SPECTRO-TEMPORAL SUBBAND WIENER FILTER FOR SPEECH ENHANCEMENT

Chung-Chien Hsu[1], Tse-En Lin[1], Jian-Hueng Chen[2] and Tai-Shih Chi[1]

[1]Department of Electrical Engineering, National Chiao Tung University, Taiwan, R.O.C.
[2]Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan, R.O.C.

ABSTRACT

In this paper, we propose a signal-channel speech enhancement algorithm by applying the conventional Wiener filter in the spectro-temporal modulation domain. The multi-resolution spectro-temporal analysis and synthesis framework for Fourier spectrograms [12] is extended to the analysis-modification-synthesis (AMS) framework for speech enhancement. Compared with conventional speech enhancement algorithms, a Wiener filter and an extended minimum mean-square error (MMSE) algorithm, our proposed method outperforms them by a large/small margin in white/babble noise conditions from both objective and subjective evaluations.

Index Terms—spectro-temporal modulation filtering, Wiener filter, speech enhancement

1. INTRODUCTION

Speech enhancement algorithms have been used to improve quality of processed speech in a wide range of applications for decades. Conventional algorithms, such as the spectral subtraction, Wiener filter, statistical-model-based methods and subspace methods, perform well in high SNR environments, but usually have unsatisfactory performance in low SNR environments [1]. These algorithms suppress noise in either the time or the frequency domain and many of them use the analysis-modification-synthesis (AMS) paradigm to enhance speech in the frequency domain. However, humans perceive and process sounds in the time and the frequency domains simultaneously so that people are not seriously affected by pure temporal or spectral interferences in noisy environments.

From psychoacoustic studies, slow temporal modulations (≤16 Hz) were shown highly related to speech intelligibility by assessing speech intelligibility after smearing the temporal envelopes of speech [2][3]. These findings inspired many engineering approaches in speech applications. For example, temporal modulations were extracted from each frame in a spectrogram to generate the bi-frequency (acoustic frequency and temporal modulation frequency) representation for audio coding [4], and for speech separation [5]. In addition, features with temporal modulation information were adopted in building a more robust automatic speech recognition (ASR) system [6] and a speaker identification system [7]. Another example was the conventional spectral subtraction method was applied in the temporal modulation domain for speech enhancement [8].

Meanwhile, neuro-physiological evidences suggest that neurons of the auditory cortex (A1) tune to different spectro-temporal (ST) modulations of input sounds and a computational auditory model was proposed accordingly [9]. From engineering viewpoints, A1 neurons decompose the spectrogram of an input sound into many spectrograms at different joint ST modulation resolution. Psychoacoustic experiments were also carried out to determine the critical ranges of ST modulation parameters of speech for speech comprehension by human listeners [10]. Not surprisingly, the concept of applying ST filters in speech related applications is observed in engineering approaches, such as applying 2-D Garbor filters to mel-spectrogram to extract robust features for ASR systems [11].

Based on the auditory model in [9], we have proposed a ST modulation analysis-synthesis framework for Fourier spectrograms [12]. Extended from that work, the AMS paradigm is adopted to the ST modulation domain for speech enhancement in this paper. The rationale of our approach is that noises would not uniformly degrade speech at all ST resolutions. The rest of this paper is organized as follows. In section 2, we briefly introduce our previous work, the ST analysis and synthesis framework for Fourier spectrograms, and demonstrate ST contents of typical speech and noise signals. In section 3, we propose a speech enhancement algorithm by adopting the conventional Wiener filter in the ST modulation AMS paradigm. Objective and subjective speech quality evaluations are also given in section 3. We end in section 4 with the conclusion and discussions.

2. SPECTRO-TEMPORAL ANALYSIS AND SYNTHESIS OF FOURIER SPECTROGRAMS

In our previous work [12], the concept of the ST analysis of the cortex was implemented for Fourier spectrograms. Outputs of the analytical stages are inverted to sounds using the overlap-and-add method in a timely manner and with

This research is supported by National Science Council, R.O.C. with Grant NSC 100-2220-E-009-004 and Chunghwa Telecom Co., Ltd.

