ADAPTIVE TWO-BAND SPECTRAL SUBTRACTION WITH MULTI-WINDOW SPECTRAL ESTIMATION

Chuang He

Los Alamos National Laboratory
T Division, MS B276
Los Alamos, NM 87545
Email: he@busco.lanl.gov

George Zweig †

Signition, Inc.
901 18th Street
Los Alamos, NM 87544
Email: zweig@signition.com

ABSTRACT

An improved spectral subtraction algorithm for enhancing speech corrupted by additive wideband noise is described. The artifactual noise introduced by spectral subtraction that is perceived as musical noise is 7 dB less than that introduced by the classical spectral subtraction algorithm of Berouti et al. Speech is decomposed into voiced and unvoiced sections. Since voiced speech is primarily stochastic at high frequencies, the voiced speech is high-pass filtered to extract its stochastic component. The cut-off frequency is estimated adaptively. Multi-window spectral estimation is used to estimate the spectrum of stochastically voiced and unvoiced speech, thereby reducing the spectral variance. A low-pass filter is used to extract the deterministic component of voiced speech. Its spectrum is estimated with a single window. Spectral subtraction is performed with the classical algorithm using the estimated spectra. Informal listening tests confirm that the new algorithm creates significantly less musical noise than the classical algorithm.

1. INTRODUCTION

In the past two decades a variety of speech enhancement algorithms have been proposed. Spectral subtraction [2] is an algorithm that has been extensively studied because of its simplicity and effectiveness.

However, the original spectral subtraction algorithm [2] introduces a perceptually annoying artifact which is commonly referred to as musical noise. The musical noise is caused by large statistical fluctuations in the spectral estimate of the noisy speech. In [1], the authors proposed an algorithm to reduce the level of perceived musical noise by subtracting an overestimate of the noise spectrum and preventing the resultant spectral components from going below a preset minimum value. By preventing the resultant spectral components from going below a preset minimum value, the level of perceived musical noise is reduced, but background noise remains. When a high signal-to-noise ratio (SNR) is required, and the preset minimum value must be reduced, the unmasked musical noise becomes distracting. Therefore it is desirable to develop an algorithm that reduces the level of musical noise. This can be achieved by reducing the variance of the spectral estimate of the stochastic component of speech with spectral smoothing.

The stochastic component consists of the high frequency part of voiced speech [6, 7] and unvoiced speech. In this paper the stochastic component of voiced speech is adaptively extracted with a high-pass filter. The spectrum of the stochastic component of speech—voiced speech at high frequencies and unvoiced speech—is smoothed with Thomson's method of multi-window spectral estimation (MWSE) [8]. The resulting reduction of variance reduces musical noise, and therefore reduces the background noise necessary for masking.

The spectrum of the deterministic component of speech —low-pass filtered voiced speech—is estimated with a single window. Multiple windows would damage the harmonic structure and decrease intelligibility.

2. STOCHASTICITY OF VOICED SPEECH

It is generally believed that there is a stochastic component present in the excitation function of voiced speech [4, 7]. As a result, the higher formants of voiced speech are excited randomly, not periodically. In [6] we defined a quantity called normalized variance that measures the frequency dependence of the relative strengths of the stochastic and deterministic components of speech. We showed that for voiced vowels the normalized variance is small at low frequencies, confirming the deterministic nature of speech. At high frequencies, the normalized variance reaches its maximum value of 1, indicating that voiced speech is primarily stochastic above some cut-off frequency f_c. Figure 1 shows the frequency dependence of the estimated normalized variance of one utterance of the vowel /I/. This utterance is essentially stochastic above 4 kHz. We also observed that the cut-off frequency (f_c) above which voiced speech becomes stochastic varies with pitch, phoneme, and speaker. Therefore in the new denoising algorithm, this cut-off frequency is estimated adaptively for every frame classified as voiced.

* This work was supported by the DARPA Information Technology Office and the DOE Applied Mathematics Program.

† George Zweig is also affiliated with Los Alamos National Laboratory, T Division, MS B276, Los Alamos, NM 87545.

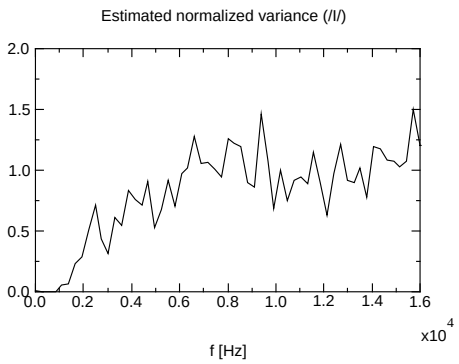

Figure 1: Estimated normalized variance of an utterance of the vowel /I/. Values of 0 and 1 indicate entirely deterministic and stochastic speech, respectively. Despite statistical fluctuations in the estimate, it is clear that this utterance is stochastically excited at high frequencies.

