Spectral Subtraction with Variance Reduced Noise Spectrum Estimates

Kamil K. Wójcicki, Benjamin J. Shannon and Kuldip K. Paliwal

Signal Processing Laboratory
Griffith University, Nathan Q4111, Australia
{K.Wojcicki, B.Shannon, K.Paliwal}@griffith.edu.au

Abstract

Spectral subtraction has the drawback that it introduces an unpleasant residual noise. This noise is a result of under-subtraction which occurs due to high variance of noise magnitude spectrum estimates. In this study we investigate a number of smoothing techniques that can be employed to reduce this variability. We extend the scope of this paper by using the phase spectrum in a novel manner along with the processed magnitude spectrum. This is based on recent findings which suggest that estimation of the phase spectrum using low dynamic range analysis windows (at 20–40ms window durations) is beneficial for speech enhancement. Using an objective speech quality measure and spectrogram analysis we show that the smoothing of noise magnitude spectrum estimates is an effective method of suppressing musical noise. We also show that the use of a low dynamic range analysis window for estimation of the phase spectrum of noisy speech results in a reduction of background noise. However, we found that combining the two techniques offers no advantage over spectral subtraction alone.

1. Introduction

Spectral subtraction is a simple and effective speech enhancement technique used for restoration of speech degraded by stationary additive noise. The subtraction is performed in the magnitude spectral domain[1] within an analysis-modification-synthesis (AMS) type of framework. Small window durations (20–40ms) are employed, so that the speech signal can be considered quasi-stationary. Unfortunately spectral subtraction has a major drawback in that it introduces an unpleasant residual noise, referred to as musical noise. This residual is attributed to spurious random peaks in the magnitude spectrum of enhanced speech. These peaks are a result of under-subtraction which is caused by the high variance of noise magnitude spectrum estimates. While magnitude spectrum of noisy speech undergoes a subtraction, the phase spectrum is typically left unmodified due to a long standing belief among speech researchers that the phase spectrum is of little importance in speech enhancement (Lim and Oppenheim 1979; Wang and Lim 1982; Vary 1985). However, recent perceptual studies (Paliwal and Alsteris 2003; Paliwal 2003; Alsteris 2005; Paliwal and Alsteris 2005; Alsteris and Paliwal 2006) have shown that the phase spectrum over small window durations (20–40ms) has a significant intelligibility when estimated using a low dynamic range analysis window. Our aim in this study is two-fold. Firstly, we investigate various smoothing techniques that can be employed to reduce the variance of noise magnitude spectrum estimates, as a means for musical noise reduction. Secondly, we investigate how the dynamic range of the analysis window affects phase spectrum estimation of noisy speech within a modified AMS procedure. This paper is organised as follows: The spectral subtraction technique is introduced in Section 2. The experimental procedure, its parameters, as well as an objective speech quality measure used in our evaluation are presented in Section 3. Finally, results and discussion are given in Section 4.

2. Spectral subtraction

In this section, we review the basics of spectral subtraction. Spectral subtraction (Boll 1979; Berouti, Schwartz, and Makhoul 1979; Lim and Oppenheim 1979) is a simple yet effective speech enhancement tool used for the removal of stationary additive noise from degraded speech. As the name suggests, the noise removal is achieved by subtraction in the spectral domain. Two main assumptions are made in the spectral subtraction method: 1) noise and speech signals are uncorrelated, and 2) the additive noise signal is stationary. The effectiveness of spectral subtraction relies on the consistency and accuracy of noise magnitude spectrum estimates. Such estimation is typically performed during non-speech portions of the noisy speech. If possible, the estimates are averaged over several frames to improve consistency of the final noise estimate. The estimate of the noise magnitude spectrum is then blindly subtracted from the magnitude spectrum of the noisy speech segments.[2] The subtraction is performed within a short-time analysis-modification-synthesis (AMS) framework (Fig. 1) over small window durations (20–40ms) at which the speech signal is assumed to be quasi-stationary. High variance of noise magnitude spectrum estimates results in under-subtraction, which produces sharp spurious peaks in the magnitude spectrum of enhanced speech. This artefact is commonly referred to as musical noise and is the major drawback of the spectral subtraction method. Typically the degraded phase spectrum is left unmodified for synthesis, as it is universally believed that for small window durations (20–40ms) the STFT phase spectrum carries no useful information (Lim and Oppenheim 1979; Wang and Lim 1982; Vary 1985). Finally, the enhanced speech signal is synthesised from overlapped segments using an overlap-add procedure (Allen and Rabiner 1977; Crochiere 1980; Portnoff 1981; Griffin and Lim 1984). Let us consider the mathematical details as well as approximations used in spectral subtraction. Spectral subtraction assumes an additive noise model

