Journal of Mobile Communication 1 (3): 97-100, 2007
ISSN: 1990-794X

A Frequency-Dependent Speech Enhancement Methods

[1,2]C. Boubakir, [2]D. Berkani and [3]F. Grenez
[1]LAMEL, Faculté des Sciences de l'Ingénieur, Université de Jijel,
BP. 98, Ouled Aissa, 18000, Jijel, Algérie
[2]Laboratoire Signal et Communications, Département d'Electronique,
Ecole Nationale Polytechnique, 10, av. Hassen Badi, El Harrach, Alger, Algérie
[3]Service Ondes et Signaux, Université Libre de Bruxelles,
Belgique ULB Campus du Solboch, Avenue Paul Héger, Bâtiment U

Abstract: The corruption of speech due to presence of additive background noise causes severe difficulties in various communication environments. Most implementations and variations of the basic spectral subtraction 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. This study explores a Multi-Band Spectral Subtraction (MBSS) approach with suitable pre-processing of the speech signal. Speech is processed into N ($1 \le N \le 8$) frequency bands and spectral subtraction is performed independently on each band using band-specific over-subtraction factors.

Key words: Communication environment, spectrum, speed, enhancement methods, PSS, MBSS

INTRODUCTION

Among the many available single channel speech enhancement algorithms, spectral subtraction has been one of the most popular methods for its simplicity and effectiveness. The limitation of the basic Power Spectral Subtraction (PSS) method is that is often results in excessive remnant residual noise and musical noise. Residual noise refers to the broadband noise that has the same perceptual characteristics as the original noise. Musical noise refers to the synthetic musical tones due to the presence of random short-duration spectral peaks in the remnant noise spectrum. Several studies have been recently proposed for the modifications of the basic method to minimize the residual noise and musical noise artifacts.

Power spectral subtraction: Spectral subtraction is a method for restoration of the power or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. It is the most common of the subtractive type algorithms, which form a family of methods based on subtraction of the noise estimate from the original speech (Boll, 1979; Berouti *et al*., 1979; Sim *et al*., 1998; Crozier *et al*., 1980). These systems form a category of algorithms that operate in the frequency domain. The noise spectrum is estimated and updated, from the periods when the signal is absent and only the noise is present. The assumption being that noise is stationary or a slowly varying process and that the noise spectrum does not change significantly between the updating periods. For restoration of the time-domain signal, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy signal and then transformed via an inverse discrete Fourier transform to the time domain. The phase of the noisy signal is not modified, as not only is it hard to get an estimation of the phase as compared to the magnitude spectrum, it is also believed that from perceptual point of view the phase does not carry any useful information for noise suppression.

Thus, if we assume that y(n), the discrete noise corrupted input signal, is composed of the clean speech signal s(n) and d(n) the uncorrelated additive noise signal, then it the noisy signal can be represented as:

$$y(n) = s(n) + d(n) \quad (1)$$

The power spectrum of the clean speech can be approximately estimated as:

$$\left|\hat{S}(k)\right|^{\gamma} = a_k \left|Y(k)\right|^{\gamma} - b_k \left|\hat{D}(k)\right|^{\gamma} \quad (2)$$

Corresponding Author: C. Boubakir, LAMEL, Faculté des Sciences de l'Ingénieur, Université de Jijel,
BP. 98, Ouled Aissa, 18000, Jijel, Algérie

Where

$$\left|\hat{S}(k)\right|, \left|Y(k)\right|, \text{ and } \left|\hat{D}(k)\right|$$

refer to speech magnitude spectrum estimate, the noisy speech magnitude spectrum and noise magnitude spectrum estimate, respectively for an input speech frame and "k" is the frequency index.

When $\gamma = 1$, $a_k = b_k = 1$, the equation reduces to the basic spectral subtraction method proposed by Boll (1979) where the subtraction is carried out by subtracting the magnitude spectra.

The over-subtraction method proposed by Berouti *et al.* (1979) is obtained when we set $\gamma = 2$, $a_k = 1$ and b_k as the over-subtraction factor. Both methods use the same parameters for all frequency bins.

The algorithm of the generalized method given by (2) is modified as:

$$\left|\hat{S}(k)\right|^2 = \begin{cases} \left|Y(k)\right|^2 - \alpha\left|\hat{D}(k)\right|^2, \text{if } \left|\hat{S}(k)\right|^2 > \beta\left|\hat{D}(k)\right|^2 \\ \beta\left|\hat{D}(k)\right|^2, \text{Otherwise} \end{cases} \quad (3)$$

Where β is the spectral floor and the over-subtraction factor α is a function of the noisy signal-to-noise ratio and calculated as:

$$\alpha = \alpha_0 - \frac{3}{20}\text{SNR} \quad (4)$$

$$-5\text{dB} \le SNR \le 20\text{dB}$$

Here α_0 is the desired value of α at 0 dB SNR. The value for α has to be carefully chosen in order to prevent both the musical noise and too much signal distortion. For power subtraction, the optimal range of α_0 is between 3 and 6.