3. EXPLOITING THE STOCHASTICITY OF SPEECH

In [7] the stochasticity of voiced speech was exploited to improve the performance of speech compression and synthesis algorithms. In [4], the author introduced the Dual Excitation model which represents speech as the sum of a voiced and an unvoiced component. The model was later applied to speech enhancement and the fundamental frequency and harmonic amplitudes of the voiced component of speech were estimated using a minimum mean-squared error approach [5, 9]. The voiced component was then constructed from these estimated parameters. The unvoiced component was obtained by subtracting the estimated voiced component from the speech signal. Different speech enhancement algorithms were applied to the voiced and unvoiced components. In this paper the stochastic component of speech is extracted by a different and computationally simpler algorithm (high-pass filtering), and a more sophisticated method of spectral smoothing is employed (MWSE).

3.1. Overview

For each frame, the energy of noisy speech is compared to a threshold to classify the frame as unvoiced (including silence) or voiced. For voiced speech, an algorithm similar to the one described in [7] is used to determine the cut-off frequency (f_c) above which speech in this frame is stochastic. The voiced speech is then divided into two bands using linear phase FIR filters with cut-off frequency f_c.

Both the high-pass part of voiced speech and unvoiced speech are stochastic. MWSE is used to estimate their spectra, thereby reducing the variance of the spectral estimates. A single window (the first discrete prolate spheroidal sequence) is used to determine the spectrum of the low-pass deterministic part of voiced speech, thereby preserving its harmonic structure.

Once the spectral estimates are obtained, the algorithm described in [1] is used to enhance the speech with spectral subtraction.

3.2. Cut-off frequency (f_c) estimation

For each frame classified as voiced, the fundamental frequency is estimated with a peak-picking algorithm performed on a smoothed spectral estimate. Here the goal of spectral smoothing is to reduce the variance in the spectral estimate near the harmonic frequencies. The cut-off frequency is then determined as the highest frequency below which the separation between adjacent peaks is approximately equal to the fundamental frequency. Only peaks that are significantly greater than background are considered and small gaps in the harmonic structure are ignored. This estimated cut-off frequency is smoothed by a median filter operating on three consecutive frames and rounded upward to the nearest multiple of 500 Hz. Pre-designed low-pass and high-pass Parks-McClellan optimal linear phase FIR filters with cut-off frequencies at multiples of 500 Hz are used to separate the stochastic and deterministic components of voiced speech.

3.3. Multi-window spectral estimation

Thomson's method [8] of multi-window spectral estimation (MWSE) is used to estimate the spectrum of high-passed voiced speech and unvoiced speech in order to minimize the variance of the spectral estimates. For a given spectral resolution, multi-window spectral estimation entails computing $K = 2NW - 1$ individual estimates of the spectrum with discrete prolate spheroidal sequence windows [8], and then combining these estimates to form a single spectral estimate. Here N is the number of points in a window and W is the frequency resolution of the spectral estimate. If the spectrum is flat within the frequency interval $[\Omega - W, \Omega + W]$ centered about frequency Ω, then the variance of the spectral estimate is reduced by a factor of K with respect to that of a single estimate. For fixed N and W, the variance of a multi-window spectral estimate is smaller than that of other spectral smoothing techniques, e.g., the weighted overlapped segment averaging spectral estimator [3].

4. SPEECH ENHANCEMENT

The new speech enhancement system is summarized in Figure 2. High-passed-voiced and unvoiced speech are enhanced by the same algorithm. A different algorithm is used to enhance low-passed voiced speech.

4.1. Enhancing low-passed voiced speech

In the low-frequency band of voiced speech, the spectrum is estimated from a windowed fast Fourier transform (FFT). The first discrete prolate spheroidal sequence is used for the window.

