DUAL CHANNEL REDUCTION OF RAPIDLY VARYING HARMONIC AND RANDOM NOISE USING A SPOT MICROPHONE

Florian Heese, Thomas Esch and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{heese,esch,vary}@ind.rwth-aachen.de

Abstract: This contribution presents a dual channel speech enhancement system that is operating in the frequency domain. State-of-the-art estimators usually rely on the assumption that the noise signal is stationary or only slightly time-varying. In this contribution we are explicitly considering rapidly time-varying harmonic noise (e.g., engine noise). Therefore we take a spot microphone into account enabling the exploitation of correlation between the main and the spot microphone signals. Our recently proposed noise power spectral density (PSD) estimator for rapidly varing harmonic and random noise [3] is modified and combined with a coherence based noise PSD estimator. The performance of the proposed system provides consistently improved performance.

1 Introduction

Speech quality and intelligibility may significantly be degraded under the presence of background noise, e.g., street noise or engine noise. One of the popular methods for enhancing degraded speech represents the noisy speech in the short-time Fourier domain and applies individual adaptive gains to each frequency bin based on a noise power spectral density (PSD) estimation, e.g., [6] [2] [7]. Speech enhancement has many applications in voice communications, speech recognition and hearing aids.

Estimation of the noise PSD remains a crucial and challenging task in every noise reduction system. State-of-the-art estimators usually rely on the assumption that the noise signal is stationary or only slightly time-varying. In this contribution, however, we are explicitly considering noise environments characterized by mainly rapidly time-varying harmonic noise as e.g., engine noise.

For this purpose we extend our recently proposed noise PSD estimator [3]. Taking a spot microphone offers the opportunity to use a coherence based approach (see Fig. 1). The second microphone is placed at a distance of about 15 cm from the main microphone. Measurements on a mock up phone placed in a car have shown that strong noise coherence between the two microphones exists for $f < f_c \approx 1.3\,\text{kHz}$. By using a noise canceller the correlation between the two microphones can be exploited to improve the initial noise PSD estimation at low frequencies ($f < f_c$) significantly, especially at low input SNR: -10dB, ..., 5dB where also wind and tire noise components are present.

The remainder of this paper is organized as follows: In Sec. 2 a brief overview of the proposed system is given. Section 3 comprises the proposed noise estimation technique. Experimental results are shown in Sec. 4 and conclusions are drawn in Sec. 5.

2 System Overview

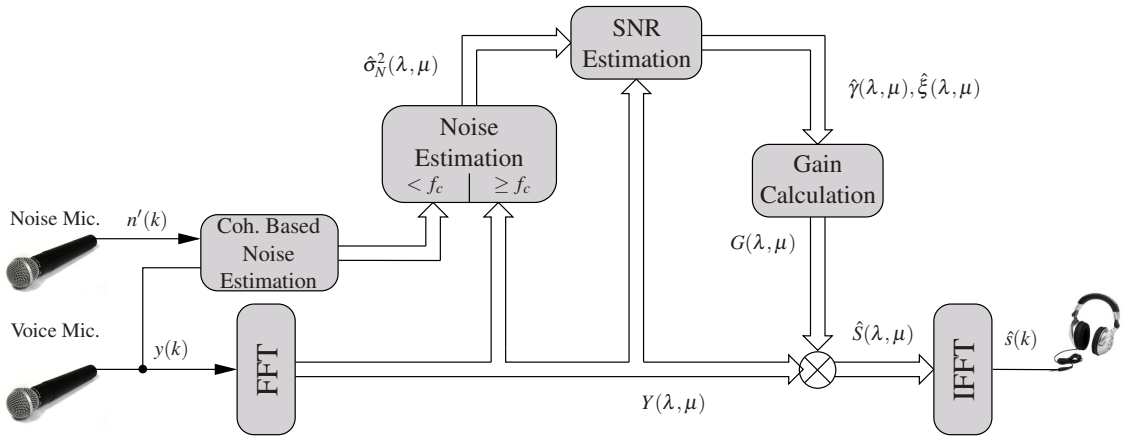

Figure 1 - Proposed noise reduction system.