The introduction of spectral floor β reduces the amount of musical noise. For higher noise levels, β should be in the range of 0.1-0.01. For lower input noise levels, β can be chosen smaller than 0.01.

MULTI-BAND SPECTRAL SUBTRACTION

The Berouti *et al.* (1979) algorithm 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, Babble noise,...). These effects are best illustrated in the plots of the Power Spectral Density (PSD) of different noise signals. Figure 1 depicts the PSD of White Gaussian Noise (WGN), which is flat over the whole spectrum. Figure 2 illustrates similar plot for babble noise.

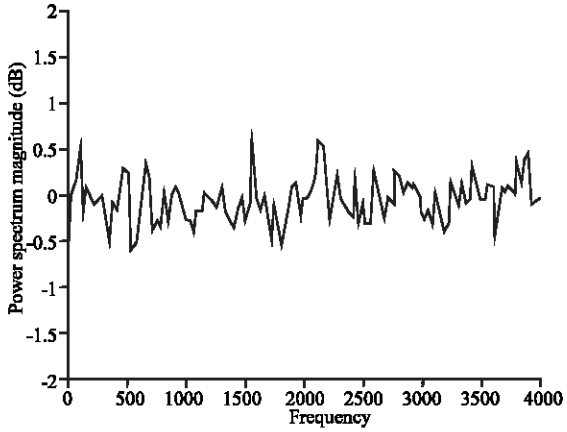

Fig. 1: PSD of WGN

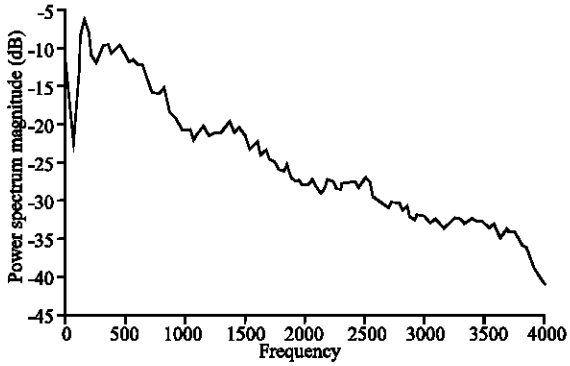

Fig. 2: PSD of babble noise

To take into account the fact that colored noise affects the speech spectrum differently at various frequencies, we use the multi-band approach to spectral subtraction (Sunil and Loizou, 2002). The speech spectrum is divided into N non-overlapping bands and spectral subtraction is performed independently in each band. The estimate of the clean speech spectrum in the i th band is obtained by:

$$\left|\hat{S}_i(k)\right|^2 = \left|Y_i(k)\right|^2 - \alpha_i\delta_i\left|\hat{D}_i(k)\right|^2, b_i \le k \le e_i \quad (5)$$

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 band-subtraction factor that can be individually set for each frequency band to customize the noise removal properties.

The band specific over subtraction factor α_i is a function of the segmental SNR_i of the ith frequency band which is calculated as:

$$SNR_i(dB) = 10\log_{10}\left(\frac{\sum_{k=b_i}^{e_i}\left|Y_i(k)\right|^2}{\sum_{k=b_i}^{e_i}\left|\hat{D}_i(k)\right|^2}\right) \quad (6)$$

Using the SNR_i value, α_i can be determined as:

$$\alpha_i = \begin{cases} 5 & SNR_i < -5 \\ 4-\frac{3}{20}SNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \quad (7)$$

The negative values in the enhanced spectrum in Eq. 5 were floored to the noisy spectrum as:

$$\left|\hat{S}_i(k)\right|^2 = \begin{cases} \left|\hat{S}_i(k)\right|^2 \text{ if } \left|\hat{S}_i(k)\right|^2 > 0 \\ \beta\left|Y_i(k)\right|^2 \quad \text{else} \end{cases} \quad (8)$$

Where the spectral floor parameter was set to β = 0.002.

RESULTS AND DISCUSSION

To measure quality of the enhanced signal, we have used the The Itakura-Saito (IS) measure and the segmental SNR (Quackenbush *et al*., 1998). Both the measures show high correlation with subjective quality.

The Itakura-Saito distance is given by,

$$d_{IS}\left(\vec{a}_d,\vec{a}_\phi\right) = \left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right] + \log\left(\frac{\sigma_d^2}{\sigma_\phi^2}\right) - 1 \quad (9)$$

Where $\vec{a}_\phi$ and $\vec{a}_d$ represent the Linear Prediction (LP) coefficient vectors for the clean and processed speech frame respectively and the σ_ϕ^2 all-pole gains.

The highest 5 % of the IS distance values were discarded, as suggested in (Quackenbush *et al*., 1998), to exclude unrealistically high spectral distance values. The lower the IS measure for an enhanced speech, the better is its perceived quality.

