Wideband Speech Enhancement Using a Robust Noise Estimation

Deepa Janardhanan, Ulrich Heute

Institute for Circuit and System Theory, University of Kiel, Kaiserstr. 2, D-24143 Kiel, Germany, Email: {dj, uh} @tf.uni-kiel.de

Introduction

Noise is an unavoidable obstacle in communication systems, especially in hands-free communication where the background noise level is high due to the distance of the speaker from the microphone. In such systems the inclusion of speech enhancement becomes mandatory, provided that it does not introduce any additional distortions to the speech and the background noise. Especially now, different issues in wideband-speech enhancement are addressed. A good noise estimation combined with a suitable spectral subtraction rule plays a key role in obtaining a good subjective quality and intelligibility of the enhanced speech. In this paper the noise estimation method [1, 2] is considered, which already for telephone-band noisy speech shows degradation in performance when the background noise spectrum has a narrowband character (for, e.g., in cars where the noise has low-pass behavior). This effect becomes stronger for wideband noisy speech and hence leads to stronger degradation of the enhanced speech. With a modified noise estimation method an improvement in performance is observed for both narrowband and wideband background noise.

Overview of Noise Estimation

The speech signal disturbed with additive noise is mathematically formulated as,

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

where $x(i,k)$ denotes the noisy speech signal, $s(i,k)$ the clean speech signal, and $n(i,k)$ the additive background noise. The noisy speech signal is processed framewise, i represents the time index and k the frame index. By transforming to the frequency domain using the discrete Fourier transform (DFT) we get the short-time spectrum as

$$X(\mu,k) = S(\mu,k) + N(\mu,k), \tag{2}$$

where μ denotes the frequency index. An estimate of the short-time speech spectrum can be obtained by means of a simple weighting rule called as the Wiener rule,

$$\hat{S}(\mu,k) = H_w(\mu,k) \cdot X(\mu,k) \tag{3}$$

where $H_w(\mu,k)$ is the Wiener weighting function which is defined as

$$H_w(\mu,k) = 1 - \frac{\hat{\phi}_{nn}(\mu,k)}{\phi_{xx}(\mu,k)} \tag{4}$$

where, $\phi_{xx}(\mu,k)$ and $\hat{\phi}_{nn}(\mu,k)$ denote the short-time power spectral density (PSD) of noisy speech and noise estimate respectively. A more complex formulation of the weighting rule was proposed in [3]. Both the rules require the knowledge of the background noise estimate $\hat{\phi}_{nn}(\mu,k)$. In this section we give a short overview of the estimation of background noise as in [1, 2]. The noise estimate is obtained by a separate estimation of the spectral shape and the gain factor. In the estimation of the noise spectral shape an initial estimate $\hat{W}_{nn}(\mu,k)$ of the noise PSD is obtained using [4] or [5]. The gain factor $\hat{G}(k)$ adjusts the spectral shape to the correct noise power, hence the real variations in the noise power spectrum are updated without any delay. The initial noise estimate $\hat{W}_{nn}(\mu,k)$ and the noisy speech PSD, $\phi_{xx}(\mu,k)$, are smoothed along the frequency direction, in order to reduce their variances. The smoothing is done by means of a median filter (or by convolving with a spectral lowpass filter):

$$\tilde{\phi}_{xx}(\mu,k) = med_D\{|X(\mu,k)|^2\} \tag{5}$$

$$\tilde{W}_{nn}(\mu,k) = med_D\{\hat{W}_{nn}(\mu,k)|\}, \tag{6}$$

where $D(=9)$ is the length of the median filter. The values of $\tilde{\phi}_{xx}(\mu,k)$ are sorted in ascending order. The gain factor $\hat{G}(k)$ is computed as the ratio of the mean power of the L smallest values in $\tilde{\phi}_{xx}(\mu,k)$ to the mean power of the corresponding L values in $\tilde{W}_{nn}(\mu,k)$:

$$\hat{G}(k) = \frac{\sum_{l=1}^{L} \tilde{\phi}_{xx}(\mu_l,k)}{\sum_{l=1}^{L} \tilde{W}_{nn}(\mu_l,k)}. \tag{7}$$