A simplified block diagram of the proposed speech enhancement system is depicted in Fig 1. The samples $y(k)$ and $n'(k)$ are obtained by analog-digital conversion with sampling frequency of $f_s = 8\,\text{kHz}$, where k is the discrete time index. It is assumed that the noisy input signal $y(k)$ of the voice microphone consists of the clean speech signal $s(k)$ which is degraded by an additive noise component $n(k)$ according to:

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

The noise microphone which records $n'(k)$ should be placed with the aim to record mainly noise components while preserving a good coherence with the noise $n(k)$ in the voice microphone. The noise is primarily dominated by rapidly time-varying harmonic noise (e.g., engine noise), but it also consists of wind and tire noise. The distance between the two microphones is approximately 15 cm. Given a strong coherence between $n(k)$ and $n'(k)$ it is possible to estimate the noise component $n(k)$ of the voice microphone.

Different noise estimation schemes are applied dependent on the coherence between the noise component in the voice microphone signal $y(k)$ and the noise microphone signal $n'(k)$. For frequencies with a strong coherence a noise canceller is utilized to estimate the PSD $\hat{\sigma}_N^2$ of $n(k)$ and a statistical based PSD estimator [3] is used for the remaining frequencies.

The noise suppression is performed in the frequency domain. Therefore $y(k)$ is segmented into overlapping frames of length L_F. After windowing and zero-padding, the fast Fourier transform (FFT) of length M_F is applied to these frames. The spectral coefficients of the noisy input signal $y(k)$ at frequency bin μ and frame λ are given by:

$$Y(\lambda,\mu) = S(\lambda,\mu) + N(\lambda,\mu), \tag{2}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ represent the spectral coefficients of the speech and the noise signal, respectively.

Based on the estimate $\hat{\sigma}_N^2(\lambda,\mu)$ of the noise PSD as described above, two SNR parameters are estimated, namely the *a posteriori SNR* $\gamma(\lambda,\mu)$ and the *a priori SNR* $\xi(\lambda,\mu)$ defined as:

$$\gamma(\lambda,\mu) = \frac{|Y(\lambda,\mu)|^2}{\sigma_N^2(\lambda,\mu)} \quad \text{and} \quad \xi(\lambda,\mu) = \frac{\mathscr{E}\{|S(\lambda,\mu)|^2\}}{\sigma_N^2(\lambda,\mu)}. \tag{3}$$

The a priori SNR is estimated using the decision-directed approach [2]. The noise suppression is achieved by spectral weighting and is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ by the weighting gains $G(\lambda,\mu)$:

$$\hat{S}(\lambda,\mu) = G(\lambda,\mu)\cdot Y(\lambda,\mu). \tag{4}$$

In order to determine the weighting gains, the well-known Wiener filter is used which is dependent on the SNR estimates. The enhanced signal $\hat{s}(k)$ in the time domain is obtained by applying an Inverse Fast Fourier Transform (IFFT) and overlap-add.

3 Noise Estimation

Different noise PSD estimation techniques are presented in this section. They are divided into statistical based noise estimation at higher frequencies $f \geq f_c$ and coherence based noise estimation at lower frequencies $f < f_c$.

3.1 Statistical Based Estimation of Time Varying Harmonic and Stationary Noise

We consider speech signals disturbed by stationary *and* harmonic noise which are characterized by (strong) spectral components at multiples of the (time varying) fundamental frequency f_0. As the fundamental frequency might change over time very fast conventional noise estimation techniques, e.g., Minimum Statistics, usually fail in tracking the spectral harmonics.