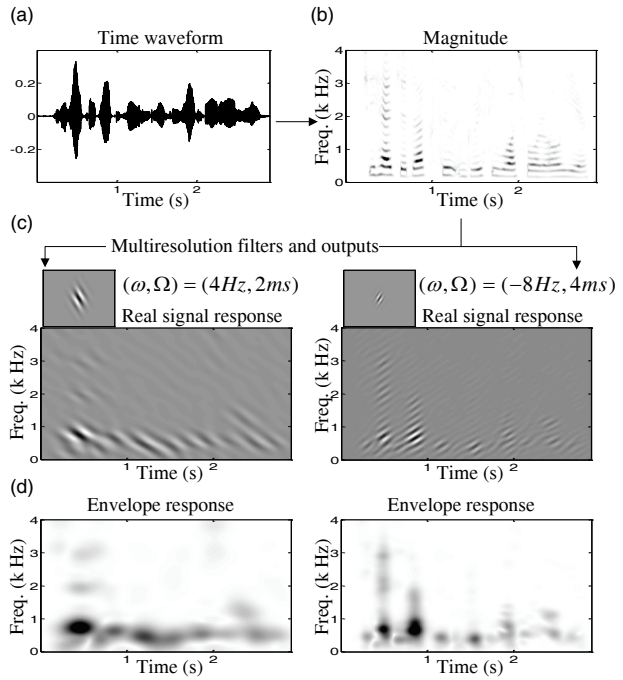

Fig. 1. Examples of STMFs and corresponding outputs.

less distortion than the projection method proposed in [9]. The ST analytical stage captures prominent speech structures, such as pitch, harmonicity, formant, amplitude modulation (AM) and frequency modulation (FM), using different ST modulation filters [12].

2.1. Spectro-temporal analysis

First, the Fourier spectrogram of a sound is obtained by the short-time Fourier transform (STFT) with a 25-ms Hamming window, 5-ms shift and the 1024-point discrete Fourier transform (DFT). Then, a bank of *constant-Q non-causal zero-phase* bandpass spectro-temporal modulation filters (STMFs) is adopted to decompose the spectrogram. Note, the STMFs are designed to possess downward and upward directivity to simulate the directional preference of A1 neurons. The frequency responses of the downward (with subscript "+") and the upward (with subscript "-") STMFs can be written as

$$STMF_{+}(\omega,\Omega) = \begin{cases} \left|\mathcal{F}\{h_{rate}(t)\} \otimes \mathcal{F}\{h_{scale}(f)\}\right|, & 0 \le \omega;\Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$STMF_{-}(\omega,\Omega) = \begin{cases} \left|\mathcal{F}\{h_{rate}(t)\} \otimes \mathcal{F}\{h_{scale}(f)\}\right|, & -\pi \le \omega \le 0;0 \le \Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathcal{F}$ is the 1-D Fourier transform; $\otimes$ is the outer product and π indicates the half of the sampling frequencies of the time and the frequency axes of the Fourier spectrogram. The *rate* (ω in Hz, as frequency) and the *scale* (Ω in ms, as quefrency) are defined as the Fourier domain of the time and frequency axis, respectively. The h_{rate} and h_{scale} are the 1-D constant-Q (Q_{3dB} = 2) temporal and spectral impulse responses derived from gammatone filters. Detailed equations of h_{rate} and h_{scale} can be found in [12]. Note, based on (1) and (2), the downward and upward STMFs have components only in the first and second quadrant of the ω-Ω space, respectively. Finally, the four-dimensional output of the bank of STMFs for an arbitrary input Fourier spectrogram $X(t,f)$ is given by:

$$C(t,f,\omega,\Omega) = \mathcal{F}_{2D}^{-1}\{\mathcal{F}_{2D}\{X(t,f)\} \cdot STMF_{\pm}(\omega,\Omega)\} \quad (3)$$

where $\mathcal{F}_{2D}$ and $\mathcal{F}_{2D}^{-1}$ denote the 2-D Fourier transform and the inverse 2-D Fourier transform.