The noise estimate is now given by,

$$\hat{\phi}_{nn}(\mu,k) = \hat{G}(k) \cdot \hat{W}_{nn}(\mu,k). \tag{8}$$

Drawbacks

There are two drawbacks to the noise estimation technique [1] that can be clearly seen in the computation of the gain factor in equation (7). Firstly, the L smallest values in $\tilde{W}_{nn}(\mu,k)$ would be clustered in specific low-energy bands of the spectrum in case of narrowband background noise. Secondly, for the case when the background noise has a low-pass (LP) character (for, e.g., noise from a car), the L smallest values would be clustered in the low energy high frequency region. The high frequency components are highly random in nature, and since the gain factor depends on the ratio of the mean of the L smallest values, $\hat{G}(k)$ fluctuates from frame to frame. This results in a fluctuating residual background noise which is unpleasant and disturbing to the listener. The gain factor was slightly modified in [2] to take into account colored noise but this modification was not sufficient to minimize the unwanted clustering of the L smallest spectral values especially in case of LP charactered background noise.

Modified Noise Estimation

In order to reduce the variance of the gain factor from frame to frame a first-order recursive smoothing of the noisy speech power spectrum $\phi_{xx}(\mu,k)$ is performed along the time direction

$$\bar{\phi}_{xx}(\mu,k) = \alpha \cdot \bar{\phi}_{xx}(\mu,k-1) + (1-\alpha) \cdot \phi_{xx}(\mu,k), \tag{9}$$

where α(=0.91) may be called the 'forgetting factor'. $\bar{\phi}_{xx}(\mu,k)$ is now smoothed along the frequency direction as in equation (5).

$$\bar{\tilde{\phi}}_{xx}(\mu,k) = med_D\{\bar{\phi}_{xx}(\mu,k)\}. \tag{10}$$

In order to minimize the clustering of the L smallest values in $\tilde{\bar{\phi}}_{xx}(\mu,k)$ and $\tilde{W}_{nn}(\mu,k)$, the search for the smallest values is performed independently in narrower bands rather than over the entire short-time spectrum. This is done by dividing the short-time PSD's $\tilde{\bar{\phi}}_{xx}(\mu,k)$ and $\tilde{W}_{nn}(\mu,k)$ of DFT length $N_{DFT}(=512)$ into smaller bands of equal length $K=\frac{N_{DFT}}{M}$. $L'(=3)$ smallest values are chosen from each of the $M(=4)$ bands in $\tilde{\bar{\phi}}_{xx}(\mu,k)$ and the corresponding spectral components in $\tilde{W}_{nn}(\mu,k)$. The values of $\tilde{\bar{\phi}}_{xx}(\mu,k)$ in each subband are sorted in ascending order. The modified gain function is given by

$$\hat{G}_{mod}(k)=\frac{\sum_{l=1}^{L'\cdot M}\tilde{\bar{\phi}}_{xx}(\mu_l,k)}{\sum_{l=1}^{L'\cdot M}\tilde{W}_{nn}(\mu_l,k)}. \qquad (11)$$

The noise estimate using the modified gain function is calculated by

$$\hat{\phi}_{nn}(\mu,k)=\hat{G}_{mod}(k)\cdot\hat{W}_{nn}(\mu,k). \qquad (12)$$

Fig. 1 shows the enhanced speech obtained by applying the noise estimation [2] (upper figure) and the modified method (lower figure), respectively, with the spectral subtraction rule [3] to wideband speech (sampling frequency of 16kHz) distorted with noise from a car (which has a low-pass spectral character). The residual background noise present in the enhanced speech has less variations and is more attenuated for the modified method than that using [2].

