Speech Enhancement Using a Modified Speech-Presence-Uncertainty Criterion

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

Introduction

The presence of environmental noise in speech communication is an unavoidable obstacle. Especially, in a car where the distance of the speaker from the microphone is large, high levels of background noise occur. Furthermore, as noise degrades the quality and intelligibility of the speech signal, it results in listener fatigue. Speech-enhancement aims at improving the perceptual quality and/or the intelligibility of such a degraded speech signal by carefully reducing the background noise. There are several approaches to speech-enhancement, one of them being spectral subtraction. A single-channel spectral-subtraction method has generally two main tasks: firstly to estimate the short-time background noise power-spectrum and secondly to estimate the short-time clean speech power-spectrum. In such a system only the noisy speech signal is available for further processing, thereby making the speech-enhancement task more challenging. It is desirable to have an approach which can carefully retain the speech signal and suppress the noise. For this purpose algorithms which incorporate the concept of speech-presence-uncertainty have shown to be successful candidates for speech enhancement.

Speech enhancement

A simplified single-channel speech-communication scenario has as input clean speech $s(k)$, and additive background noise $n(k)$, and as output noisy speech $y(k) = s(k) + n(k)$. Due to the short-time stationarity of the speech signal, the noisy speech signal is segmented into short-time frames by means of a suitable window function (e.g., Hann window) with half-overlapping. Each speech time-frame is then transformed to the frequency-domain by means of the discrete-Fourier transform (DFT). So, now the short-time DFT can be written as $Y(\mu,i) = S(\mu,i) + N(\mu,i)$, where μ denotes the frequency index and i the frame index. Generally a gain factor, $H(\mu,i)$, is derived as a function of $|Y(\mu,i)|^2$ and the estimate of the short-time noise power-spectrum [1]. An estimate of the short-time clean speech spectrum is given by $\hat{S}(\mu,i) = Y(\mu,i) \cdot H(\mu,i)$.

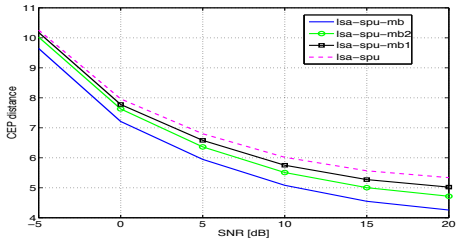
Speech-presence-uncertainty criterion

A spectral weighting function known as the log-spectral amplitude estimator (LSA) in [2] showed good performance, and was able to remove the common artifact of a conventional spectral-subtraction algorithm, know as "musical-noise". This gain factor, $H_{LSA}(\mu,i)$, based on LSA was modified in [3] by incorporating the probability of speech presence and absence, so that background noise is sufficiently attenuated and weak speech components are preserved. This optimally modified LSA (OM-LSA) estimator was obtained under the assumption that the short-time Fourier transform (STFT) coefficients of speech and noise have a Gaussian distribution. With this consideration a conditional speech presence probability was derived as

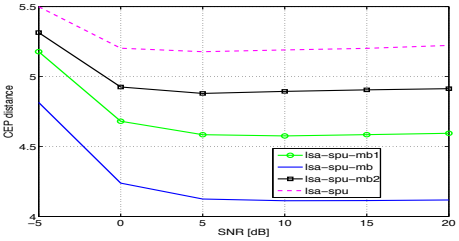
$$p(\mu,i) = \left\{1 + \frac{q(\mu,i)}{1-q(\mu,i)}(1+\xi(\mu,i)) \cdot exp(-\upsilon(\mu,i))\right\}^{-1}, \quad (1)$$

where $q(\mu,l)$ represents the *a-priori* speech-absence probability (SAP), $\xi(\mu,i)$ denotes the *a-priori* signal-to-noise ratio (SNR) which is given by $\xi(\mu,i) = |S(\mu,i)|^2/|N(\mu,i)|^2$, and $\upsilon(\mu,i) = \gamma(\mu,i) \cdot \xi(\mu,i)/(1+\xi(\mu,i))$ where $\gamma(\mu,i)$ is the a-posteriori SNR given by $\gamma(\mu,i) = |Y(\mu,i)|^2/|N(\mu,i)|^2$. The OM-LSA is written as $H_{OM-LSA}(\mu,i) = \{H_{LSA}(\mu,i)\}^{p(\mu,i)} \cdot \{H_{min},\}^{1-p(\mu,i)}$, where H_{min} is a constant spectral floor which avoids the complete suppression of spectral components in case $p(\mu,i) \approx 0$; this results in a natural sounding background noise. An estimate of the *a-priori* SNR, $\hat{\xi}(\mu,i)$, was obtained by the decision-directed approach as proposed in [4] The *a-priori* SAP was estimated by considering three parameters to make a final soft-decision on the probability of speech absence. To start with, a second *a priori* SNR was computed by performing the following averaging: $\zeta(\mu,i) = \beta\zeta(\mu,i-1) + (1-\beta)\hat{\xi}(\mu,i-1)$. The three parameters were based on: (a) mean $\zeta(\mu,i)$ of a frame (b) local average of $\zeta(\mu,i)$ and (c) global average of $\zeta(\mu,i)$. The parameters based on (b) and (c) were meant to exploit the correlation of speech presence in the neighboring frequency bins. The local and global averaging for (b) and (c), respectively, were performed along frequency bins in each frame by using windows (e.g., Hann window) of suitable length. Similarly, based on local and global averages, a soft-decision was also assigned to the two parameters (b) and (c) which were represented by $P_{local}(\mu,i)$ and $P_{global}(\mu,i)$, respectively. Finally, the *a-priori* SAP estimate was obtained as $\hat{q}(\mu,i) = 1 - P_{local}(\mu,i) \cdot P_{global}(\mu,i) \cdot P_{frame}(i)$.

