interspeech 2008 incorporating SST 08

Noise Driven Short-Time Phase Spectrum Compensation Procedure for Speech Enhancement

Anthony P. Stark, Kamil K. Wójcicki, James G. Lyons and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane Queensland 4111, Australia
{a.stark, k.wojcicki, j.lyons, k.paliwal}@griffith.edu.au

Abstract

Typical speech enhancement algorithms operate on the short-time magnitude spectrum, while keeping the short-time phase spectrum unchanged for synthesis. Recently, a novel approach to speech enhancement has been proposed where the noisy magnitude spectrum is recombined with a changed phase spectrum to produce a modified complex spectrum. During synthesis the low energy components of the modified complex spectrum cancel out more than the high energy components, thus reducing background noise. In the present work, a procedure that employs noise estimates to compensate the phase spectrum for additive noise distortion is formulated. The proposed approach is objectively evaluated against several popular speech enhancement methods under various noise conditions and is shown to compare favourably.

Index Terms: speech enhancement, magnitude spectrum, phase spectrum, phase spectrum compensation

1. Introduction

In the field of speech enhancement we are interested in the reduction of noise from noise-corrupted speech in order to improve its intelligibility and quality. Various methods have been investigated in the literature for performing speech enhancement. These can be grouped into spectral subtraction [1], MMSE estimation [2], Wiener filtering (linear MMSE) [3], Kalman filtering [4] and subspace [5] methods. Several of these methods employ the analysis-modification-synthesis (AMS) framework [6].

Let us consider an additive noise model

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

where $x(n)$, $s(n)$ and $d(n)$ denote discrete-time signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analysed frame-wise in the AMS framework through short-time Fourier analysis. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal $x(n)$ is given by

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)\mathrm{e}^{-j2\pi km/N}, \tag{2}$$

where k denotes the kth discrete-frequency of N uniformly spaced frequencies and $w(n)$ is an analysis window function. In speech processing, the Hamming window with 20–40 ms duration is typically employed. Using DSTFT analysis we can, subject to constraints described in [7], represent Eq. (1) as

$$X(n,k) = S(n,k) + D(n,k), \tag{3}$$

where $X(n,k)$, $S(n,k)$, and $D(n,k)$ are the DSTFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of the DSTFT magnitude spectrum and the DSTFT phase spectrum. For instance, the DSTFT of the noisy speech signal can be written in polar form as

$$X(n,k) = |X(n,k)|\mathrm{e}^{j\angle X(n,k)}, \tag{4}$$

where $|X(n,k)|$ denotes the magnitude spectrum and $\angle X(n,k)$ denotes the phase spectrum.[1]

Traditional AMS-based speech enhancement methods modify only the magnitude spectrum while keeping the noisy phase spectrum unchanged for synthesis. In the present work we take the opposite approach. We modify the noisy phase spectrum and leave the noisy magnitude spectrum unchanged. Noise suppression is achieved by altering the DSTFT phase spectrum in such a way as to induce large synthesis cancellation among noise components during the inverse DSTFT operation. A preliminary study of this novel technique has shown encouraging results [8]. In the present work, we formulate a procedure that employs a noise estimate to compensate the phase spectrum for additive noise distortion. Using an objective speech quality measure and spectrogram analysis, we demonstrate that the proposed method compares favourably to other popular speech enhancement techniques.

2. Proposed method

The proposed speech enhancement method is based on the analysis-modification-synthesis (AMS) framework commonly employed in speech processing. The AMS framework consists of three stages: 1) the analysis stage, where the input speech is processed using DSTFT analysis (see Eq. (2)); 2) the modification stage, where the noisy complex spectrum undergoes some kind of modification; and 3) the synthesis stage, where the inverse discrete short-time Fourier transform (IDSTFT) operation is followed by the overlap-add (OLA) synthesis to construct the output signal. A block diagram of the proposed approach is shown in Fig. 1. The noisy speech signal, used in the analysis stage of the AMS framework, is a real-valued signal and therefore its DSTFT is conjugate symmetric, i.e. $X(n,k) = X^*(n,N-k)$. In our approach, we control the degree to which the conjugates reinforce or cancel by altering their angular relationship. An antisymmetry function is used for this purpose. We make the degree of phase spectrum compensation dependent on the magnitude of noise spectral estimates. Our formulation facilitates the handling of time and/or

[1] In our discussions, when referring to the magnitude spectrum, phase spectrum and complex spectrum the DSTFT modifier is implied unless otherwise stated.

