SPEECH ENHANCEMENT BASED ON FILTERING THE SPECTROTEMPORAL MODULATIONS

Nima Mesgarani, Shihab Shamma

Electrical and Computer Engineering Department
Institute for System Research
University of Maryland, College Park, MD 20742

ABSTRACT

A monaural noise suppression algorithm is proposed based on filtering the spectro-temporal modulations of noisy speech. The modulations are estimated from a multiscale representation of the signal spectrogram generated by a model of sound processing in the auditory system. A significant advantage of this method is its ability to suppress noise that has distinctive modulation patterns, despite being spectrally overlapping with the speech. The performance of the algorithm is evaluated using subjective and objective tests and compared to the Optimal Smoothing and Minimum Statistics approach (R. Martin 2001). The results demonstrate the efficacy of the spectro-temporal filtering approach in the conditions examined.

1. INTRODUCTION

Noise suppression to enhance speech quality or intelligibility is necessary in a wide range of applications including mobile communication, hearing aids and speech recognition. It has been an active research area for over fifty years, mostly framed as a statistical estimation problem in which the goal is to estimate speech from its sum with other independent processes (noise). This approach requires an underlying statistical model of the speech and noise, as well as an optimization criterion. In some of the earliest work, the speech waveform itself was estimated (summarized in [1]). When the distortion is expressed as a minimum mean square error, the problem reduces to the design of an optimum Wiener filter.

Estimation can also be done in the frequency domain, as is the case with such methods as spectral subtraction [1] and its derivatives such as the signal subspace approach [2] and the estimation of the short-term spectral magnitude [3]. Estimation in the frequency domain is superior to the time domain as it offers better initial separation of the speech from noise, which (1) results in easier implementation of optimal/heuristic approaches, (2) simplifies the statistical models because of the decorrelation of the spectral components, (3) facilitates integration of psychoacoustic models [4].

Recent psychoacoustic and physiological findings in mammalian auditory systems, however, suggest that the spectral decomposition is only the first stage of several interesting transformations in the representation of sound. Specifically, it is thought that neurons in the auditory cortex decompose the spectrogram further into its spectro-temporal modulation content [5]. This finding has inspired a multiscale model representation of speech modulations that has proven useful in assessment of speech intelligibility [9], discriminating speech from non-speech signals [13], and in accounting for a variety of psychoacoustic phenomena [10].

The focus of this article is an application of this model to the problem of speech enhancement. The rationale for this approach is the finding that modulations of noise and speech have a very different character, and hence they are well separated in this multiscale representation, more so than is normally the case at the level of the spectrogram. Modulation frequencies have been used in noise suppression before (e.g. [8]), however this study is different in several ways: (1) the proposed method is based on filtering not only the temporal modulations, but the joint spectro-temporal modulations of speech (2) Modulations are not used to obtain the weights of frequency channels, but the filtering itself is done in spectro-temporal modulation domain, and (3) the filtering is done only on slow temporal modulations of speech (below 32Hz) which are important for intelligibility.

A key computational component of this approach is an *invertible* auditory model which captures the essential auditory transformations from the early stages up to the cortex, and provides an algorithm for inverting the "filtered representation" back to an acoustic signal. Details of this model are available elsewhere, and hence only a brief summary is provided next.

2. AUDITORY MODEL

The auditory model was inspired by psychoacoustical and neurophysiological findings in the early and central stages of the auditory pathway. The early stage converts the sound waveform into an *auditory spectrogram* - roughly akin to a

