IDIAP RESEARCH REPORT

EFFECTIVE POST-PROCESSING FOR SINGLE-CHANNEL FREQUENCY-DOMAIN SPEECH ENHANCEMENT

Weifeng Li [a]

IDIAP–RR 07-71

JANUARY 2008

SUBMITTED FOR PUBLICATION

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-71

Effective post-processing for single-channel frequency-domain speech enhancement

Weifeng Li

January 2008

submitted for publication

Abstract. Conventional frequency-domain speech enhancement filters improve signal-to-noise ratio (SNR), but also produce speech distortions. This paper describes a novel post-processing algorithm devised for the improvement of the quality of the speech processed by a conventional filter. In the proposed algorithm, the speech distortion is first compensated by adding the original noisy speech, and then the noise is reduced by a post-filter. Experimental results on speech quality show the effectiveness of the proposed algorithm in lower speech distortions. Based on our isolated word recognition experiments conducted in 15 real car environments, a relative word error rate (WER) reduction of 10.5% is obtained compared to the conventional filter.

1 Introduction

Modern communication systems employ some speech enhancement algorithms at the pre-processing stage prior to further processing (such as speech coding or automatic speech recognition (ASR)). Over the past three decades, frequency domain enhancement methods have received significant interest due to their relatively good performance and low computational cost. The first one is the well-known "spectral subtraction" method [1]. There have also been other development methods, e.g., Wiener filter, short-time spectral amplitude (STSA) analysis with different estimation techniques, such as maximum likelihood (ML) [2], minimum mean square error (MMSE) [3], and maximum a posteriori (MAP). While most of the above speech estimators improve the signal-to-noise ratio (SNR), they also produce speech distortions, mainly due to inaccurate or erroneous noise or SNR estimation. In fact, as indicated in [4], generally no or hardly any improvements regarding speech intelligibility are found with single-microphone speech enhancement algorithms.

Perceptually motivated speech enhancement methods have been proposed to lower speech distortion by exploiting the masking properties from psycho-acoustics. These methods, however, are largely dependent on the accurate estimation of the masking threshold in noise. In low SNR conditions, the estimated masking thresholds might deviate from the true ones resulting in additional residual noise [5]. Moreover, trying to mask the distortions of the residual noise leads into a variable speech distortion [6].

In this paper, we propose a novel post-processing algorithm for reducing the speech distortion caused by the use of conventional filters, while maintaining the noise reduction abilities. The proposed algorithm consists of two stages. In the first stage, the speech processed (or enhanced) by a conventional filter is compensated by adding the original noisy speech. The second stage incorporates a Wiener filter to remove additional residual noise using the cross-spectrum between the original speech and the speech processed by the conventional filter. The proposed post-processing algorithm is universal and may be applied to different types of conventional speech enhancement filters to achieve better performance.

The organization of this paper is as follows: In Section 2, we formulate the proposed filter. In Section 3, we present the performance evaluation. Section 4 summarizes this paper.

2 Algorithms

2.1 Formulation of the proposed filter

Let the corrupted speech signal $x(i)$ be represented as

$$x(i) = s(i) + n(i), \tag{1}$$

where $s(i)$ is the clean speech signal and $n(i)$ is the noise signal. By using the short-time Fourier transform (STFT), in the time-frequency domain we have

$$X(k,l) = S(k,l) + N(k,l), \tag{2}$$

where k and l denote frequency index and frame index, respectively. For compactness, we will drop both the frequency bin index k and the frame index l in this section.

Fig. 1 shows a diagram of the proposed filtering operation. After the noise estimation we apply a conventional (original) filter with a multiplicative nonlinear gain function G_1 to the amplitude of X, and by incorporating the phase of X we obtain

$$\begin{aligned} \hat{S}_1 &= G_1 \cdot X && (3)\\ &= S + \tilde{N}, && (4) \end{aligned}$$

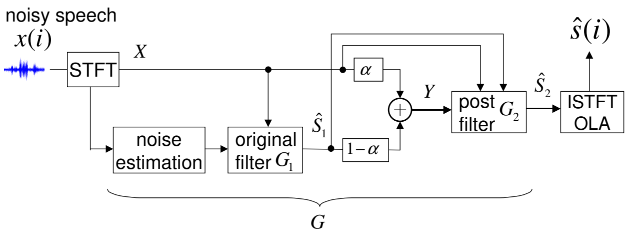

Figure 1: Diagram of the proposed algorithms.