Accepted after peer review of full paper

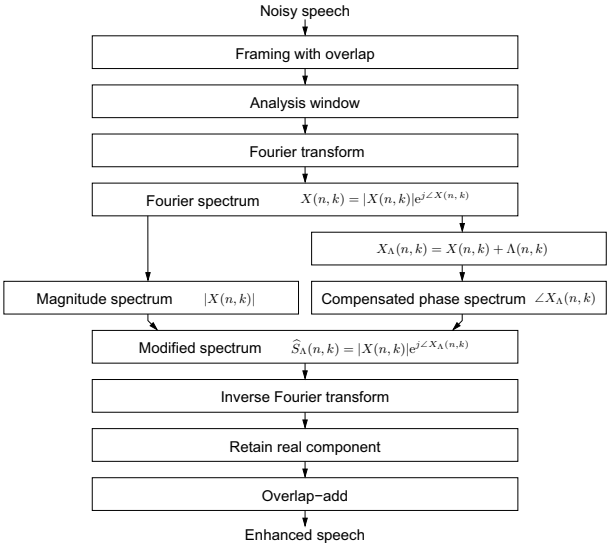

Figure 1: *Block diagram of the proposed method.*

frequency varying noise conditions. The compensated short-time phase spectrum is computed as follows. First, we obtain the phase spectrum compensation function given by

$$\Lambda(n,k) = \lambda\Psi(k)|\widehat{D}(n,k)|, \tag{5}$$

where λ is a real-valued empirically determined constant, $\Psi(k)$ is the antisymmetry function and $|\widehat{D}(n,k)|$ is an estimate of the short-time magnitude spectrum of the noise.[2] The time-invariant antisymmetry function is given by

$$\Psi(k) = \begin{cases} 1, & \text{if } 0 < k/N < 0.5 \\ -1, & \text{if } 0.5 < k/N < 1 \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where zero weighting is given to the values corresponding to non-conjugate vectors of DSTFT (i.e. the $k{=}0$ value and possible singleton at $k{=}N/2$ for $N{=}$even). Since noise magnitude estimate $|\hat{D}(n,k)|$ is symmetric, multiplication by $\Psi(k)$ produces an antisymmetric $\Lambda(n,k)$ function. It is this antisymmetry that forms the primary basis for noise cancellation during synthesis. The next step in the computation of the compensated phase spectrum is to offset the complex spectrum of the noisy speech by the additive real-valued frequency-dependent $\Lambda(n,k)$ compensation function

$$X_\Lambda(n,k) = X(n,k) + \Lambda(n,k). \tag{7}$$

The compensated phase spectrum is then obtained through

$$\angle X_\Lambda(n,k) = ARG\left[X_\Lambda(n,k)\right], \tag{8}$$

where ARG is the complex angle function. Note that the compensated phase spectrum may not possess the properties of a true phase spectrum, i.e. one that is computed from a real-valued signal. The compensated phase spectrum is recombined with the noisy magnitude spectrum to produce a modified complex spectrum

$$\widehat{S}_\Lambda(n,k) = |X(n,k)|e^{j\angle X_\Lambda(n,k)}. \tag{9}$$

[2]Note that setting the noise estimate $|\widehat{D}(n,k)|$ in Eq. (5) to unity reduces the proposed algorithm to the approach studied in [8].