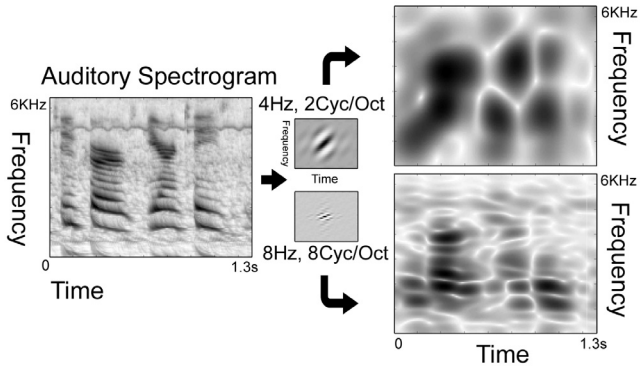

Fig. 1. Demonstration of the cortical processing stage of the auditory model. The auditory spectrogram (left) is decomposed into its spectro-temporal components using a bank of spectro-temporally selective filters. The impulse responses (spectro-temporal receptive fields or STRF) of two such filters are shown in the center panels. The upper filter is tuned to 4Hz temporal (rate) and 2 cycle/octave spectral (scale) modulation whereas the bottom one is tuned to faster rate (8Hz) and finer scale (8 cycle/octave). The multiresolution (cortical) representation is computed by (2-dimensional) convolution of the spectrogram with each STRF, generating a *family* of spectrograms with different spectral and temporal resolutions, i.e., the cortical representation is a 4-dimensional function of time, frequency, rate and scale. A complete set of STRFs guarantees an invertible map which is needed to reconstruct a spectrogram back from a modified cortical representation.

time-frequency distribution along a tonotopic (logarithmic frequency) axis [6]. The second (cortical) stage performs a two dimensional wavelet transform of the auditory spectrogram, thus providing an estimate of its spectral and temporal modulation content. It is computationally implemented by a bank of two-dimensional (spectro-temporal) filters that are selective to different spectro-temporal modulation parameters ranging from slow to fast $rates$ temporally, and from narrow to broad $scales$ spectrally. The spectro-temporal impulse responses (or "receptive fields") of these filters are centered at different frequencies along the tonotopic axis. Therefore, the basic mathematical formulation of the model can be summarized as followed:

$$y_{cochlea}(t,f) = s(t) * h_{cochlea}(t,f) \quad (1)$$

$$y_{an}(t,f) = g_{cochlea}(\partial_t y_{cochlea}(t,f)) * \mu_{haircell} \quad (2)$$

$$y(t,f) = max(\partial_f y_{an}(t,f), 0) * \mu_{midbrain} \quad (3)$$

$$r(t,f;\omega,\Omega,\theta,\phi) = \quad (4)$$
$$y(t,f) *_{tf} [h_{rate}(t;\omega,\theta).h_{scale}(f;\Omega,\phi)]$$

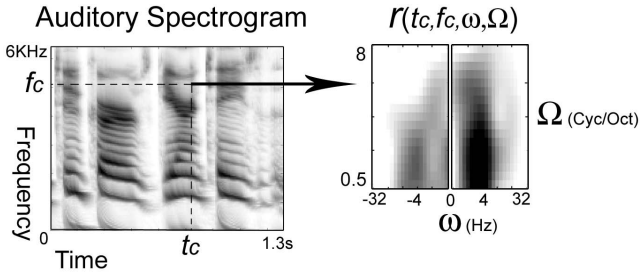

Fig. 2. Rate-Scale representation of clean speech. Spectro-temporal modulations of speech are estimated by a bank of modulation selective filters, and are depicted at a particular time instant and frequency (t_c and f_c) by the 2-dimensional distribution on the right.

where, $y_{cochlea}(t,f)$ is the cochlear filter output, $y_{an}(t,f)$ is auditory nerve patterns, $y(t,f)$ is the auditory spectrogram and $r(t,f;\omega,\Omega,\theta,\phi)$ is the rate-scale representation.

Since the cortical stage (Equation (4)) is linear and invertible, we can readily reconstruct the auditory spectrogram $y(t,f)$ from its modified rate-scale representation, $r(t,f;\omega,\Omega,\theta,\phi)$. The reconstruction of an audio waveform from the auditory spectrogram is more difficult to derive directly because of the two nonlinear functions, $g_{cochlea}$ and $max(.,0)$. Instead, an iterative method is used based on a convex projection algorithm proposed in [7]. The central stage processing is illustrated with an example in Figure 1.