The original Minimum Statistics approach [8] is based on two assumptions: speech and noise are statistically independent and the power of the noisy signal $y(k)$ often decays to the power level of the noise signal $n(k)$. Using a smoothed PSD of the noisy signal $y(k)$ it is possible to track the minimum separately for each frequency bin within a time window of length D. The duration of the time window for the minimum search states a trade-off between fast noise tracking and the speech distortions after spectral weighting. As the minimum is always smaller or equal to the mean noise power a bias correction according to:

$$\hat{\sigma}_N^2(\lambda,\mu) = B(\lambda,\mu)\hat{\sigma}_{Y,\min}^2(\lambda,\mu), \tag{5}$$

is necessary. The correction factor $B(\lambda,\mu)$ [8] is mainly dependent on the variance of the noisy signal.

Minimum Statistics performs well in stationary and slowly changing noise conditions as the minimum at each frequency bin within a search time window provides a good estimate of the actual noise power. However, when it comes to a sudden rise in the noise power in one specific frequency bin, Minimum Statistics is not able to track this rise due to the large window length D which typically corresponds to a duration of approximately 1.5 seconds [8].

In our system, we use a modified Minimum Statistics procedure [4] to estimate the harmonic noise power $\sigma_{N,\mathrm{h}}^2(\lambda,\mu)$. Instead of tracking the spectral minimum over time at *one* specific frequency bin (see Fig. 2, method a), we adaptively 'look back' inclined according to the evolution of the harmonics in the time-frequency plain (see Fig. 2, method b). Following one specific harmonic oscillation over time, the harmonic components are no longer fluctuating that much but relatively stationary and we can apply the original Minimum Statistics concept.

Facing the problem of stationary and harmonic noise we modify our previous noise reduction system [3] which consists of two stages. In the first stage, the harmonic noise power $\hat{\sigma}_{N,\mathrm{h}}^2(\lambda,\mu)$ is estimated and attenuated using a modified Minimum Statistics approach. A conventional noise reduction is applied in a second stage in order to reduce the random components $\hat{\sigma}_{N,\mathrm{r}}^2(\lambda,\mu)$ of the noise spectrum.

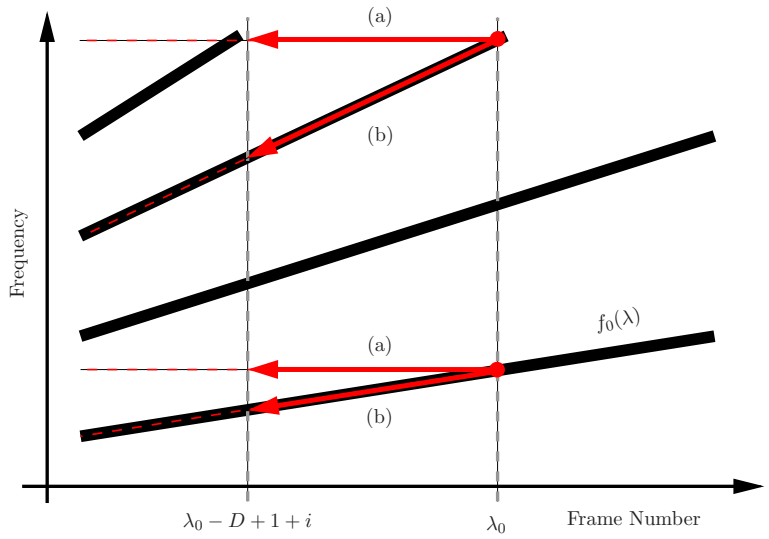

Figure 2 - 'Direction of view' of (a) original Minimum Statistics and (b) modified Minimum Statistics.

