A MULTI-BAND SPECTRAL SUBTRACTION METHOD FOR ENHANCING SPEECH CORRUPTED BY COLORED NOISE

Sunil D. Kamath and Philipos C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688
{skamath, loizou}@utdallas.edu

ABSTRACT

The spectral subtraction method is a well-known noise reduction technique. Most implementations and variations of the basic technique advocate subtraction of the noise spectrum estimate over the entire speech spectrum. However, real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. In this paper, we propose a multi-band spectral subtraction approach which takes into account the fact that colored noise affects the speech spectrum differently at various frequencies. This method outperforms the standard power spectral subtraction method resulting in superior speech quality and largely reduced musical noise.

1. INTRODUCTION

The spectral subtraction method as proposed by Boll [1] is a popular noise reduction technique due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. The technique is based on the direct estimation of the short-term spectral magnitude. The basic principle of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal.

The conventional power spectral subtraction method substantially reduces the noise levels in the noisy speech. However, it also introduces an annoying distortion in the speech signal called musical noise. Due to the inaccuracies in the short-time noise spectrum estimate, large spectral variations exist in the enhanced spectrum causing these distortions. Also, occasional negative estimates of the enhanced power spectrum can occur. In such cases, the negative spectral components are floored to zero or to some minimal value, causing further distortions in the time signal.

Recent studies have focused on a non-linear approach to the subtraction procedure [2] [3] [4] [5]. This approach has been justified due to the variation of signal-to-noise ratio across the speech spectrum. Unlike white gaussian noise, which has a flat spectrum, the spectrum of real-world noise is not flat. Thus, the noise signal does not affect the speech signal uniformly over the whole spectrum. Some frequencies are affected more adversely than others. In multi-talker babble, for instance, the low frequencies, where most of the speech energy resides, are affected more than the high frequencies. Hence it becomes imperative to estimate a suitable factor that will subtract just the necessary amount of the noise spectrum from each frequency bin (ideally), to prevent destructive subtraction of the speech while removing most of the residual noise.

In this paper, we propose a multi-band approach to the spectral subtraction method that accomplishes just that, i.e., it reduces the above-mentioned distortions to a large extent while maintaining a high level of speech quality.

Section II presents the proposed approach, section III describes the implementation of the proposed method and section IV gives the experimental results. Conclusions and comments are given in section V.

2. MULTI-BAND SPECTRAL SUBTRACTION

Assuming the additive noise to be stationary and uncorrelated with the clean speech signal, the resulting corrupted speech can be expressed as:

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

where $y(n), s(n)$ and $d(n)$ are the corrupted speech signal, clean speech signal and the noise respectively. The power spectrum of the corrupted speech can be approximately estimated as:

$$|Y(k)|^2 \approx |S(k)|^2 + |D(k)|^2 \tag{2}$$

where $S(k)$ and $D(k)$ are the magnitude spectra of the clean speech and the noise respectively. Since the noise spectrum cannot be directly obtained, an estimate $\hat{D}(k)$ is calculated during periods of silence. Since Boll's original work, many

different variations of the spectral subtraction have been proposed [2] [3] [5] [6] [7]. Most, if not all, implementations of the spectral subtraction approach are variants of the approach proposed by Berouti et.al. [6]. In the implementation proposed by Berouti et.al. [6], the estimate of the clean speech spectrum is obtained as:

$$|\hat{S}(k)|^2 = |Y(k)|^2 - \alpha|\hat{D}(k)|^2 \quad (3)$$

where α is an over-subtraction factor [6] which is a function of the segmental SNR. This implementation assumes that the noise affects the speech spectrum uniformly and the over-subtraction factor α subtracts an over-estimate of the noise over the whole spectrum. That is not the case, however, with real-world noise (e.g., car noise, cafeteria noise, etc.). This is best illustrated in Figure 1 showing the segmental SNR estimated for four (linearly-spaced) frequency bands of speech corrupted by speech-shaped noise. In this particular example, the segmental SNR of the high frequency band (band 4) was significantly lower than the SNR of the low frequency band (band 2), by as much as 15 dB in some cases.