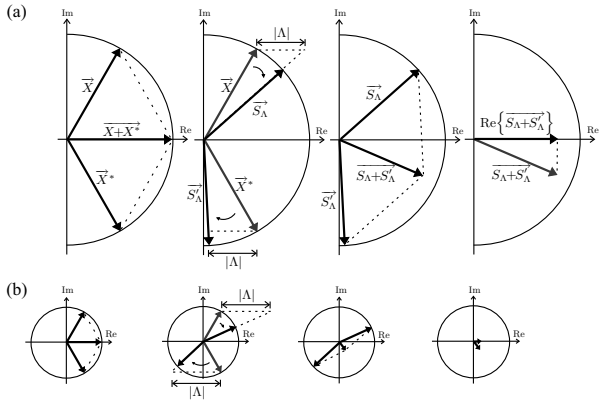

Figure 2: *Vector diagrams: modification of DSTFT conjugate symmetry of a single conjugate pair. Top row (a):* $|\overrightarrow{X}|{>}|\Lambda|$. *Bottom row (b):* $|\overrightarrow{X}|{<}|\Lambda|$. *Column one: conjugate vectors,* $\overrightarrow{X}$ *and* $\overrightarrow{X}^*$*, as well as their addition vector,* $\overrightarrow{X{+}X^*}$*. Column two: the real parts of the conjugate vectors are offset by* $|\Lambda|$ *and* $-|\Lambda|$*. Thus, the angles of vectors* $\overrightarrow{X}$ *and* $\overrightarrow{X}^*$ *are altered, while their magnitudes are kept unchanged to produce vectors* $\overrightarrow{S_\Lambda}$ *and* $\overrightarrow{S'_\Lambda}$*, respectively (see Eq. (9)). Column three: the resulting vectors are added to produce the* $\overrightarrow{S_\Lambda{+}S'_\Lambda}$ *vector. Column four: the imaginary part of the* $\overrightarrow{S_\Lambda{+}S'_\Lambda}$ *addition vector is discarded. For clarity both time and frequency indexes have been omitted in this figure.*

In the synthesis stage, the IDSTFT is used to convert the frequency-domain frames, $\widehat{S}_\Lambda(n,k)$, to the time-domain representation. Due to the additive offset introduced in Eq. (7), the resulting time-domain frames may be complex. In the proposed method the imaginary component is discarded. The enhanced time-domain signal, $\hat{s}(n)$, is produced by employing the OLA procedure. We refer to the proposed speech enhancement method as noise driven short-time phase spectrum compensation procedure, or PSC.

Figure 2 demonstrates the PSC procedure using vector diagrams for a single conjugate pair. Since $\Lambda(n,k)$ is antisymmetric, the angles of the conjugate pair being considered are pushed in opposite directions, one toward 0 radians and the other toward π radians. The further they are pushed apart, the more out of phase they become. The strength of the compensation is dependent on the magnitudes of both the DSTFT vectors and the $\Lambda(n,k)$ function. A more in-depth vector-based explanation is given in [8].

3. Experimental evaluation

3.1. Empirical search for optimal λ

The scaling factor λ is a tunable parameter that governs the degree to which noise is suppressed. As such, it is important to find a value of λ that suppresses noise while having minimal impact on speech. For this purpose, an experiment was designed to determine the optimal λ values over a range of SNRs in white noise. More specifically, the objective was to empirically determine the values of λ that would maximise objective speech quality in terms of the PESQ measure [9]. The core test set of the TIMIT speech corpus [10], which consists of 192 files from 24 speakers, was used for this task.

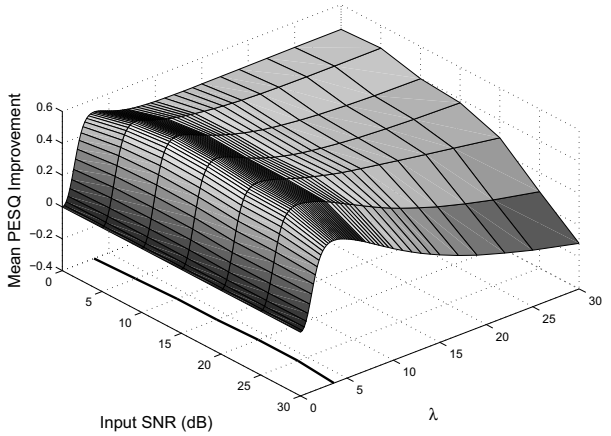

Figure 3: *Mean PESQ improvement scores as a function of λ scaling factor and input SNR (dB) for white noise. The bold line on the base of of the plot indicates λ values that produce maximum mean PESQ improvement scores as a function of SNR.*

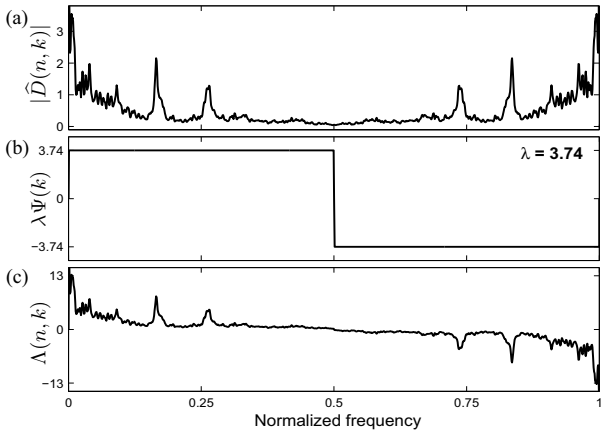

Figure 4: *Generation of the phase spectrum compensation function: (a) noise magnitude spectrum estimate, $|\widehat{D}(n,k)|$, for F16 noise; (b) antisymmetric $\lambda\Psi(k)$ weighting function; and (c) $\Lambda(n,k)$ phase spectrum compensation function.*

From the results shown in Fig. 3, it can be seen that λ=3.74 produces maximum objective speech quality across all input SNRs. This is an encouraging result, as it demonstrates the robustness of the tuning parameter λ to changing noise intensity.