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

where $y(n)$, $s(n)$ and $d(n)$ denote discrete-time signals: noisy speech, clean speech, and noise, respectively. Using discrete-time short-time Fourier transform (STFT) analysis we can equivalently represent (1) as

$$Y(n,\omega) = S(n,\omega) + D(n,\omega), \tag{2}$$

where $Y(n,\omega)$, $S(n,\omega)$, and $D(n,\omega)$ are the STFTs of noisy speech, clean speech, and noise, respectively.[3] Each of these can be expressed in terms of its STFT magnitude spectrum and STFT phase spectrum. For instance, the STFT of noisy speech signal can be written as

$$Y(n,\omega) = |Y(n,\omega)|\mathrm{e}^{j\phi(n,\omega)}, \tag{3}$$

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

$$|\widehat{S}(n,\omega)| = \begin{cases} \left(|Y(n,\omega)|^{\alpha} - \beta|\widehat{D}(n,\omega)|^{\alpha}\right)^{\frac{1}{\alpha}}, & \text{if } \left(|Y(n,\omega)|^{\alpha} - \beta|\widehat{D}(n,\omega)|^{\alpha}\right)^{\frac{1}{\alpha}} > \gamma|\widehat{D}(n,\omega)| \\ \gamma|\widehat{D}(n,\omega)|, & \text{otherwise} \end{cases} \quad (4)$$

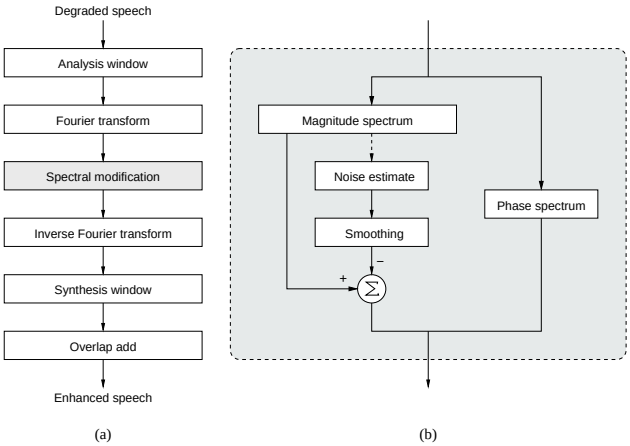

Figure 1: *Enhancement procedure: (a) analysis-modification-synthesis (AMS) framework commonly used in speech processing; (b) spectral modification stage for spectral subtraction with incorporated smoothing of noise magnitude spectrum estimate.*

where $|Y(n,\omega)|$ is the STFT magnitude spectrum and $\phi(n,\omega) = \angle Y(n,\omega)$ is the STFT phase spectrum. From the above it follows that we can approximate the magnitude spectrum of clean speech $|S(n,\omega)|$ using the generalised expression for spectral subtraction given in (4), where α, β, and γ are (typically) positive constants, and $|\widehat{D}(n,\omega)|$ is an estimate of noise magnitude spectrum, $|D(n,\omega)|$. The parameter α selects the subtraction domain. For unity α, the subtraction is performed in the magnitude spectral domain. Similarly, for α equal to two, the subtraction occurs in the power spectral domain. The choice of α has some correlation with speech intelligibility. Lower values of α result in less intelligible speech with the benefit of reduced residual noise. On the other hand higher values of α give higher intelligibility at the expense of increased musical artefact (Goh, Tan, and Tan 1998). The β parameter controls the amount of over-subtraction (Berouti, Schwartz, and Makhoul 1979). Strong over-subtraction results in strong noise suppression for both additive and residual noises; however, at the same time there is a reduction in the speech content. On the other hand, lower values of β result in suppression of additive background noise but introduce very audible musical noise. The γ parameter is used to prevent spectral subtraction from producing negative spectral magnitudes. This is achieved by setting spectral magnitude values falling below the spectral floor ($\gamma|\widehat{D}(n,\omega)|$) to that spectral floor. This reduces the dynamic range of the enhanced speech and, hence, also lowers the dynamic range of the spurious spectral peaks. As a result the intensity of musical noise is also reduced. However, raising the spectral floor too high results in an audible hum. On the other hand, very low spectral floor values make the speech sound unnatural due to the absence of ambient noise. Various extensions of spectral subtraction have been proposed in the literature. To name just a few Whipple (1994), Goh, Tan, and Tan (1998), Kamath and Loizou (2002), and Denda, Nishiura, and Kawahara (2003).