To take into account the fact that colored noise affects the speech spectrum differently at various frequencies, we propose a multi-band approach to spectral subtraction. The speech spectrum is divided into N non-overlapping bands, and spectral subtraction is performed independently in each band. So, the estimate of the clean speech spectrum in the ith band is obtained by:

$$|\hat{S}_i(k)|^2 = |Y_i(k)|^2 - \alpha_i\delta_i|\hat{D}_i(k)|^2 \quad b_i \le k \le e_i \quad (4)$$

where b_i and e_i are the beginning and ending frequency bins of the ith frequency band, α_i is the over-subtraction factor of the ith band and δ_i is a tweaking factor that can be individually set for each frequency band to customize the noise removal properties. The band specific oversubtraction factor α_i is a function of the segmental SNR_i of the ith frequency band which is calculated as:

$$SNR_i(dB) = 10\log_{10}\left(\frac{\sum_{k=b_i}^{e_i}|Y_i(k)|^2}{\sum_{k=b_i}^{e_i}|\widehat{D}_i(k)|^2}\right) \quad (5)$$

Using the SNR_i value calculated in Eq. (5), α_i can be determined as:

$$\alpha_i = \begin{cases} 5 & SNR_i < -5 \\ 4 - \frac{3}{20}(SNR_i) & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \quad (6)$$

While the use of the over-subtraction factor α_i provides a degree of control over the noise subtraction level in each band, the use of multiple frequency bands and the use of the

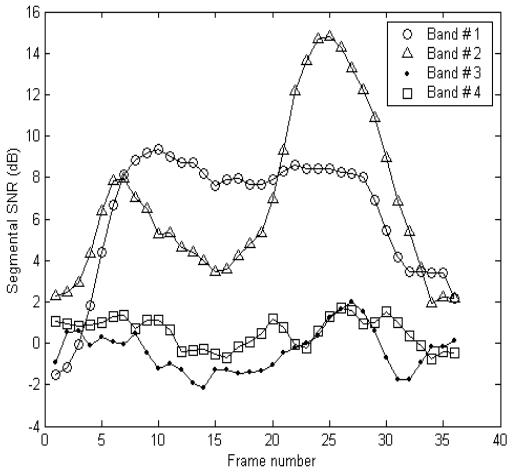

Fig. 1. The segmental SNR of four (linearly-spaced) frequency bands of speech corrupted by speech-shaped noise.

δ_i weights provide an additional degree of control within each band.

The negative values in the enhanced spectrum in Eq. 4 were floored to the noisy spectrum as:

$$|\hat{S}_i(k)|^2 = \begin{cases} |\hat{S}_i(k)|^2 & |\hat{S}_i(k)|^2 > 0 \\ \beta|Y_i(k)|^2 & else \end{cases} \quad (7)$$

where the spectral floor parameter was set to $\beta = 0.002$.

3. IMPLEMENTATION

Speech was Hamming windowed using a 20-ms window and a 10-ms overlap between frames. The Fast Fourier Transform (FFT) of the windowed speech was smoothed as per [8] and a weighted spectral average [9] [10] is taken over preceding and succeeding frames of data as:

$$\bar{Y}_j(k) = \sum_{l=-M}^{M} W_l Y_{j-l}(k) \quad (8)$$

where j is the frame index. The number of frames, M, was limited to 2 to prevent spectral smearing. The filter weights W_l were empirically determined and set to $W = [0.09, 0.25, 0.32, 0.25, 0.09]$.