Objective Tests

The absolute error of the noise estimate is computed as in [1, 2], where two different measures are used for computing the error, one for the case when speech is absent and the other one when speech is present. When speech is present psychoacoustics (masking) is taken into consideration. The errors in the noise estimates were compared for different noise estimation techniques, [2, 4, 5, 6] and the proposed modified noise estimation technique. The upper plot in fig. 2 shows the error curves for speech distorted with white artificial background noise and the lower plot for the case when speech is distorted with real background car noise (LP character). The proposed method performs better than [2] for narrowband noise, and it is superior even for wideband disturbances.

Conclusion

This paper describes the enhancement of wideband speech using a noise estimation as in [2], which shows degradation in performance when the background noise has narrowband characteristics. A modification of this noise estimation was proposed which when combined with a spectral subtraction method [3] was shown to perform better than the method in [2], for wideband speech (16kHz sampling frequency) distorted with narrowband or wideband background noise characteristics.

References

[1] T. G"ulzow, *Spectral-Subtraction-Based Speech Enhancement Using a New Estimation Technique for Non-Stationary Noise*, in Proc. of the IWAENC, pp. 76-79, 1999.

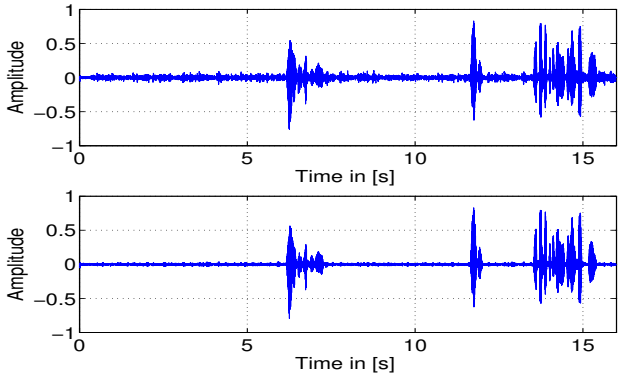

Figure 1: *Enhanced speech using spectral subtraction rule [3] combined with [2] (upper figure) or modified noise estimation method (lower figure)*

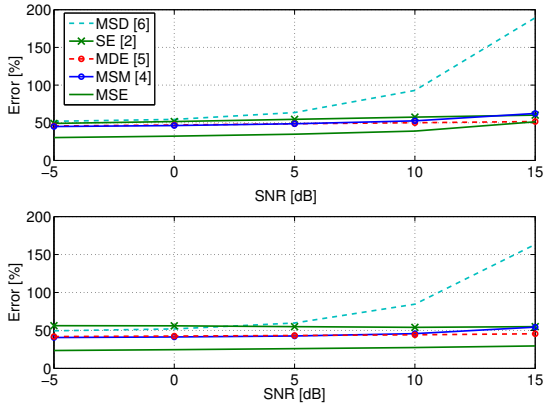

Figure 2: *Normalized estimation-error in [%] for wideband speech distorted with white artificial background noise (upper figure) and low-pass real background car noise (lower figure); MSM: Minimum Statistics Martin, SE: Separate Estimation, MDE: Modified Direct Estimation, MSD: Minimum Statistics Doblinger, MSE: Modified Separate Estimation (proposed method)*

[2] T. G"ulzow, *Verbesserung der Qualit"at stark gest"orter Sprachsignale-Detektion eines Tr"agerversatzes und Unterdr"uckung additiver St"orungen*, PhD thesis, Christian-Albrechts-Universit"at zu Kiel, Nr. 20 in Arbeiten "uber Digital Signalverarbeitung, Hrsg. U. Heute, Shaker Verlag, 2001.

[3] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. on ASSP,Vol. 32, pp. 1109-1121, 1984.

[4] R. Martin, *Spectral Subtraction Based on Minimum Statistics*, in Proc. of EUSIPCO, pp. 1182-1185, 1994.

[5] L. Arslan, A. McCree and V. Viswanathan, *New Methods for Adaptive Noise Suppression*, in Proc. of the ICASSP, vol. 1, pp. 812-815, 1995.

[6] G. Doblinger, *Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands*, in Proc. of the EUROSPEECH, pp. 1513-1516, 1995.