3. Evaluation framework

3.1. Experimental setup

Spectral subtraction, as described in Section 2, modifies the magnitude spectrum of degraded speech while retaining the unmodified phase spectrum. This is due to a long standing belief among speech researchers that for small window durations (20–40ms) the STFT magnitude spectrum is important, while the STFT phase spectrum carries no useful information (Lim and Oppenheim 1979; Wang and Lim 1982; Vary 1985). However, recent perceptual experiments (Paliwal and Alsteris 2003; Paliwal 2003; Alsteris 2005; Paliwal and Alsteris 2005; Alsteris and Paliwal 2006) have shown that phase spectrum contributes as much to intelligibility as magnitude spectrum, provided that a low dynamic range analysis window is used in estimation of phase spectrum.

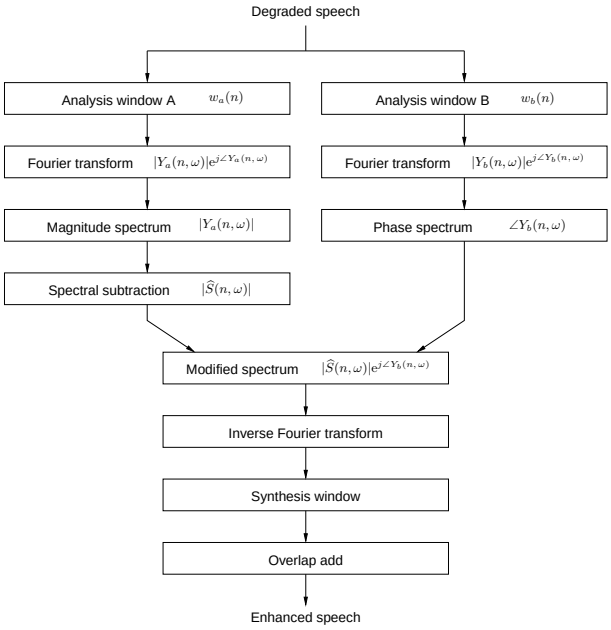

Figure 2: *Modified analysis-modification-synthesis procedure used in our evaluation.*

Recently, Shannon and Paliwal (2006) employed these findings for speech enhancement. They performed oracle type experiments, where both clean and noisy speech were made available for processing.[4] The Hamming window was used in the estimation of the magnitude spectrum of degraded speech within a modified AMS framework. No magnitude spectrum modification was performed. For estimation of the phase spectrum of clean speech, analysis windows with dynamic ranges lower than that of the dynamic range of the Hamming window were used. Shannon et al. showed that this enhancement procedure, when applied to speech degraded with additive white Gaussian noise (WGN), results in background noise reduction and an improved speech quality. Shannon et al. employed the Chebyshev window (Dolph 1946; Harris 1978; Kabal 2005) in their investigations.[5] The Chebyshev window is characterised by adjustable equi-ripple sidelobe attenuation as well as having the narrowest mainlobe for a given sidelobe attenuation. These properties make the Chebyshev window very suitable for our evaluation. Our experimental procedure is very similar to the typical AMS procedure shown in Fig. 1(a). We incorporate noise magnitude spectrum smoothing within the spectral modification stage of AMS (as per Fig. 1(b)). We also create a second analysis branch to allow for estimation of the phase spectrum using analysis window that is different from the one used for estimation of the magnitude spectrum. Our modified AMS pro-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 1: *The parameters used in our evaluation. The FFT analysis length was set to $2N = 512$, where N is the frame length in samples. Noise magnitude spectrum estimates were computed over the initial four frames of each file, by averaging. The T_w parameter is the frame duration in ms. For the stimuli D1–F5, the dynamic range of Chebyshev w_b is examined between 5dB and 65dB in 5dB increments.*