In [5] it was shown that the SNR is not the same for all frequency bins; this behavior of the SNR is due to the fact that the background noise has a non-uniform effect on the speech spectrum. This consideration calls for modifying the parameter $P_{frame}(i)$ so that it is a function of the average *a-priori* SNR within frequency bands instead of the entire speech spectrum. The next section describes the proposed multi-band spectral subtraction method.



(a)



(b)

Figure 1: CEP distances for wideband speech distorted with (a) factory noise and (b) car noise.

Proposed multi-band spectral subtraction

For the implementation of the spectral subtraction method discussed in the previous section independently in linearly-spaced speech bands , the *a-priori* SAP estimate is re-formulated as

$$\hat{q}^b(\mu_b, i) = 1 - P^b_{local}(\mu_b, i) \cdot P^b_{global}(\mu_b, i) \cdot P^b_{frame}(i) \quad (2)$$

where b denotes the band number, μ_b $(s_b \leq \mu_b \leq e_b)$ is the frequency index in band b with s_b and e_b marking the begin and end of band b. It was observed in the multi-band case that by neglecting the redundant parameters $P^b_{global}(\mu_b, i)$ and $P^b_{local}(\mu_b, i)$ for the estimation of the *a-priori* SAP resulted in a better enhancement of speech. Hence, in a multi-band spectral subtraction approach the *a-priori* SAP estimate is reduced to $\hat{q}^b(i) = 1 - P^b_{frame}(i)$.

In order to compensate for the absence of $P^b_{global}(\mu_b, i)$ and $P^b_{local}(\mu_b, i)$ which had the function of exploiting the local and global correlation of the frequency bins, $P^b_{frame}(i)$ is now computed as follows: The recursive average of the band-wise *a-priori* SNR $\zeta^b(\mu_b, l)$ is subject to a second averaging by applying an averaging window (e.g. Hann window). This was done to exploit the correlation in the adjacent speech bands.

The gain factor is now defined in the b^{th} band as

$$H^b_{OM-LSA}(\mu, i) = \{H^b_{LSA}(\mu, i)\}^{p^b(\mu,i)} \cdot \{H^b_{min}\cdot\}^{1-p^b(\mu,i)} \quad (3)$$

Experimental results and discussion

In order to assess the objective quality of the enhanced speech, cepstral (CEP) distance measure was employed [6]. In figure 1 the CEP distances for wideband speech degraded by different additive background noises (factory and car noise) are shown at varying noise levels. For each type of background noise 4 variations of a speech-presence-uncertainty criterion were tested. Each type is described as follows:

- lsa-spu: (log-spectral amplitude - speech-presence-uncertainty) represents the method in [3] with the approach in [1] for the noise estimate.
- lsa-spu-mb1: (lsa-spu-multiband1) is similar to lsa-spu but the *a-priori* SAP estimate was computed within 5 linearly spaced bands of the speech spectrum.
- lsa-spu-mb2: (lsa-spu-multiband2) in this case the *a-priori* SAP estimate is computed as $\hat{q}^b(\mu_b, l) = 1 - P^b_{frame}(l)$, where $P^b_{frame}(l)$ is determined in a manner similar to that in lsa-spu-mb1.
- lsa-spu-mb: (lsa-spu-multiband) is the newly proposed method.

It is observed from the experimental results in figure 1 that the proposed method shows lower distance values when compared to the other variations of the *a-priori* SAP estimate. For car noise the distance is constant after 5dB SNR (maximum recovery capacity reached), but still the recovery of the spectral components is better with the proposed method. The perceptual quality of the enhanced wideband speech (by informal subjective listening) was also found to be better for the proposed approach.

Conclusion

In this paper, a new multi-band spectral-subtraction method incorporating a modified speech-presence-uncertainty criterion is proposed. The improved performance of this method was shown by means of experimental results. The presented approach has the ability to preserve speech spectral components due to the adoption of a multi-band *a-priori* SAP estimate.

References

[1] D. Janardhanan and U. Heute. Wideband Speech Enhancement Using a Robust Noise Estimation. In *Proceedings of DAGA '05*, pages 159–160, 2005.

[2] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. *IEEE Transactions on ASSP*, ASSP-33:443–445, 1985.

[3] I. Cohen. Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator. *IEEE Signal Processing Letters*, 9:113–116, 2002.

[4] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Transactions on ASSP*, ASSP-32:1109–1121, 1984.

[5] Sunil D. Kamath and Philipos C. Loizou. A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise. In *Proceedings of ICASSP*, volume 4, pages IV–4164, 2002.

[6] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Taylor & Francis Group, 2007.