2.1. Multiresolution representation of speech and noise

The multiresolution energy representation of sound is a 4-dimensional function of time (t), frequency (f), rate (ω) and scale (Ω). One can think of each point in the spectrogram as having a 2-dimensional rate-scale representation, $r(t_c,f_c,\omega,\Omega)$ which indicates the modulation strength at all ω's and Ω's for that channel and instant. Figure 2 illustrates the spectro-temporal modulation energy of clean speech at time t_c and frequency f_c. Temporal modulations in speech tend to concentrate near $4Hz$, while spectral modulations span a wide range reflecting at its high end the harmonic structure due to voicing (2-6 cycle/octave) and at its low end the spectral envelope or formants (less than 2 cycle/octave). Figure 3 illustrates the different modulations due to noise and speech in this representation. The speech signal has been corrupted by additive Buccaneer One noise from Noisex [11] database. The resulting rate-scale modulation at the same point as before (t_c and f_c) is shown in the top-center panel. In the bottom-center panel the modulations due to the *noise-only* signal are shown. Unlike speech, this noise has a strong temporally modulated energy (at a rate of 10 Hz) near f_c which overlaps with a more stationary speech spectrum. Noise and speech also differ substantially in their

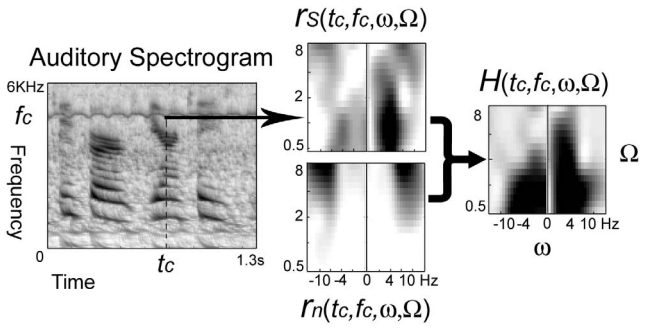

Fig. 3. Filtering the rate-scale representation: Modulations due to the noise are filtered out by weighting the rate-scale representation of noisy speech with the function, $H(t, f, \omega, \Omega)$. In this example, the Buccaneer One noise from Noisex was added to clean speech at SNR 10dB. The rate-scale representation of the signal, $r_s(t_c, f_c, \omega, \Omega)$ and the rate-scale representation of noise, $r_n(t_c, f_c, \omega, \Omega)$ were used to obtain the necessary weighting as a function of ω and Ω (equation 5). This weighting was applied to the rate-scale representation of the signal, $r_s(t_c, f_c, \omega, \Omega)$ to restore modulations typical of clean speech. The restored modulation coefficients were then used to reconstruct the cleaned auditory spectrogram, and from it the corresponding audio signal.