Treatment	α	β^* [dB]	γ^* [dB]	w_a	w_b	T_w [ms]	**Frame Shift**	**Smoothing Method**
A1 (*Original*)	–	–	–	Hamming	Hamming	32	$T_w/8$	–
A2 (*Degraded*)	–	–	–	Hamming	Hamming	32	$T_w/8$	–
B1	1	0	-27	Hamming	Hamming	32	$T_w/8$	–
B2	1	6	-27	Hamming	Hamming	32	$T_w/8$	–
B3	1	12	-27	Hamming	Hamming	32	$T_w/8$	–
C1	1	6	-27	Hamming	Hamming	32	$T_w/8$	Cepstral smoothing (5 cepstral coefficients kept)
C2	1	6	-27	Hamming	Hamming	32	$T_w/8$	Moving average filtering (span=5)
C3	1	6	-27	Hamming	Hamming	32	$T_w/8$	Moving average filtering (span=55)
C4	1	6	-27	Hamming	Hamming	32	$T_w/8$	MVDR spectral estimate (all-pole model order $p=12$)
C5	1	6	-27	Hamming	Hamming	32	$T_w/8$	LPF by keeping scale of 7-level db5 wavelet decomposition
D1	–	–	–	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	–
D2	–	–	–	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	–
E1	1	0	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	–
E2	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	–
E3	1	12	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	–
F1	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	Cepstral smoothing (5 cepstral coefficients kept)
F2	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	Moving average filtering (span=5)
F3	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	Moving average filtering (span=55)
F4	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	MVDR spectral estimate (all-pole model order $p=12$)
F5	1	6	-27	Hamming	Chebyshev (5–65dB)	32	$T_w/8$	LPF by keeping scale of 7-level db5 wavelet decomposition

*Note that α and γ are given above in [dB] for convenience; however, before substitution in (4) these parameters have to be converted back to linear scale.

cedure is shown in Fig. 2. In our discussions, we will use w_a to refer to the analysis window used for estimation of the discrete-time STFT magnitude spectrum within the modified AMS procedure (Fig. 2). Similarly, we will use w_b to refer to the analysis window used for estimation of the discrete-time STFT phase spectrum. Detailed settings of our experimental procedure are given in Table 1.

3.2. Speech corpus and noise type

In our evaluations we use the NOIZEUS (noisy) speech corpus (Hu and Loizou 2006b).[6] NOIZEUS is composed of 30 phonetically-balanced sentences belonging to six speakers (three males and three females). The corpus is sampled at 8kHz and filtered to simulate receiving frequency characteristics of telephone handsets. The NOIZEUS corpus comes with non-stationary noises at different SNRs. For our purposes, we use the clean speech part of the NOIZEUS corpus. We create the degraded speech by adding artificially generated WGN at different SNRs.

3.3. Smoothing techniques

As part of this study we evaluate speech quality resulting from spectral subtraction with variance reduced noise magnitude spectrum estimates. For this purpose we investigate a number of smoothing and estimation techniques which are listed below.

1. **Cepstral smoothing:** Cepstral smoothing is achieved by low-pass liftering in quefrency domain (Oppenheim and Schafer 1989). In our study we keep the first 5 cepstral coefficients.
2. **Moving average (MA):** Moving average is a simple and effective smoothing filter. The input signal is smoothed by averaging each sample over a span of adjacent samples (Smith 1999). In our investigations we employ 5 and 55 point moving average filters. We apply a given MA filter directly to the estimated noise magnitude spectrum.
3. **Minimum variance distortionless response (MVDR):** MVDR is an auto-regressive (AR) spectral estimation technique that produces smooth estimates (Wolfel and McDonough 2005). We employ an AR model order $p=12$.
4. **Low pass filtering:** Our initial aim was to employ wavelet thresholding of noise magnitude spectrum estimates for variance reduction. However, over the course of our investigations, it became evident that for the case of WGN, the thresholding of wavelet details, in addition to the keeping of approximation (or scale), does not yield any improvements over using the approximation alone. Consequently, in this study we keep only the approximation based on a 7 level wavelet decomposition, using the Daubechies db5 wavelet. As a result, we are left with a low-pass filtered noise magnitude spectrum.

3.4. Objective speech quality evaluation

Speech quality measures can be broadly categorised into subjective and objective measures. The objective measures can be further grouped into time domain, spectral domain, and perceptual domain measures (Fig. 3). The aim of the objective speech