Figure 4 shows an example of $\Lambda(n,k)$ function generation. In this example, a noise magnitude spectrum estimate is generated from a 120 ms sample of F16 noise. The antisymmetry function scaled by $\lambda = 3.74$ is then applied to the noise estimate to produce the phase spectrum compensation function (see Eq. (5)).

3.2. Enhancement experiments

The enhancement experiments were carried out on the core test set of the TIMIT corpus [10]. Three additive noise cases were investigated: white noise, F16 noise and babble noise. The noise signals were taken from the NOISEX-92 noise database [11] and downsampled to 16 kHz. The TIMIT utterances were corrupted with each noise type at several SNR levels. Noise estimates were computed from the initial 120 ms of each utterance. The corrupted files were enhanced using the proposed (PSC) method with λ set to 3.74 in all cases. Three other popular speech enhancement techniques were also used, namely the spectral subtraction method [1], the MMSE method [2] and the logMMSE method [12]. Mean PESQ scores were calculated for each method and each noise case.

4. Results and discussion

The results of the enhancement experiments, in terms of mean PESQ scores as well as mean PESQ improvement scores, are shown in Table 1 and Fig. 5, respectively. Compared to other methods, the proposed method performed well, providing consistent improvements across all SNRs. As a comparison, logMMSE peaks early, providing good improvements at low SNRs, but not working as well at high SNRs. Results for F16 noise are very similar to the white noise case, showing that the empirically determined λ also works well for coloured noises. For the babble noise case, the proposed method performed slightly better than the other three methods in the comparison.

The spectrograms in Fig. 6 show a single TIMIT utterance corrupted by white noise, F16 noise and babble noise at 10 dB SNR, as well as their corresponding PSC enhanced versions. The enhanced versions show little loss of speech information along with background noise suppression. For the white noise case, it should be noted that although some background noise remains, it is fairly white and lacks most of the musical noise artifacts introduced by some of the other speech enhancement methods.

5. Conclusion

In this paper a novel approach for speech enhancement has been presented, where the noisy magnitude spectrum is recombined with a phase spectrum compensated for additive noise distortion to produce a modified complex spectrum. Noise estimates are incorporated into the phase spectrum compensation procedure. During synthesis the low energy components of the modified complex spectrum cancel out more than the high energy components, thus reducing background noise. The presented method was objectively evaluated using the PESQ measure and it was shown to perform well under a variety of noisy conditions.

6. References

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, vol. 4, 1979, pp. 208–211.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32(6), pp. 1109–1121, Dec 1984.

[3] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley, 1949.

[4] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. ICASSP*, vol. 12, 1987, pp. 297–300.

[5] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 251–266, Jul. 1995.

[6] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[7] L. Alsteris and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Process.*, vol. 17, pp. 578–616, 2007.

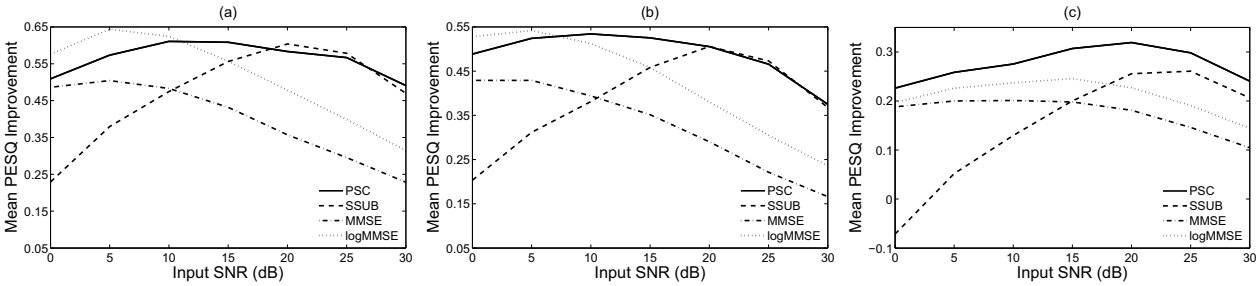

Figure 5: *Mean PESQ improvement scores for the proposed (PSC) method, spectral subtraction (SSUB) method, MMSE method and logMMSE method. Improvement scores for three noise cases are shown: (a) white noise; (b) F16 noise; and (c) babble noise.*