In contrast to [3] we perform the estimation of stationary *and* harmonic noise separately (not sequentially) for each noise type resulting in $\hat{\sigma}^2_{N,\mathrm{h}}(\lambda,\mu)$ and $\hat{\sigma}^2_{N,\mathrm{r}}(\lambda,\mu)$ respectively. By combining the noise PSDs according to:

$$\hat{\sigma}^2_{N,\mathrm{stat}}(\lambda,\mu) = \max\left\{\hat{\sigma}^2_{N,\mathrm{h}}(\lambda,\mu),\ \hat{\sigma}^2_{N,\mathrm{r}}(\lambda,\mu)\right\}, \tag{6}$$

it is possible to track stationary and harmonic noise jointly. For stationary noise estimation any conventional estimation technique can be applied. In this paper the original Minimum Statistics and the MMSE based noise tracking algorithm [5] are investigated in the following evaluation.

3.2 Coherence Based Noise Estimation

At low frequencies, the correlation between the signals $n(k)$ and $n'(k)$ taken by the voice and noise microphone is exploited to estimate the noise power spectrum. A system overview is depicted in Fig. 3. It is assumed that the noise components of the voice and noise microphone have strong coherence for $f < f_c$.

The Adaptive Noise Canceller system (see [10]) is able to suppress signal parts with strong coherence between the noise microphone signal $n'(k)$ and the voice microphone signal $y(k) =$

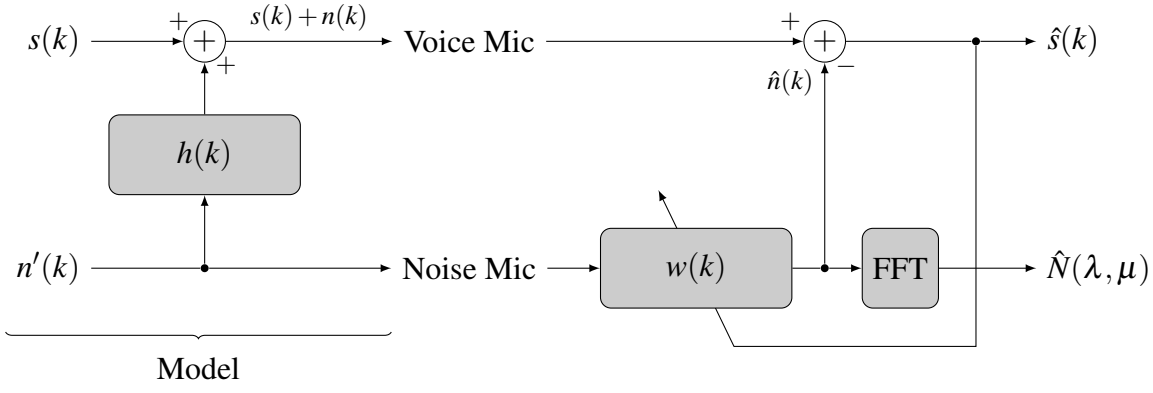

Figure 3 - Noise Canceller

$s(k)+n(k)$. Therefore, the microphone signals are at first lowpass filtered with a cut off frequency f_c resulting in the correlated frequency parts. The transfer function of $h(k)$ is unknown and approximated by $w(k)$. The filter coefficients $w(k)$ are determined using the normalized least mean square (NLMS) algorithm:

$$w(k+1) = w(k) + \frac{\alpha \hat{s}(k)\mathbf{n}'(\mathbf{k})}{||\mathbf{n}'(k)||^2}, \ \alpha \text{ stepsize factor}, \tag{7}$$

where $||\cdot||^2$ denotes the L2-norm. $\hat{n}(k)$ is an estimate of the noise component of the voice microphone. It can be used to calculate an estimation for the noise PSD $\hat{\sigma}^2_{N,\text{coh}}(\lambda,\mu)$ in the proposed noise reduction system (see Fig. 1) by taking magnitude square of FFT of $\hat{n}(k)$.