Fig. 1 shows examples of the STMFs and corresponding outputs. The Fourier spectrogram of an utterance from a female speaker is presented in Fig. 1(b). The smaller panels in Fig. 1(c) depict the real part of the impulse responses of a downward STMF tuned to (ω_c = 4Hz, Ω_c = 2ms) and a upward STMF tuned to (ω_c = -8Hz, Ω_c = 4ms). The real part of outputs of corresponding STMFs (i.e., $\Re\{C(t,f;\omega_c,\Omega_c)\}$) are shown below in the bigger panels. The corresponding envelopes $|C(t,f;\omega_c,\Omega_c)|$ are given in Fig. 1(d).

2.2. Spectro-temporal synthesis

Eq. (3) shows that the ST analysis is a pure linear operation such that a Fourier spectrogram $X'(t,f)$ can be perfectly reconstructed from the four-dimensional representation $C(t,f,\omega,\Omega)$ by

$$X'(t,f) = \Re\{\mathcal{F}_{2D}^{-1}\{\frac{\sum_{\omega,\Omega}\mathcal{F}_{2D}\{C(t,f,\omega,\Omega)\}}{\sum_{\omega,\Omega} STMF_{\pm}(\omega,\Omega)}\}\} \quad (4)$$

The reconstructed Fourier spectrogram can be inverted to a sound using the overlap-and-add (OLA) method. In [12], we have shown this Fourier spectrogram based analysis and synthesis framework not only provides a similar ST analytical process for sounds as the auditory model [9] but also synthesizes sounds with better quality in a timely manner.

2.3. Spectro-temporal contents of speech and noise

In our ST analysis and synthesis framework, the rate ω was selected as 1~64 Hz and the scale (Ω) was selected as 0.25~16 ms to cover critical temporal modulations related to speech intelligibility and the speaking rate, and spectral modulations related to harmonicity and formant spacing [12].

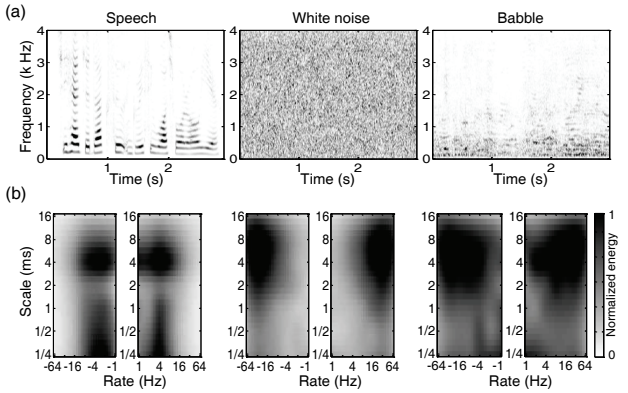

Fig. 2. Spectro-temporal modulation energy profiles of clean speech, white and babble noises, in the rate-scale domain.

The four-dimensional output $C(t, f, \omega, \Omega)$ can be further reduced to a two-dimensional representation by integrating over the time and the frequency axis as

$$P(\omega, \Omega) = \sum_t \sum_f \left| C(t, f, \omega, \Omega) \right| \quad (5)$$

where $P(\omega, \Omega)$ represents the modulation energy profile of the input sound in the joint rate-scale domain. Fig. 2 (a) demonstrates Fourier spectrograms of a clean speech sample, a white noise and a babble noise (extracted from the NOISEX-92 database) from left to right. Fig. 2(b) shows their corresponding rate-scale domain energy profiles. The prominent peaks in the energy profile of the speech signal reveal the dominant speaking rate at 4 Hz, 250 Hz harmonic spacing (4 ms) and averaged formant spacing between 1000 Hz to 4000 Hz (1~0.25 ms). On the other hand, the rate-scale energy profiles of noises show strong energies at higher rates and higher scales. This suggests speech signals possess smoother ST modulations than noises.