Table 1: *Mean PESQ scores for the proposed (PSC), spectral subtraction (SSUB), MMSE and logMMSE methods. Results for three noise cases: (a) white noise; (b) F16 noise; and (c) babble noise; for various input SNRs (dB) are shown.*

METHOD	MEAN PESQ						
SNR	**0**	**5**	**10**	**15**	**20**	**25**	**30**
				(a)			
Noisy	1.55	1.90	2.26	2.62	2.97	3.32	3.65
PSC	2.06	2.47	2.87	3.23	3.56	3.88	4.14
SSUB	1.78	2.28	2.73	3.18	3.58	3.89	4.12
MMSE	2.03	2.40	2.74	3.05	3.33	3.61	3.87
logMMSE	2.12	2.54	2.88	3.18	3.45	3.71	3.96
				(b)			
Noisy	1.67	2.03	2.38	2.73	3.08	3.42	3.72
PSC	2.16	2.55	2.92	3.26	3.59	3.88	4.10
SSUB	1.88	2.34	2.76	3.19	3.59	3.89	4.09
MMSE	2.10	2.46	2.78	3.09	3.37	3.64	3.89
logMMSE	2.20	2.57	2.90	3.19	3.46	3.72	3.96
				(c)			
Noisy	1.75	2.08	2.43	2.77	3.10	3.43	3.74
PSC	1.97	2.34	2.71	3.08	3.42	3.73	3.98
SSUB	1.68	2.13	2.56	2.97	3.36	3.69	3.94
MMSE	1.94	2.28	2.64	2.97	3.28	3.58	3.84
logMMSE	1.94	2.31	2.67	3.02	3.33	3.62	3.88

[8] K. Wójcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement," *IEEE Signal Process. Lett.*, vol. 15, pp. 461–464, 2008.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27 403–+, 1993.

[11] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33(2), pp. 443–445, 1985.

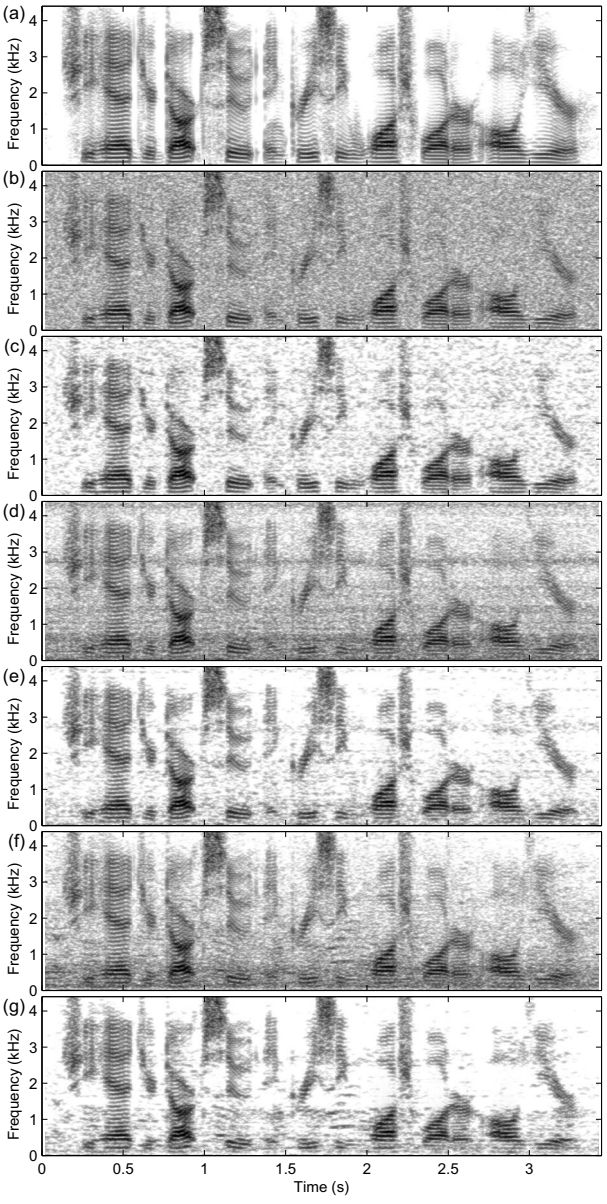

Figure 6: *Spectrograms of a sa1 TIMIT utterance "She had your dark suit in greasy wash water all year" belonging to a male speaker: (a) clean speech; (b,d,f) noisy speech at 10 dB SNR for white, F16 and babble noise cases, respectively; and (c,e,g) corresponding PSC enhanced speech.*