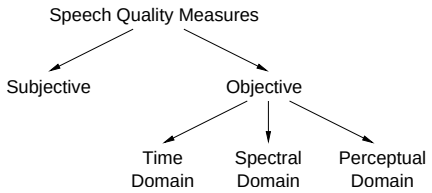

Figure 3: *Broad categories of speech quality measure types. For detailed discussion refer to Quackenbush, Barnwell, and Clements (1988), Yang (1999), and Hu and Loizou (2006a).*

quality measures is to achieve high correlation with subjective speech quality measures such as Mean Opinion Scores (MOS), or Degradation Mean Opinion Scores (DMOS). The subjective measures are based on human responses from listening experiments. For our evaluation we selected a perceptually motivated objective speech quality measure, namely the Perceptual Evaluation of Speech Quality (PESQ). The PESQ (Rix, Beerends, Hollier, and Hekstra 2001b; Rix, Beerends, Hollier, and Hekstra 2001a) algorithm is a fusion of two other perceptually motivated objective speech quality measures: PAMS and PSQM99. PESQ is based

on a model designed to have a high prediction accuracy across many types of codecs and network conditions. PESQ produces robust estimates of speech quality in the presence of a wide range of noise types, including background noises such as those used in our evaluation. PESQ prediction maps MOS[7] to a range between -0.5 and 4.5, where 1.0 corresponds to *bad* and 4.5 corresponds to *distortionless.* In our evaluation we compute mean PESQ scores over the entire NOIZEUS corpus.

3.5. Spectrogram analysis

As part of evaluation of our results we employ spectrogram analysis. The spectrogram of clean speech is given in Fig. 4 as a reference.

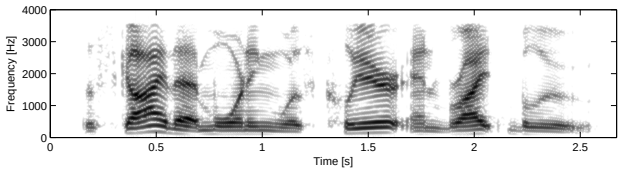

Figure 4: *Spectrogram of NOIZEUS utterance (sp10.wav): "The sky that morning was clear and bright blue". The utterance belongs to a male speaker.*

4. Results and discussion

The mean PESQ scores at different SNRs are shown in Table 2. We separate the analysis and discussion of results into the following three sections.

4.1. Spectral subtraction

In this section, we look at the results for classical spectral subtraction, as well as its extension via the smoothing of noise magnitude spectral estimates (Fig. 1(b)). We make the following observations.

Table 2: *Mean PESQ scores. For stimuli A1–C5 w_b Hamming window was used. For stimuli D1–F5, a subset of results is shown for Chebyshev w_b that gives highest mean PESQ score (as per Fig. 5(a)). Hence, for SNRs 0dB, 10dB, and 20dB, Chebyshev w_b windows with dynamic ranges of 10dB, 10dB, and 20dB are used, respectively.*

	SNR [dB]				SNR [dB]		
Treatment	0	10	20	**Treatment**	0	10	20
A1 (*Original*)	4.50	4.50	4.50	D1	3.54	3.54	3.90
A2 (*Degraded*)	1.57	2.14	2.80	D2	1.78	2.42	2.98
B1	1.87	2.56	3.26	E1	1.93	2.55	3.16
B2	2.00	2.71	3.40	E2	1.95	2.57	3.22
B3	1.76	2.46	3.22	E3	1.79	2.46	3.14
C1	2.07	2.80	3.52	F1	1.96	2.62	3.31
C2	2.04	2.75	3.45	F2	1.97	2.62	3.28
C3	2.05	2.76	3.47	F3	1.98	2.67	3.33
C4	2.08	2.79	3.50	F4	2.00	2.69	3.35
C5	2.06	2.76	3.47	F5	2.00	2.69	3.33

- The classical spectral subtraction produces an annoying spectral artefact (stimuli B1). Spectral over-subtraction (by 6dB) reduces the intensity of the musical noise (stimuli B2), however, the artefact is still very prominent. Strong over-subtraction (by 12dB) eliminates the musical noise (stimuli B3), but much of the speech content is also removed. The resulting speech loses its original quality and is less intelligible. This is reflected in the mean PESQ scores shown in Table 2.
- Smoothing of the noise magnitude spectrum combined with 6dB over-subtraction results in an improvement in speech quality and reduction of musical noise. This can be seen in the spectrograms of C5 stimuli shown in Fig. 6(c,g,k). The mean PESQ scores suggest that all of the smoothing techniques achieve comparable improvements in speech quality. However, informal subjective listening tests, as well as spectrogram analysis, revealed that cepstral smoothing (stimuli C1) is not as effective at suppression of musical noise as the other smoothing techniques used in this evaluation. The PESQ measure fails to take into account some of the artefacts produced by speech enhancement. This is because the PESQ measure was not specifically designed for speech enhancement framework. More research is needed into objective speech quality measures for speech enhancement. This need has been emphasised in the recent study by Hu and Loizou (2006a).