3. SPEECH ENHANCEMENT AND EVALUATIONS

The modulation energy profiles shown in Fig. 2(b) support the rationale of our approach that noises do not uniformly degrade speech in the rate-scale domain. Based on this observation, we propose a rate-scale subband Wiener filter approach to suppress noises. In our implementations, the bank of 2-D STMFs covers all rate-scale components and the DC value of the spectrogram. We do not pre-filter any rate-scale component in advance. These 2-D STMFs decompose a spectrogram based on different rate-scale parameter combinations. And a generalized Wiener filter [1] is applied in the rate-scale subband (ω_i, Ω_j) as follows:

$$W(f; t_n, \omega_i, \Omega_j) = \left(\frac{P_S(f; t_n, \omega_i, \Omega_j)}{P_S(f; t_n, \omega_i, \Omega_j) + \alpha P_N(f; \omega_i, \Omega_j)} \right) \quad (6)$$

where P_S and P_N are the estimated power spectra of clean speech and noise in each rate-scale subband. $P_N(f; \omega_i, \Omega_j)$ can be obtained by collapsing $P_N(t, f; \omega_i, \Omega_j)$, which is

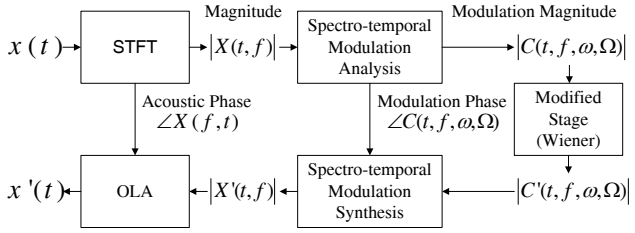

Fig. 3. Block diagram of proposed spectro-temporal AMS speech enhancement algorithm.

computed from the beginning of the input noisy signal. However, in order to have a valid estimate of $P_N(f)$ in low rate (1 Hz) subbands, a longer duration (> 1 sec) of noise was assumed present at the beginning of the input noisy signal. $P_S(f; t_n, \omega_i, \Omega_j)$ can be then obtained by subtracting the estimated $P_N(f; \omega_i, \Omega_j)$ from the power spectrum $P_{S+N}(f; t_n, \omega_i, \Omega_j)$ of the noisy signal. The parameter α is the noise attenuation factor and it affects degrees of attenuations at both high and low SNR levels [1].

The modified 4-D representation was obtained by applying the gain function $W(f; t_n, \omega_i, \Omega_j)$ in each rate-scale subband (ω_i, Ω_j) as follows:

$$C'(f; t_n, \omega_i, \Omega_j) = W(f; t_n, \omega_i, \Omega_j) \cdot C(f; t_n, \omega_i, \Omega_j) \quad (7)$$

The modified spectrogram is then reconstructed using (4) and the enhanced speech signal is obtained by the overlap-and-add method. Our proposed ST AMS speech enhancement algorithm is summarized in Fig. 3.

Speech samples in NOIZEUS corpus [1], which contains thirty phonetically-balanced sentences spoken by three male and three female speakers (five sentences per speaker), were used in our evaluations. In our experiments, noisy signals were generated by adding white and babble noises from NOISEX-92 at four SNR levels (15dB, 10dB, 5dB, 0dB). Our algorithm was evaluated objectively and subjectively by comparing with a Wiener filter [13], which continuously updates the prior SNR estimate, and an extended minimum mean-square error (MMSE) method [14], which was designed to reduce musical noise generated by the original MMSE method. The PESQ speech quality score [15], which has been shown more reliable and with a higher correlation with subjective listening test results than other objective measures [16], was used in our objective tests. In subjective listening tests, the sounds produced by all enhancement methods at all SNR levels were mixed in a random order and presented to 10 subjects (22~26 years old with normal hearing) via a Sennheiser HD 380 Pro headphone. Subjects were requested to rate quality of perceived speech in a five-point scale (1: bad; 2: poor; 3: fair; 4: good; 5: excellent). The average score from all subjects is referred to as the mean opinion score (MOS).