3.3 Combination of Coherence and Statistical Based Noise PSD Estimation

As mentioned before different noise PSD estimators are applied according to the noise coherence between the voice microphone and noise microphone signal. As depicted in Fig. 4, however, a strong coherence only exists for low frequencies $f < f_c$. Therefore, the combination of the two PSD estimates is expressed as:

$$\hat{\sigma}^2_N(\lambda,\mu) = \begin{cases} \hat{\sigma}^2_{N,\text{coh}}(\lambda,\mu) & \mu < \mu_c \\ \hat{\sigma}^2_{N,\text{stat}}(\lambda,\mu) & \mu \geq \mu_c, \end{cases} \tag{8}$$

where μ_c is the corresponding frequency bin according to f_c.

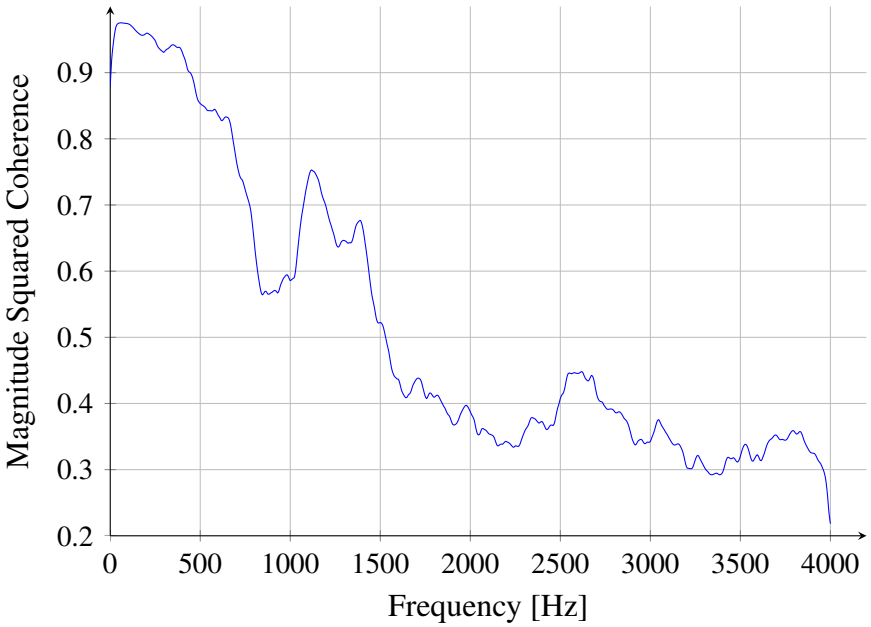

Figure 4 - Measured coherence between voice and noise microphone

4 Results

The proposed noise estimation technique is compared with the results of state-of-the art single-channel noise estimator techniques. Among them, a noise PSD estimator which is developed for reduction of time varying harmonic and stationary noise [3]. Therefore, the speech enhancement system depicted in Fig. 1 was used. Referring to Fig. 1, the following configuration of

noise estimation techniques (A, ..., E) are compared:

	Noise Estimation Technique		
	Statistical Based $\hat{\sigma}^2_{N,\text{stat}}$		**Coherence Based $\hat{\sigma}^2_{N,\text{coh}}$**
	Harmonic Noise Estimation	Stationary Noise Estimation	
A	disabled	Original Minimum Statistics	disabled ($f_c = 0\,\text{Hz}$)
B	Modified Minimum Statistics	Original Minimum Statistics	disabled ($f_c = 0\,\text{Hz}$)
C	Modified Minimum Statistics	MMSE based PSD noise tracking	disabled ($f_c = 0\,\text{Hz}$)
D	Modified Minimum Statistics	Original Minimum Statistics	enabled ($f_c = 1.3\,\text{kHz}$)
E	Modified Minimum Statistics	MMSE based PSD noise tracking	enabled ($f_c = 1.3\,\text{kHz}$).

For the evaluation, four different (real) noise recordings taken from a sports-car were each added to three male and two female speech sequences (each with a length of 8 s taken randomly from the NTT speech database) at an input SNR varying between -10 dB and 15 dB with an increment of 5 dB. The parameters that are used in the simulations are listed in Tab. 1.