where we model $\tilde{N}$ as the short-time spectrum of residual noise $\tilde{n}$ in the processed speech. Then the speech processed by a conventional filter is compensated by adding the original noisy speech, i.e.,

$$\begin{aligned} Y &= \alpha X + (1-\alpha)\hat{S}_1 &(5)\\ &= \alpha(S+N) + (1-\alpha)(S+\tilde{N}) &(6)\\ &= S + \alpha N + (1-\alpha)\tilde{N} &(7)\\ &= [\alpha + (1-\alpha)G_1]\cdot X, &(8)\end{aligned}$$

where α is the parameter that controls the degree of the added noisy speech ($0 \leq \alpha \leq 1$). This kind of compensation is expected to reduce the speech distortion caused by the conventional filter G_1. In order to reduce the additive noise in the compensated speech Y, we propose a post-filter

$$\begin{aligned} G_2 &= \frac{P_{X\hat{S}_1}}{P_{YY}} &(9)\\ &= \frac{G_1}{[\alpha + (1-\alpha)G_1]^2}, &(10)\end{aligned}$$

which utilizes the cross-spectrum between X and $\hat{S}_1$, to be applied to the new noisy speech Y. Here we derive Eq. (10) using Eqs. (3) and (8). As a whole, the proposed filter (gain function) can be formulated as

$$G = \frac{G_1}{\alpha + (1-\alpha)G_1}. \tag{11}$$

Finally, the enhanced speech $\hat{s}(i)$ is obtained through the inverse short-time Fourier transform (ISTFT) and overlap-add (OLA) synthesis.

2.2 Analysis of the proposed filter

With the real value of G, we can formulate the error between the spectrum of the clean signal and the estimated one as

$$\begin{aligned} \mathcal{E} &= E[|G\cdot X - S|^2]\\ &= E[|G\cdot(S+N) - S|^2]\\ &= (G-1)^2\cdot E[|S|^2] + G^2\cdot E[|N|^2]\\ &+ (G-1)G\cdot E[S\cdot N^* + S^*\cdot N], &(12)\end{aligned}$$

where $E[\cdot]$ denotes the expectation operator and $*$ indicates the complex conjugate operator. If we assume that the speech and noise are uncorrelated, the third term in the above equation can be negligible. The first term describes the speech distortion while the second term indicates the

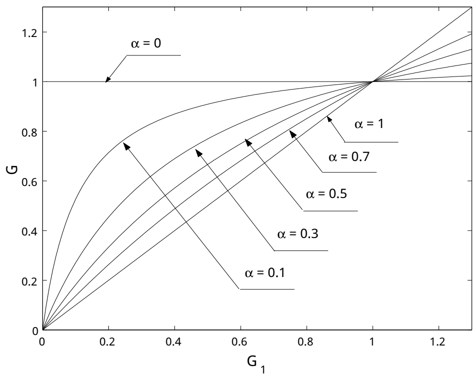

Figure 2: Parametric gain curves of resulted filter G as a function of the original filter G_1.

noise distortion. As shown in [6], complete masking of both speech and noise distortions can not be guaranteed and we must settle for a trade-off between the two distortions (For example, perceptually motivated methods try to mask noise distortion by allowing a variable speech distortion [6]). When $G_1 < 1$, our method aims to reduce the speech distortion compared to the original filter, since G is always larger than G_1 (see Fig. 2). When $G_1 > 1$ (e.g., may arise in Ephraim-Malah algorithms), using the presented post-filter results in the reduction of both speech and noise distortions compared to the original filter. The parameter α provides a soft transition between the original noisy speech ($\alpha = 0$) and the speech processed with the original filter ($\alpha = 1$), and plays the role of controling the trade-off between noise reduction and speech distortion.

Compared to two-stage Wiener filtering [7], in the second stage we use cross-spectrum and avoid estimating the noise or SNR, which may introduce additional errors. Moreover, in [7] Wiener filters are designed in the frequency domain, whereas the filters are applied in the time domain using convolution operations. The proposed one implements the two filters consistently in the frequency domain, which avoids the re-calculation of the power spectrum in time-frequency switches and improves computational efficiency.

3 Performance Evaluation

For evaluation purposes, 100 utterances from Aurora-2J database are used (Aurora-2J is the same as Aurora-2, but uttered in Japanese [8]). The speech signals are sampled at 8 kHz and degraded by three types of noise (subway, babble, car) at different SNR levels from 0 dB to 20 dB in 5 dB steps. The spectral analysis is implemented with hamming windows of 32 ms and a frame shift of 16 ms. A *minimum mean-square error log-spectral amplitude* (MMSE-LSA) estimator [3] is used as an original filter as shown in Fig. 1 (Other estimators can also be applied). An *improved minima controlled recursive averaging* (IMCRA) method [9] was used to estimate the noise. The *a priori* SNR was calculated using "decision-directed" approach. The following three types of speech signals were evaluated:

1. noisy: degraded noisy speech ($\alpha = 0$);
2. original filter: speech enhanced using MMSE-LSA estimator ($\alpha = 1$);
3. presented methods: speech enhanced using the proposed algorithm by cascading the original MMSE-LSA estimator with different values of α ([0.1 0.3 0.5 0.7 0.9]).