Segmental SNR measure: Since the correlation of SNR with subjective quality is so poor. Instead, we choose the frame-based segmental SNR by averaging frame level SNR estimates and is defined by (Quackenbush *et al*., 1998).

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{t=Nm}^{Nm+N-1} s_\phi^2(n)}{\sum_{t=Nm}^{Nm+N-1}\left[s_d(n)-s_\phi(n)\right]^2} \quad (10)$$

Where M denotes the number of frames. The lower and upper thresholds are selected to be -10 and +35 dB, respectively.

Implementation: The input noisy speech sampled at 8 kHz, is first windowed by a half-overlapped Hamming window of length 256 points and then spectrally decomposed by the FFT with the same length. For the two methods, instead of directly using the power spectra of the signal, a smoothed version of the power spectra can be used.

The power spectral subtraction is implemented as in Eq. 3, with the parameter values introduced in this study.

The parameter values that have been suggested by the authors (Sunil and Loizou, 2002) are used in our implementation. Four bands have been used with α_i (Eq. 7), β = 0.002 and the tweaking factor (δ_i) are fixed empirically as '1.0 for first band (0-1kHz)', '2.5 for second and third bands (1-3 kHz)' and '1.5 for the fourth band (3-4 kHz)'.

A critical component in any frequency domain enhancement algorithm is the estimation of the noise power spectrum. A common technique is to use a VAD (Nathalie, 1999) and update the estimated noise spectrum during periods of speech absence in the input signal.

For restoration of the time-domain signal, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy signal and then transformed via an inverse discrete Fourier transform to the time domain. The phase of the noisy signal is not modified.

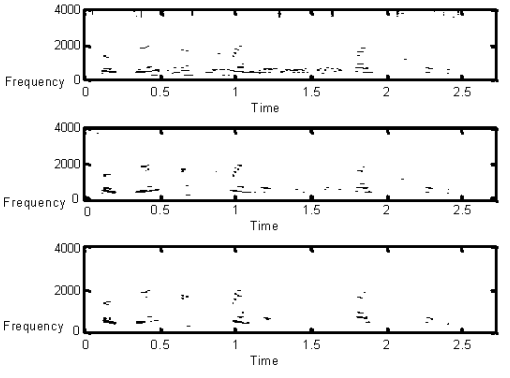

Fig. 3: Spectrogram of the sentence"" at SNR = 0 dB

Table 1: Object quality scores under babble noise

	5 dB SNR		0 dB SNR	
SNRseg	SNRseg	ISM	SNRseg	ISM
Noisy	-1.959	2.1899	-4.3883	2.8596
PSS	-0.7990	3.1451	-3.0690	3.9101
MultiB	-0.4610	6.7982	-2.5281	8.3674

Finally, the standard overlap-and-add method is used to obtain the enhanced signal.

Sentences from the noisy database (NOIZEUS) (http://www.utdallas.edu/~loizou/speech/noizeus) at 5 dB and 0 dB SNR were used to evaluate the two methods. From the informal listening tests, the objective quality measures (Table 1) and spectrograms (Fig. 3) we can conclude that the multi-band method yields good speech quality with minimal musical noise then the basic PSS.

The top spectrogram is the noisy speech, the middle is the enhanced speech by power spectral subtraction and the bottom spectrogram is the enhanced speech with multi-band spectral subtraction.

CONCLUSION

The multi-band approach, a method which essentially takes into account the non-uniform effect of noise on the spectrum of the noisy speech, gives the best overall quality improvement over the basic power spectral subtraction. This demonstrates the potential of frequency dependent methods in reducing the remnant noise-speech distortion trade-off of linear spectral subtraction methods.

REFERENCES

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoustic, Speech and Signal Processing, ASSP., 27: 113-120.

Berouti, M., R. Schwartz and J. Makhoul, 1979. Enhancement of speech corrupted by acoustic noise, Proc. ICASSP., pp: 208-211.

Crozier, R.M., B.M.G. Cheetham, C. Holt and E. Munday, 1980. Speech enhancement employing spectral subtraction and linear predictive analysis, IEEE Trans. On Acoustic, Speech and Signal Processing, ASSP., 28: 137-145.

http://www.utdallas.edu/~loizou/speech/noizeus/.

Nathalie Virag, 1999. Single channel speech enhancement based on masking properties of the human auditory system, IEEE. Trans. Speech and Audio Processing, 7: 126-137.

Quackenbush, S.R., T.P. Barnwell and M.A. Clements, 1998. Objective Measures of Speech Quality, Prentice-Hall, NJ.

Sim, B.L., Y.C. Tong, J.S. Chang and C.T. Tan, 1998. A Parametric formulation of the generalized spectral subtraction method, IEEE Trans. Speech and Audio Processing, 6: 328-337.

Sunil D. Kamath and P.C. Loizou, 2002. A multi-band spectral subtraction method for speech enhancement, Proc. ICASSP., pp: 4-4164.

Signal Processing Information Base, 2003. Noise data http://spib.rice.edu/spib/select_noise.html.