4.2. Low dynamic range w_b phase spectrum estimation

In this section, we look at the results for w_a set to the Hamming window, while the dynamic range of w_b is varied. No magnitude spectral modification is performed, i.e. no spectral subtraction.

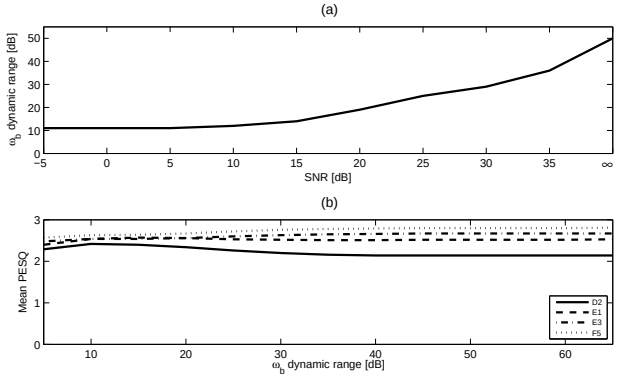

Figure 5: *Some further results on the effect of a dynamic range of an analysis window on STFT phase spectrum estimates: (a) dynamic range [dB] of w_b Chebyshev analysis window that produces highest mean PESQ score as a function of SNR [dB]; (b) mean PESQ scores for stimuli types D2, E1, E3, and F5 as a function of a dynamic range [dB] of w_b Chebyshev analysis window (SNR=10dB is shown). The Hamming window as used for w_a throughout.*

The following observations were made.

- Perfect reconstruction of clean speech within the modified AMS framework is achieved when w_a equals w_b. This is shown by maximum PESQ scores for A1 stimuli.[8]
- Analysis-modification-synthesis of clean speech with w_a different to w_b results in speech distortion (stimuli D1).
- Analysis-modification-synthesis of speech degraded with additive WGN, with w_a set to the Hamming window and w_b set to low dynamic range Chebyshev window, results in background noise reduction (stimuli D2). In addition to improvements in PESQ scores (between A2 and D2 stimuli), the noise reduction can also be seen by comparing spectrograms of noisy speech (Fig. 6(a,e,i)) with corresponding spectrograms of enhanced speech (Fig. 6(b,f,j)). Interestingly, the strongest noise reduction occurs in high energy regions, i.e. during voiced speech.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

- It was observed that the dynamic range of w_b which results in highest mean PESQ score, is SNR dependant (Fig. 5(a)). Low dynamic range w_b windows were found to be best suited for phase estimation in low SNR noisy speech. For example, mean PESQ scores at 10dB SNR for different Chebyshev windows are shown in Fig. 5(b). The solid line (stimuli D2) peaks for Chebyshev w_b with dynamic range between 10 and 15dB.
- Shannon and Paliwal (2006) used noisy speech for estimation of magnitude spectrum and clean speech for estimation of phase spectrum. They found that the Chebyshev w_b with dynamic range set to 30dB consistently produced highest mean PESQ scores when combined with magnitude spectra of noisy speech (0–15dB SNR). In our experiments, we assume that only degraded speech is available for processing. Consequently, we find that our results fall short of the best-case-scenario mean PESQ scores achieved by Shannon et al. Interestingly, however, our highest mean PESQ scores for w_b Chebyshev 10dB at 0dB SNR, and for w_b Chebyshev 10dB at 10dB SNR match closely with mean PESQ scores obtained by Shannon et al. using the same windows and SNRs in their oracle type experiments.

4.3. Spectral subtraction combined with low dynamic range w_b phase spectrum estimation

In this section, we look at the results from combining spectral subtraction (incorporating smoothed noise magnitude spectral estimates) with estimation of noisy phase spectra using low dynamic range analysis window.

- By comparing spectrograms for C5 stimuli (Fig. 6(c,g,k)) with spectrograms for F5 stimuli (Fig. 6(d,h,l)) it can be seen that both methods are effective in removal of background noise. However, the speech content of F5 stimuli is suppressed more than that of C5 stimuli. This drawback is reflected in slightly lower mean PESQ scores of F5 stimuli.
- There is some musical residual left, for stimuli C5 and F5, at very low SNRs (0dB). The residual is less notable for F5 stimuli than it is for C5 stimuli.

5. Conclusions

Based on an objective speech quality measure and spectrogram analysis, this paper shows that the smoothing of magnitude noise spectrum estimates is an effective method of musical noise suppression. This paper also shows that using a low dynamic range analysis window for the estimation of phase spectrum of noisy speech results in background noise reduction. However, combining spectral subtraction with phase estimated using a low dynamic range analysis window offers no advantage over spectral subtraction alone.