We adopt the algorithm described in [1] to perform spectral subtraction. The denoised low-passed voiced speech signal is given by

$$r_{d,l}[n] = \text{IFFT}\left\{\sqrt{\hat{S}_l[m]} \cdot e^{j\theta_l[m]}\right\}, \qquad (1)$$

where IFFT$\{\cdot\}$ denotes the inverse fast Fourier transform, $\theta_l[m]$ is the phase of the FFT of the windowed low-passed

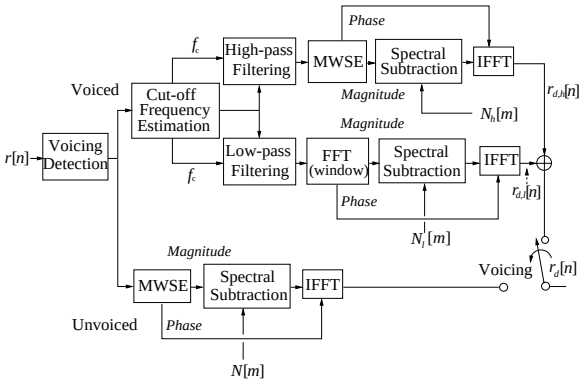

Figure 2: Adaptive two-band spectral subtraction system. The noisy and denoised speech samples are denoted by $r[n]$ and $r_d[n]$.

noisy speech, and

$$\hat{S}_l[m] = \begin{cases} R_l[m] - \alpha N_l[m], & \text{if} \quad R_l[m] > (\alpha+\beta)N_l[m] \\ \beta N_l[m], & \text{otherwise.} \end{cases} \tag{2}$$

The frequency and time indices are m and n, respectively. $R_l[m]$ is the squared magnitude of the FFT of the windowed low-passed noisy speech, $N_l[m]$ is the spectral estimate of the low-passed noise obtained during silences, α is a positive "subtraction factor," and β is a positive "spectral floor parameter" [1]. The subtraction factor α decreases with the segmental SNR in a manner specified in [1]. Its value at 0 dB segmental SNR is denoted by α_0, and together with the values for β, is given in Table 1. MWSE is used to compute $N_l[m]$.

4.2. Enhancing high-passed-voiced and unvoiced speech

For high-passed-voiced and unvoiced speech, MWSE is used to reduce the variance of the spectral estimate. Specifically, $R_l[m]$ in equation (2) is replaced by

$$R_h[m] = \sum_{k=1}^{K} \gamma_h^k[m] R_h^k[m] \quad \text{or} \quad R[m] = \sum_{k=1}^{K} \gamma^k[m] R^k[m], \tag{3}$$

where $R_h^k[m]$ and $R^k[m]$ are the squared magnitudes of the FFT of the windowed high-passed-voiced and unvoiced noisy speech with the kth discrete prolate spherical sequence as the window, $\gamma_h^k[m]$ and $\gamma^k[m]$ are the frequency-dependent weighting factors calculated adaptively [8], and K is the number of windows. Values for K are given in Table 1. N and W are chosen to be 512 (the sampling frequency is 16000 Hz) and 1/256. The number of windows used is as large as possible, subject to the constraint that the intelligibility of speech is not noticeably degraded.

The phase $\theta_l[m]$ is replaced by the phase corresponding to $R_h^1[m]$ or $R^1[m]$. The noise spectral estimate $N_l[m]$ is replaced either by $N_h[m]$, the spectral estimate of the high-passed noise, or $N[m]$, the spectral estimate of the wide-band noise. Both $N_h[m]$ and $N[m]$ are computed during silences using MWSE.

Short gaps (silences) can be created where quiet unvoiced vowels exist in the clean speech. In this case, short quiet noise bursts are added to fill the gaps. This helps most when $\beta = 0$.

		Classical algorithm of Berouti et al.	New algorithm (low-passed voiced speech)	New algorithm (high-passed voiced speech & unvoiced speech)
α_0		4	4	3.25
β	SNR = 5 dB	0.005	0.001	0.001
	SNR = 0 dB	0.02	0.004	0.004
	SNR = -5 dB	0.04	0.008	0.008
K		1	1	3

Table 1: The spectral subtraction and spectral estimation parameters.

5. RESULTS

Figure 3 shows spectrograms of clean and noisy TIMIT speech. For the noisy speech, the average segmental SNR is 0 dB. The noise is white.

Figure 4 shows spectrograms of speech denoised with the algorithm of [1], and with the new algorithm, both with $\beta = 0$. The musical noise, which appears as small dark "islands," is clearly visible in the top panel, but is less noticeable in the middle panel, corresponding to a reduction of 7 dB. The estimated cut-off frequency f_c is shown at the bottom of the Figure. Note that there are significant portions of speech that lie above f_c.

To mask the remaining musical noise (which some listeners prefer), the values of β given in Table 1 are used.

Informal listening tests indicate that the intelligibility of speech processed by both algorithms is about the same, but the speech denoised by the new algorithm contains significantly less musical noise.

All figures were created with the SigniScope^{R} and are plotted on a compressed z-axis scale.