In the evaluation, speech and noise can be filtered separately with weighting gains adapted for the noisy signal. Hence, the output signal can additionally be stated as $\hat{s}(k) = \tilde{s}(k) + \tilde{n}(k)$, where $\tilde{s}(k)$ is merely the filtered speech signal and $\tilde{n}(k)$ the residual noise. Based on these quantities, the speech and noise attenuation (SA and NA; e.g., Chap. 4 in [1]) and the Short-time Objective Intelligibility (STOI) [9] were calculated.

The averaged results are depicted in Figs. 5 and 6. Figure 5 shows the difference between noise and speech attenuation. In Fig. 6 the Short-time Objective Intelligibility is plotted over the input SNR. In all figures, higher scores indicate a better performance of the respective approach. As reference for conventional noise reduction a (unmodified original) Minimum Statistics based system (method A) is incorporated in the evaluation process.

The objective measurements show that the proposed system using the statistical and coherence based noise PSD estimation (2ch systems) consistently improves the results of the conventional and recently proposed estimation techniques for harmonic and random noise (1ch systems). All systems with enabled harmonic noise estimation (methods B, C, D and E) are superior compared to the original Minimum Statistics approach.

Comparing 1ch and 2ch systems in Fig. 5, it can be seen that systems using the MMSE based noise PSD tracking for random noise components (methods C, E) perform slightly better than systems using the corresponding modified Minimum Statistics approach (methods B, D) for the entire SNR range. Regarding STOI (Fig. 6), this behavior is valid only for SNR values lower than 10 dB and changes for higher SNR values. However, the tendency that the 2ch systems are better than the corresponding 1ch systems holds in this measurement for all SNR values as well. The objective measurements were confirmed by informal listening tests.

Parameter	**Settings**
Sampling frequency	8 kHz
Frame length L_F	512 ($\hat{=}$ 64 ms)
FFT length M_F	1024 (including zero-padding)
Frame overlap	87.2% (Hann-window)
Input SNR	-10 dB ... 15 dB (step size: 5 dB)
SNR estimation	Decision-directed approach [2]
Partition frequency (f_c)	1300 Hz

Table 1 - System settings.