The values for δ_i in Eq. 4 were empirically determined and set to:

$$\delta_i = \begin{cases} 1 & f_i \le 1\ kHz \\ 2.5 & 1\ kHz < f_i \le \frac{F_s}{2} - 2\ kHz \\ 1.5 & f_i > \frac{F_s}{2} - 2\ kHz \end{cases} \quad (9)$$

where f_i is the upper frequency of the ith band, and F_s is the sampling frequency. The motivation for using smaller δ_i

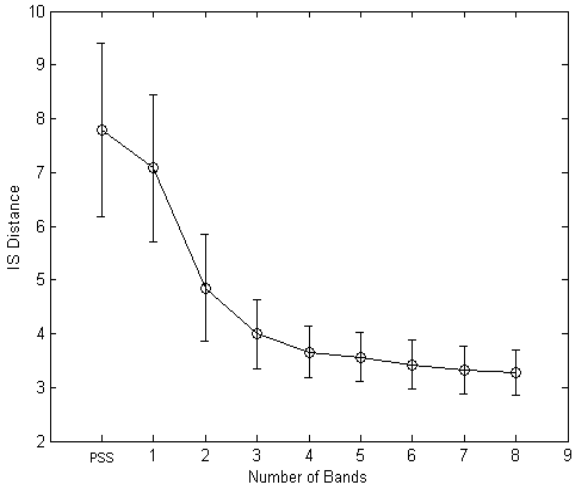

Fig. 2. The performance, in terms of mean IS distance measure, of the multi-band spectral subtraction approach as a function of the number of bands for 10 sentences embedded in speech-shaped noise at 5 dB SNR. The performance obtained with power spectral subtraction is indicated with 'PSS'. The error bars indicate standard deviations.

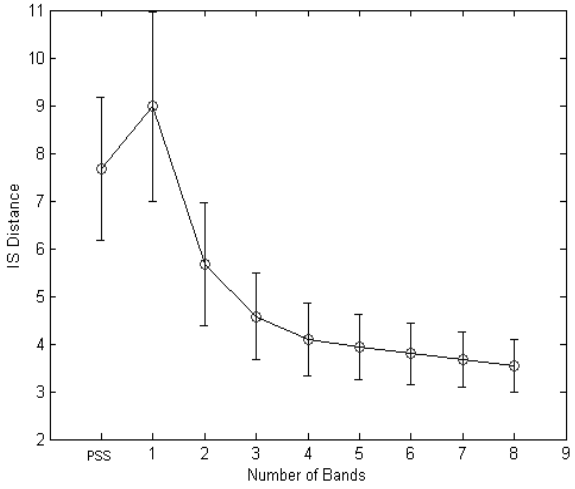

Fig. 3. The performance, in terms of mean IS distance measure, of the multi-band spectral subtraction approach as a function of the number of bands for 10 sentences embedded in speech-shaped noise at 0 dB SNR. The performance obtained with power spectral subtraction is indicated with 'PSS'. The error bars indicate standard deviations.

values for the low frequency bands, is to minimize speech distortion, since most of the speech energy is present in the lower frequencies. Relaxed subtraction was also used for the high frequency bands.

The enhanced spectrum within each band (Eq. 4) is combined, and the enhanced signal is obtained by taking the inverse Fourier transform of the enhanced spectrum using the phase of the original noisy spectrum. Finally, the standard overlap-and-add method is used to obtain the enhanced signal.

4. EXPERIMENTAL RESULTS

Ten sentences from the HINT (Hearing In Noise Test) database [11] uttered by a male speaker were used to evaluate the proposed multi-band spectral subtraction approach. Speech-shaped noise at 5 dB and 0 dB SNR was added to the sentences after downsampling them to 8 kHz. This noise was generated from the long-term spectrum of all the sentences in the database and resembles the spectral characteristics of the male speaker. The Itakura-Saito (IS) distance method was used as the objective measure to evaluate the performance of the algorithm. The highest 5% of the IS distance values were discarded, as suggested in [12], to exclude unrealistically high spectral distance values. This method ensured a reasonable overall measure of performance.

To determine the optimal (in terms of speech quality) number of bands, we varied the number of bands from 1 to 8 and examined speech performance using both objective and subjective measures. Linear frequency spacing was used. Figures 2 and 3 plot the mean IS distance values for 10 sentences at 5 and 0 dB SNR respectively, as a function of the number of bands used. For comparative purposes, we also plot the performance of the traditional power spectral subtraction (PSS) method as implemented by Berrouti *et al.* [6]. The proposed multi-band spectral subtraction approach consistently outperformed the PSS approach for both SNRs. Note that when the total number of bands is one, then our approach reduces to the traditional power spectral subtraction approach [6]. This is also seen in Figures 2 and 3. The mean IS measure for the PSS approach was almost identical to the IS distance obtained with one band. The small difference can be attributed to differences in pre-processing.