To determine the parameter α, we calculated the average PESQ score over all SNR conditions with added

white noise. The average PESQ score climbed from 2.45 (α = 1), plateaued at 2.67 (α = 7), and began to drop when α > 10. Therefore, we set α = 7 for all rate-scale subbands in our proposed method. The mean and standard deviation of the PESQ and MOS scores are plotted in top two panels and bottom two panels of Fig. 4, respectively, for each enhancement method, each SNR level, and each noise type. Our proposed method achieved significant improvement under all SNR conditions in white noise and slight improvement in babble noise. This is because babble noise are highly overlapped with speech in the rate-scale domain as shown in Fig. 2. Sound examples from all enhancement methods in our comparisons are available at http://perception.cm.nctu.edu.tw/sound-demo/.

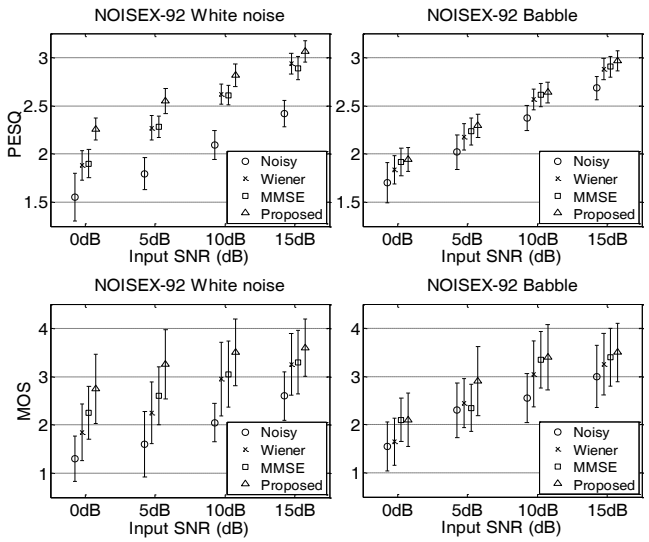

Fig. 4. Objective PESQ and subjective MOS scores of enhanced speech for various enhancement methods, SNR levels, and noise types.

4. CONCLUSION AND DISCUSSIONS

In this paper, we proposed a single-channel speech enhancement algorithm which adopts the AMS paradigm in ST modulation domain. Simulation results showed our proposed ST Wiener filter achieved higher performance in objective and subjective tests.

As indicated by Fig. 3, only the rate-scale subband magnitudes are modified while the ST modulation phase and acoustic phase remain unchanged in our method. Phase retrieval algorithms [17] may further enhance quality of reconstructed speech by estimating an appropriate phase response for a given modified magnitude response. Furthermore, the α parameter is set the same in all subbands currently. We believe different parameters are needed in different subbands based on the energy distribution of the background noise in the rate-scale domain. A time-varying estimator for ST contents of noise will be incorporated in the future to further boost the performance of our proposed method.

5. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC, New York, 2007).

[2] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053-1064, 1994.

[3] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science,* vol. 270, no. 5234, pp. 303-304, 1995.

[4] J. K. Thompson and L. E. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," in *Proc. ICASSP*, vol. 5, pp. 397-400, 2003.

[5] S. M. Schimmel, L. E. Atlas and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. ICASSP*, vol. 4, pp. 605-608, 2007.

[6] J. Tchorz and B. Kollmeier , "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 2040-2050, 1999.

[7] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 90-100, 2009.

[8] K. Paliwal, K. Wójciciki and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Comm.* , vol. 52, no. 5, pp. 450-475, 2010.

[9] T. Chi, P. Ru and S. A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.

[10] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comp. Bio.*, vol. 5, no. 3, e1000302, 2009.

[11] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Comm.*, vol. 53, no. 5, pp. 753-767, 2011.

[12] T.-S. Chi and C.-C. Hsu, "Multiband analysis and synthesis of spectro-temporal modulations of Fourier spectrogram," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. EL190-EL196, 2011.

[13] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, vol. 2, pp. 629-632, 1996.

[14] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Process.,* vol. 2, no. 2, pp. 345-349, 1994.

[15] "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, ITU-T, Geneva, Switzerland, p. 862 (2001).

[16] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no.1, pp. 229-238, 2008.

[17] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236-243, 1984.