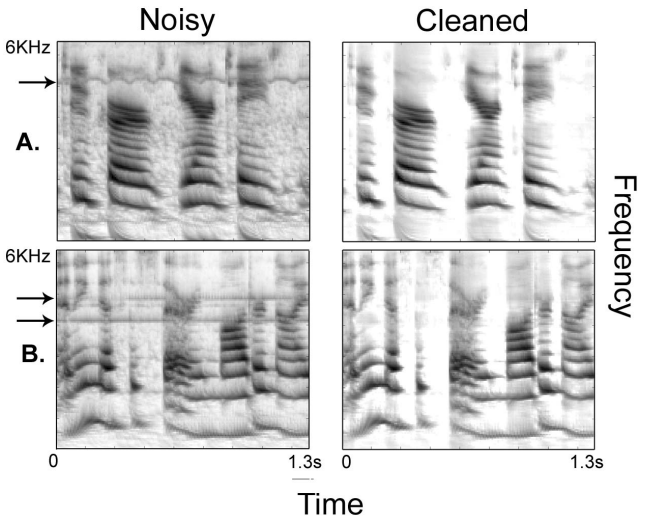

Fig. 4. Examples of restored spectrograms after "filtering" of spectro-temporal modulations. (A) (Top) Buccaneer One from Noisex and SI1347 (male) speech from TIMIT [12] added at SNR 15dB (left panel). (Right panel) The clean speech spectrum has been successfully restored although the noise has a strong temporally modulated tone (10 Hz) mixed in completely with the speech signal near 2 kHz (indicated by the arrow). (B) (Bottom panels) Destroyer Engine Room noise from Noisex and SI948 (female) speech from TIMIT added at SNR 15dB (Left panel). (Right panel) Speech spectrum has been restored even at frequencies (arrows) where the noise introduced significant modulations that have a totally different modulation character from the clean speech.

spectral modulation content because the noise at f_c is very narrow and hence spreads its energy up to relatively high scales ($6 - 8$ cycles/octave). (Figures 2 3).

2.2. Estimation of noise modulations

A crucial factor in affecting the performance of any noise suppression technique is the quality of the background noise estimation. In spectral subtraction algorithms, several techniques have been proposed that are based on three assumptions: (1) speech and noise are statistically independent, (2) speech is not always present and (3) the noise is more stationary than speech [4]. These methods include Voice Activity Detection (VAD), soft-decision method, and tracking of spectral minima [4]. We implemented two methods to perform this estimation: a VAD and an adaptive procedure. One of the common problems with VADs is their poor performance at low SNRs. To overcome this limitation, we employed a recently formulated speech detector (also based on the cortical representation) which detected speech reliably at SNR's as low as -5dB [13]. An alternate approach to VAD is to use an adaptive model to track and emphasize the salient modulations of speech and suppress irrelevant ones. The average spectro-temporal modulations of clean speech have proven to be a distinctive property that can be reliably used to detect speech and assess its intelligibility [13, 9].

In the remainder of this paper, we will focus on a modified VAD approach using [13] in which the spectro-temporal modulations of noise are estimated from a 500ms noise-only frame.

3. NOISE SUPPRESSION

The exact rule for suppressing noise coefficients is a determining factor in the subjective quality of the reconstructed enhanced speech, specially with regards to the reduction of musical noise [4]. Many techniques can be used to estimate the clean speech or noise coefficients, including linear approaches like Wiener Filter, as well as nonlinear methods. For this study, a generalized Wiener Filter was used as followed:

$$H(t, f, \omega, \Omega) = \left(\frac{SNR(t, f, \omega, \Omega)}{\alpha + SNR(t, f, \omega, \Omega)}\right)^{\beta} \qquad (5)$$

The above equation gives a gain factor for every frequency, rate and scale at each time instant (Figure 3) which then used to weight the spectro-temporal representation of noisy speech.

Noise type, SNR	*Degraded*	*OSMS[4]*	*STMF*
White, +15dB	1.6	2.5	3.3
Buccaneer, +15dB	1.1	2.4	3.2
Destroyer, +15dB	1.6	2.7	2.8
White, +5dB	1.0	2.0	2.3
Buccaneer, +5dB	0.7	2.0	1.9
Destroyer, +5dB	1.0	2.1	1.9

Table 1. Mean Opinion Score on a scale of 1 to 5 averaged over 10 subjects for three conditions: (1) Original noisy speech (2) enhanced speech using OSMS [4] and (3) Spectro-Temporal Modulation Filtering (STMF).