We compute two objective measures, the segmental SNR and the weighted cepstral distance (WCD). Fig. 3 summarizes the results of the segmental SNR for various noise types (averaged over [0, 20] dB for each type). As can be seen, the segmental SNRs are significantly improved in all three

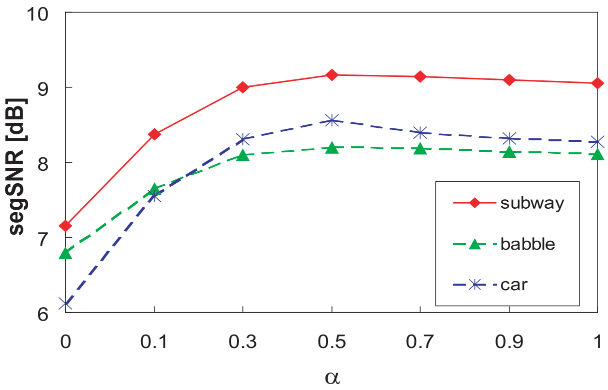

Figure 3: Segmental SNR performance as a function of α.

noise types compared to the noisy speech. The segmental SNR of the proposed algorithms depends on the parameter α. When α increases up to 0.3 or above, the proposed algorithms can perform as well as the original filter ($\alpha = 1$). In informal listening, compared to the speech processed by the original filter, the speech signals reconstructed using the proposed method are judged to be more "crisp" and involve less "musical" artifacts although a little original noise is introduced. Fig. 4 shows an example of spectrograms for different speech, demonstrating that the missed spectrograms in the speech processed by the original filter are partly recovered by using the proposed post-processing algorithm.

We also evaluate the enhanced speech using the weighted cepstral distance (WCD) measure, which is defined as

$$WCD = \frac{1}{L}\sum_{l=1}^{L}\sum_{j}^{p} w_j[c(l,j) - \hat{c}(l,j)]^2, \tag{13}$$

where c and $\hat{c}$ are cepstral coefficients corresponding to the clean signal and the estimated signal, respectively. p is the order of the model (chosen equal to 14) and w_j is the weight for the i^{th} order coefficient. L is the number of frames in one utterance. As Fig. 5 shows, in subway and babble nonstationary noise cases, the original filter does not provide significant improvement in the WCD measure. Compared to the original filter, the incorporation of the proposed post-processing provides considerable improvement (with $\alpha = 0.3$). The above two figures illustrate that with a suitable value of α the proposed algorithms can reduce speech distortions while maintaining noise reduction abilities of the original filter.

In order to evaluate the proposed algorithms, we also performed speech recognition experiments using realistic data. CIAIR in-car speech corpus [10] was used. The test data were based on 50

Table 1: 15 driving conditions (3 driving environments $\times$ 5 in-car states)

driving environment	idling city driving expressway driving
in-car state	normal CD player on air-conditioner (AC) on at low level air-conditioner (AC) on at high level window (near driver) open

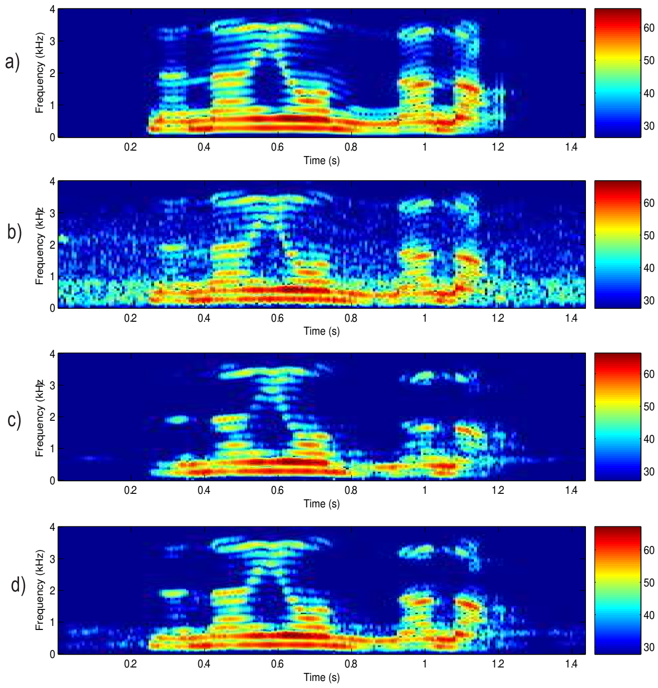

Figure 4: Spectrograms of the "747" uttered in Japanese. a) clean speech; b) corrupted speech with car noise (10 dB); c) enhanced speech obtained using the original filter (MMSE-LSA); d) enhanced speech obtained using the proposed method ($\alpha = 0.5$).