Notes

[1] When this paper refers to phase or magnitude spectrum, the use of short-time Fourier transform (STFT) over small window durations (20–40ms) is implied, unless otherwise stated. Also, we use the term "small" to refer to the length of window duration, and the term "short-time" to indicate analysis over a finite duration.

[2] The subtraction can instead be performed in power spectral domain (or higher α could be used, see (4)). The subtraction in power spectral domain may be advantageous in a presence of strong noise.

[3] We employ discrete-time STFT in our discussions. The "discrete-time" modifier is implied where not specifically stated.

[4] Typically, this kind of experiment is useful for finding some kind of upper limit of performance.

[5] The Chebyshev window is also know in the literature as the Dolph-Chebyshev window.

[6] NOIZEUS is publicly available at *http://www.utdallas.edu/~loizou/speech/noizeus/*

[7] MOS has a range of ratings from 1 to 5 corresponding to the following subjective speech quality levels: *bad*, *poor*, *fair*, *good*, and *excellent*, respectively.

[8] While for A1 stimuli, both w_a and w_b were set to the Hamming window, the perfect reconstruction can be achieved with an arbitrary pair of identical, nonzero analysis windows. For example, w_a and w_b set to the Chebyshev 10dB window also achieves perfect reconstruction.

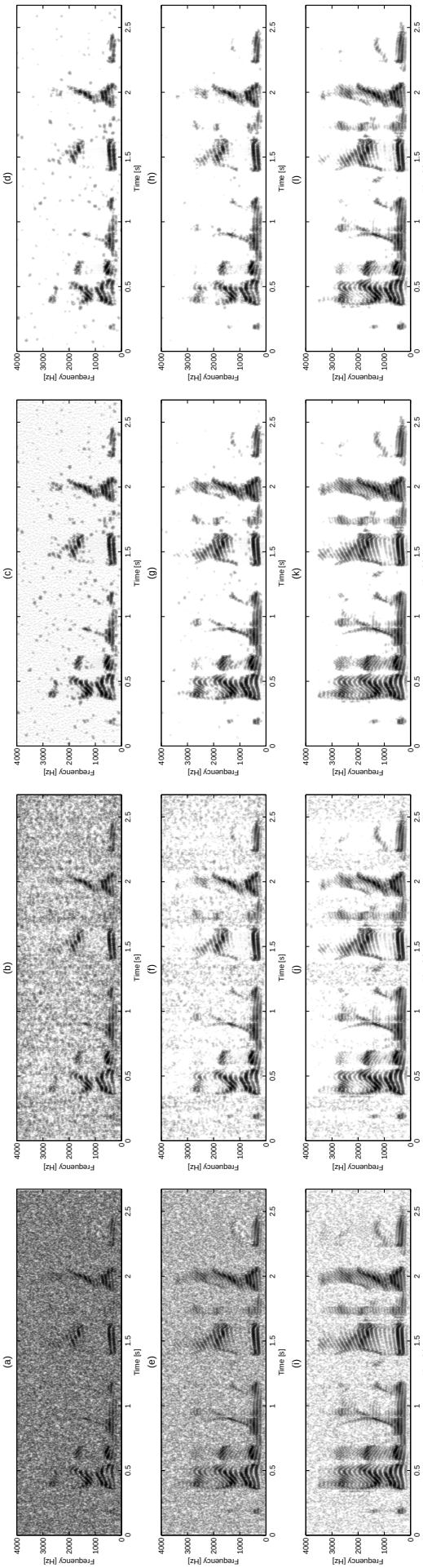

Figure 6: *Enhancement spectrograms of the sp10.wav utterance from the NOIZEUS corpus: row one (a, b, c, d) corresponds to degraded speech at 0dB SNR; row two (e, f, g, h) corresponds to degraded speech at 10dB SNR; row three (i, j, k, l) corresponds to degraded speech at 20dB SNR; column one (a, e, i) shows A2 stimuli; column two (b, f, j) shows D2 stimuli; column three (c, g, k) shows C5 stimuli; column four (d, h, l) shows F5 stimuli.*

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

References

Allen, J. and L. Rabiner (1977). A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE 65*(11), 1558–1564.

Alsteris, L. (2005). *Short-time phase spectrum in human and automatic speech recognition*. Ph. D. thesis, Griffith University, Brisbane, Australia.

