ICASSP-95*, No.5, pp. 3055-3058, Detroit, MI, May 1995.
14. T. Houtgast and H. J. M. Steeneken, "A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria," J. Acoust. Soc. Am., vol.*77*, No.3, pp. 1069-1077, March 1985.