6. CONCLUSIONS

An adaptive two-band spectral subtraction algorithm is described. Gains over conventional algorithms come primarily from exploiting the stochasticity of speech. The stochastic component consists of the high frequency part of voiced speech and unvoiced speech. The stochastic component of voiced speech is adaptively extracted with a high-pass filter. The spectrum of the stochastic component of speech—voiced speech at high frequencies and unvoiced speech—is smoothed with Thomson's method of multi-window spectral estimation. The resulting reduction of variance reduces both the musical noise and the background noise necessary for masking musical noise by approximately 7 dB.

7. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In Proceedings of ICASSP, pages 208–211, 1979.

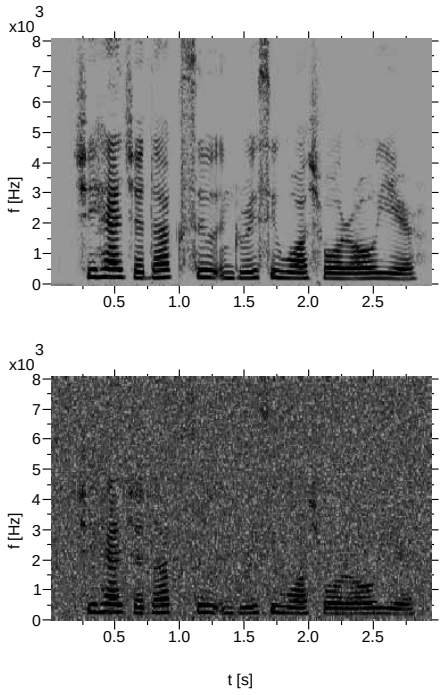

Figure 3: Top: Spectrogram of clean speech. Bottom: Spectrogram of noisy speech (average segmental SNR = 0 dB).

[2] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Processing, 27(2):113–120, April 1979.

[3] T. Bronez. On the performance advantage of multitaper spectral analysis. IEEE Trans. Signal Processing, 40(12):2941–2946, December 1992.

[4] J. Hardwick. The Dual Excitation Speech Model. Ph.D. thesis, MIT, EECS Department, June 1992.

[5] J. Hardwick, C. D. Yoo, and J. S. Lim. Speech enhancement using the dual excitation speech model. In Proceedings of ICASSP, pages 367–370, 1993.

[6] C. He and G. Zweig. Determination of the stochasticity of the excitation function of speech. To be presented at 136th Meeting of Acoustical Society of America, 1998.

[7] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins. A mixed-source model for speech compression and synthesis. J. Acoust. Soc. Amer., 64(6):1577–1581, December 1978.

[8] D. J. Thomson. Spectrum estimation and harmonic analysis. Proc. IEEE, 70(9):1055–1096, September 1982.

[9] C. D. Yoo and J. S. Lim. Speech enhancement based on the generalized dual excitation model with adaptive analysis window. In Proceedings of ICASSP, pages 832–835, 1995.

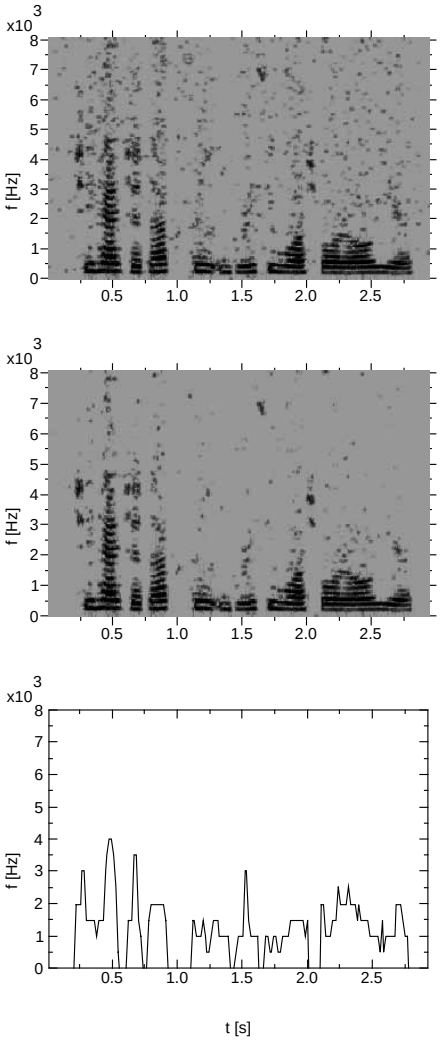

Figure 4: Top: Spectrogram of denoised speech, algorithm of [1], β=0. Middle: Spectrogram of denoised speech, new algorithm, β=0. Bottom: Estimated cut-off frequencies f_c.