Alsteris, L. and K. Paliwal (2006). Further intelligibility results from human listening tests using the short-time phase spectrum. *Speech Communication 48*(6), 727–736.

Berouti, M., R. Schwartz, and J. Makhoul (1979). Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'79)*, Washington, DC, USA, pp. 208–211.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing ASSP-27*(2), 113–120.

Crochiere, R. (1980). A weighted overlap-add method of short-time Fourier analysis / synthesis. *IEEE Trans. Acoust., Speech, Signal Processing ASSP-28*(2), 99–102.

Denda, Y., T. Nishiura, and H. Kawahara (2003). Speech enhancement with microphone array and Fourier / wavelet spectral subtraction in real noisy environments. In *Proc. European Conf. Speech Communication and Technology (EUROSPEECH'03)*, Geneva, Switzerland, pp. 2153–2156.

Dolph, C. (1946). A current distribution for broadside arrays which optimizes the relationship between beam width and sidelobe level. *Proc. IRE 34*, 335–348.

Goh, Z., K.-C. Tan, and B. T. G. Tan (1998). Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Transactions on Speech and Audio Processing 6*, 287–292.

Griffin, D. and J. Lim (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing ASSP-32*(2), 236–243.

Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE 66*(1), 51–83.

Hu, Y. and P. Loizou (2006a). Evaluation of objective measures for speech enhancement. In *Proc. Intern. Conf. Spoken Language Processing (ICSLP'06)*, pp. 1447–1450.

Hu, Y. and P. Loizou (2006b). Subjective comparison of speech enhancement algorithms. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'06)*, Volume 1, Toulouse, France, pp. 153–156.

Kabal, P. (2005). Time windows for linear prediction of speech, Version 2. Technical report, Department Electrical & Computer Engineering, McGill University, Montreal, Quebec, Canada.

Kamath, S. and P. Loizou (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02)*.

Lim, J. and A. Oppenheim (1979). Enhancement and bandwith compression of noisy speech. *IEEE Proceedings 67*(12), 1586–1604.

Oppenheim, A. and R. Schafer (1989). *Discrete-time signal processing*. Prentice Hall Signal Processing Series. London, England: Prentice Hall.

Paliwal, K. (2003). Usefulness of phase in speech processing. In *Proc. IPSJ Spoken Language Processing Workshop*, Gifu, Japan, pp. 1–6.

Paliwal, K. and L. Alsteris (2003). Usefulness of phase spectrum in human speech perception. In *Proc. European Conf. Speech Communication and Technology (EUROSPEECH'03)*, Geneva, Switzerland, pp. 2117–2120.

Paliwal, K. and L. Alsteris (2005). On the usefulness of STFT phase spectrum in human listening tests. *Speech Communication 45*(2), 153–170.

Portnoff, M. (1981). Short-time Fourier analysis of sampled speech. *IEEE Trans. Acoust., Speech and Signal Processing ASSP-29*(3), 364–373.

Quackenbush, S., T. Barnwell, and M. Clements (1988). *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice Hall.

Rix, A., J. Beerends, M. Hollier, and A. Hekstra (2001a). Perceptual Evaluation of Speech Quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'01)*, Volume 2, Salt Lake City, Utah, USA, pp. 749–752.

Rix, A., J. Beerends, M. Hollier, and A. Hekstra (2001b). Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T recommendation P.862.

Shannon, B. and K. Paliwal (2006). Role of phase estimation in speech enhancement. In *Proc. Intern. Conf. Spoken Language Processing (ICSLP'06)*, Pittsburgh, PA, USA, pp. 1423–1426.

Smith, S. (1999). *The Scientist & Engineer's Guide to Digital Signal Processing* (2nd ed.). San Diego, CA, USA: California Technical Publishing.

Vary, P. (1985). Noise suppression by spectral magnitude estimation – mechanism and theoretical limits. *Signal Processing 8*(4), 387–400.

Wang, D. and J. Lim (1982). The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Processing 30*, 679–681.

Whipple, G. (1994). Low residual noise speech enhancement utilizing time-frequency filtering. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'94)*, Montreal, Quebec, Canada, pp. 5–8.

Wolfel, M. and J. McDonough (2005). Minimum variance distortionless response spectral estimation. *Signal Processing Magazine, IEEE 22*(5), 117–126.

Yang, W. (1999). *Enhanced Modified Bark Spectral Distortion (EMBSD): an objective speech quality measure based on audible distortion and cognition model*. Ph. D. thesis, Temple University, Philadelphia, PA, USA.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review