The IS distance shows marked improvement when the number of bands increased from 1 to 4. The improvement in speech quality is also marked. While the IS distance does show a slight increase in performance for higher number of bands, there was no perceivable improvement in speech quality. Informal listening tests indicated that the multi-band approach yielded very good speech quality with very little trace of musical noise and with minimal, if any, speech distortion. The lack of muical noise can also be seen in Figure 4, which shows the spectrograms of enhanced speech obtained with multi-band spectral subtraction (4 bands) and

enhanced speech obtained with power spectrum subtraction.

5. CONCLUSIONS

The multi-band spectral subtraction method provides a definite improvement over the conventional power spectral subtraction method. We believe that the improvement is due to the fact that the multi-band approach takes into account the non-uniform effect of colored noise on the spectrum of speech. The added computational complexity of the algorithm is minimal. We found that four linearly-spaced frequency bands were adequate in obtaining good speech quality.

6. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.27, pp. 113-120, Apr. 1979.

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, Vol. 11, Nos. 2-3, pp. 215-228, 1992.

[3] I. Soon, S. Koh and C. Yeo, "Selective magnitude subtraction for speech enhancement," *Proceedings. The Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region*, vol.2, pp. 692-695, 2000.

[4] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free application," *International Conference on Consumer Electronics*, vol. 2, pp. 220-221, 2001.

[5] C. He and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation," *ICASSP*, vol.2, pp. 793-796, 1999.

[6] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 208-211, Apr. 1979.

[7] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, pp 126-137, vol. 7, March 1999.

[8] L. Arslan, A. McCree and V. Viswanathan, "New methods for adaptive noise suppression," *ICASSP*, vol.1, pp. 812-815, May 1995.

[9] Y. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *ICASSP*, vol.2, pp. 961-964, Apr. 1991.

[10] J. Deller, Jr., J. Hansen and J. Proakis, Discrete-Time Processing of Speech Signals, *NY: IEEE Press*,2000.

[11] M. Nilsson, S. Soli and J. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol.95, pp. 1085-1099, 1994.

[12] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancements algorithms," *Inter. Conf. On Spoken Language Processing*, vol.7, pp. 2819-2822, Sydney, Australia, Dec.1998.

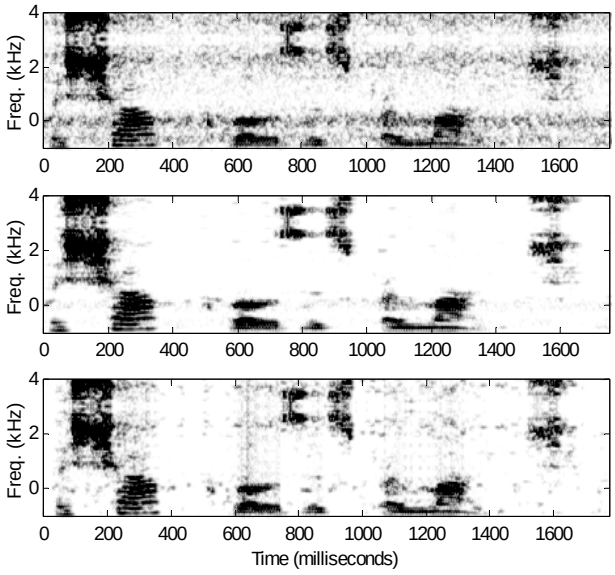

Fig. 4. Spectrogram of the sentence "The shop closes for lunch." at 5 dB SNR. The top spectrogram is the corrupted signal, the middle spectrogram is the enhanced signal obtained by the multi-band spectral subtraction method using 4 linearly-spaced frequency bands, and the bottom spectrogram is the enhanced signal obtained by the power spectral subtraction method.