isolated word sets collected under 15 real driving conditions (listed in Table 1) using a microphone set on the visor position to the driver. 1,000-state triphone Hidden Markov Models (HMM) with 32 Gaussian mixtures per state were used for acoustical modeling. They were trained over a total of 7,000 phonetically balanced sentences collected in the idling-normal and city-normal conditions. The feature vector was a 25-dimensional vector (12 CMN-MFCC + 12 Δ CMN-MFCC + Δ log energy).

For comparison, we also performed recognition experiments using ETSI advanced front-end [11]. The acoustical model used for ETSI advanced front-end experiments was trained over the training data processed with ETSI advanced front-end. Fig. 6 shows the recognition performance averaged over the 15 driving conditions (0.3 and 0.5 are used for α in the proposed method). We found that all the enhancement methods outperformed the original noisy speech. ETSI advanced front-end marginally outperformed the original filter (MMSE-LSA), while the proposed method achieved a relative word error rate (WER) reduction of 10.5% compared to ETSI advanced front-end.

4 Summary

In this paper, we have proposed a post-processing algorithm for the improvement of the quality of speech processed by a conventional filter. Our experiments demonstrated that the proposed post-processing with a suitable value of α can reduce speech distortion caused by the original filter. The proposed algorithm is universal and may be applied to different types of conventional speech enhancement filters. Since α should be changed in time-frequency, the adaptive optimization of α is worth exploiting and will be the direction of our future work. On the other hand, during speech absence the proposed method is not effective, and speech presence uncertainty may be combined to achieve better performance.

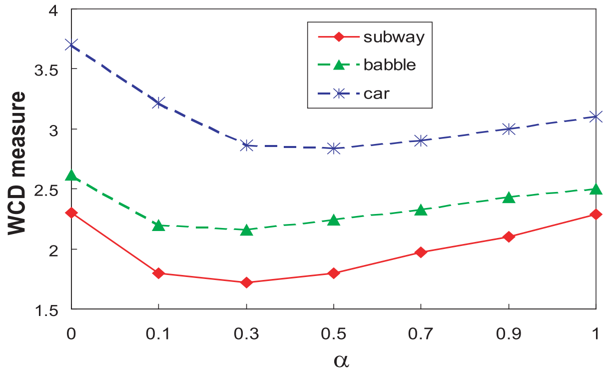

Figure 5: Weighted cepstral distance (WCD) performance as a function of α.

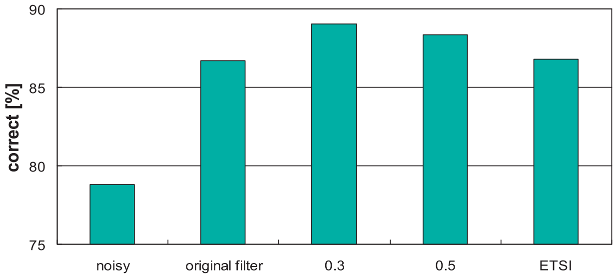

Figure 6: Recognition performance for different methods.

Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-033812) and the Swiss National Science Foundation through the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.ASSP-27, no.2, pp.113-120, 1979.

[2] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE. Trans. Acoustics, Speech, and Signal Processing* vol.ASSP-28, no.2, pp.137-145, 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.ASSP-33, no.2, pp.443-445, 1985.

[4] G.A. Studebaker and I. Hochberg (Editors). *Acoustical Factors Affecting Hearing Aid Performance*, second edition, Boston: Allyn and Bacon, 1993.

[5] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol.11, no.5, pp.457-465, 2003.

[6] S. Gustafsson, P. Jax and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving Background noise characteristics," in *Proc. ICASSP*, pp. 397-400, 1998.

[7] A. Agarwal and Y.M. Cheng, "Two-stage mel-warped wiener filter for robust speech recognition," In *Proc. IEEE ASRU workshop*, pp. 67-70, 1999.

[8] S. Nakamura, K. Takeda, *et al.*, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," *IEICE Trans. Information and Systems*, vol. E88-D, no.3, pp. 535-544, 2005.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol.11, no.5, pp.466-475, 2003.

[10] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," in *Proc. ICSLP*, pp.362-365, 2000.

[11] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.1, 2002.