Noise type, SNR	*Degraded*	*OSMS [4]*	*STMF*
White, +15dB	2.696	2.857	2.962
Buccaneer, +15dB	2.625	2.670	2.725
Destroyer, +15dB	2.684	2.848	3.023
White, +5dB	1.972	2.477	2.523
Buccaneer, +5dB	2.020	2.279	2.288
Destroyer, +5dB	1.985	2.258	2.349

Table 2. Objective PESQ scores [14] transformed to a scale of 1 to 5 for three different conditions: (1) Original noisy speech (2) enhanced speech using OSMS [4] and (3) Spectro-Temporal Modulation Filtering (STMF).

4. RESULTS FROM EXPERIMENTAL EVALUATIONS

Noisy speech data were generated by adding three different kinds of noise from Noisex [11] to eight clean speech samples from TIMIT [12]. The noise signals were: White Noise, Buccaneer Noise One and Destroyer Engine. The test material was prepared at two SNR ranges, +5 and +15 dB. The performance of the proposed algorithm was evaluated and compared against another system based on Optimal Smoothing and Minimum Statistics (OSMS) [4]. Test conducted included subjective quality evaluation using mean opinion score (MOS) test and objective Perceptual Evaluation of Speech Quality (PESQ)[14]. Table 1 shows the average MOS results of ten subjects for three conditions: degraded speech, enhanced using OSMS and Spectro-Temporal Modulation Filtering (STMF). The results are reported separately for different SNR and noise type. Considering the limited number of subjects used, the results show a comparable performance of the two methods. Table 2 shows the result of objective PESQ [14] test for these two approaches. In this case, STMF shows a small improvement over OSMS method.

5. ACKNOWLEDGEMENT

The authors wish to thank Dr. David Anderson of the Georgia Institute of Technology for stimulating discussions during the Telluride Neuromorphic Engineering Workshop. Partial funding for this project was obtained from the Southwest Research Institute, Air Force Office of Scientific Research, and the National Science Foundation (ITR, 1150086075).

6. REFERENCES

[1] J. S. Lim, A. V. Oppenheim, "Enhancement and bandwith compression of noisy speech", Proc. IEEE, Vol 67, pp.1586-1604, Dec. 1979.

[2] Y. Ephraim, H. L. Van Trees, "A signal subspace approach for speech enhancement", IEEE Trans. Speech and Audio Proc., Vol 3, pp.251-266, July 1995.

[3] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean square error Log-spectra amplitude estimator", IEEE Trans. Acoust., Speech and Signal Proc., vol. ASSP-33, pp. 443-445, Apr. 1985.

[4] R. Martin, "Statistical methods for the enhancement of noisy speech", Inter. Workshop on Acoust. Echo and Noise Control, Kyoto, Japan, Sept. 2003.

[5] S. Shamma, "Encoding sound timbre in the auditory system", IETE J. Res. 49(2), 193-205, 2003.

[6] K. Wang, and S. A. Shamma, *Spectral shape analysis in the central auditory system*, IEEE Trans. Speech Audio Process. 3 (5), pp. 382–395, 1995.

[7] X. Yang, K. Wang, and S. A. Shamma, *Auditory represntation of acoustic signals*, IEEE Trans. Inf. Theory, 38 (2), pp. 824–839, (Special issue on wavelet transforms and multi-resolution signal analysis), 1992.

[8] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with application to noise suppression", IEEE trans. Speech and Audio Proc., Vol. 11, No.3, May 2003

[9] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility", Speech communication, vol. 41, pp. 331–348, 2003.

[10] R. P. Carlyon, S. A. Shamma, *An account of monaural phase sensitivity*, Journal of Acoust Soc Amer. vol. 114(1), pp. 333–48, 2003.

[11] A. Varga, H .J .M Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition", Documentation included in the NOISEX-92 CD-ROMs, 1992.

[12] S. Seneff, and V. Zue, "Transcription and alignment of the timit database", J. S. Garofolo, Ed. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[13] N. Mesgarani, M. Slaney, S. Shamma, "Speech discrimination based on multiscale spectrotemporal modulations", ICASSP 2004.

[14] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", ITU-T Recommendation P.862, Feb. 2001.