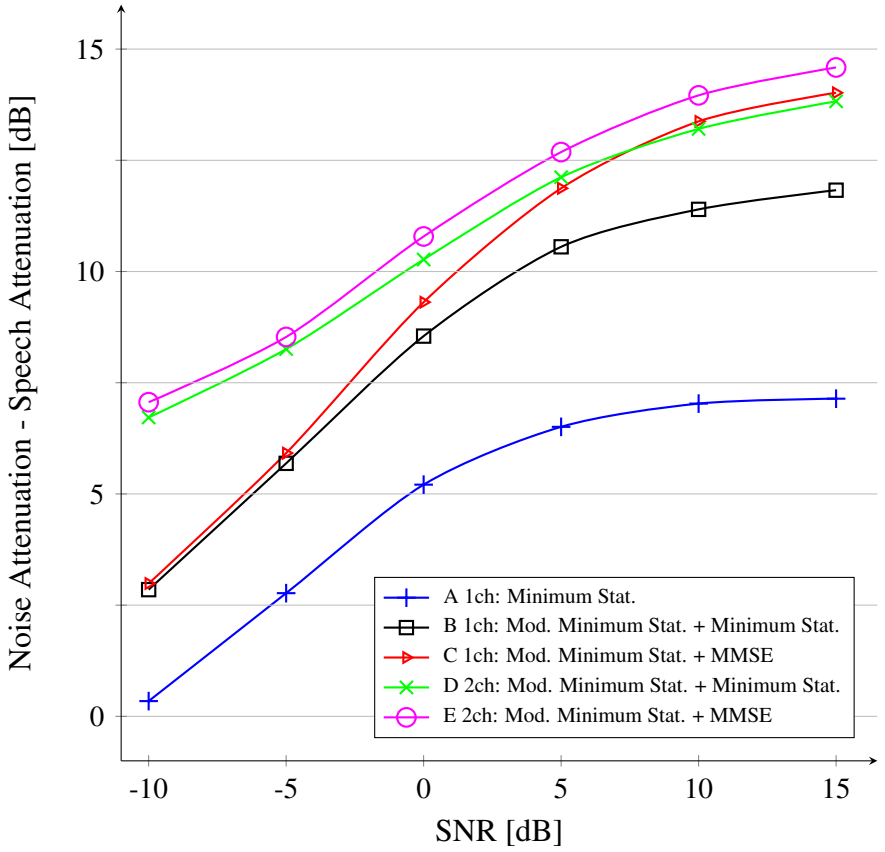

Figure 5 - Difference between noise attenuation and speech attenuation plotted over input SNR

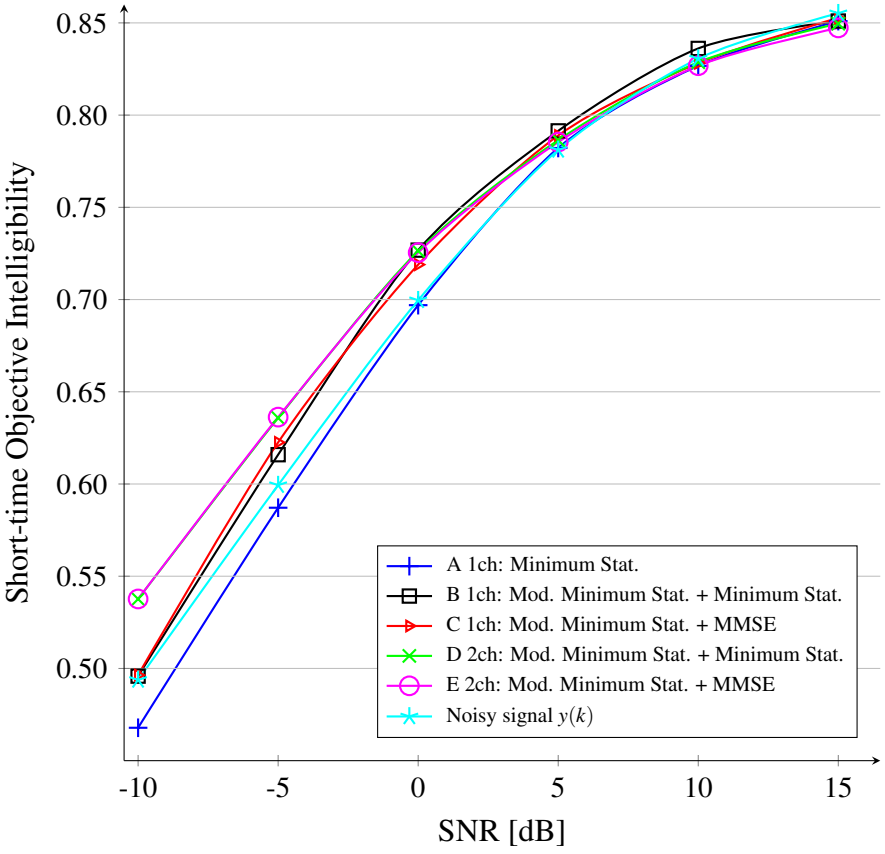

Figure 6 - Short-time Objective Intelligibility (STOI) [9] plotted over input SNR

5 Conclusion

An approach to rapidly time-varying harmonic and random noise estimation has been presented in this paper that exploits correlation between two microphones. While a noise canceller is used to estimate the noise PSD for frequencies with strong coherence a modified statistical based approach to estimate rapidly time-varying harmonic and random noise is used for the remaining frequencies. The obtained coherence and statistical based noise PSDs are combined and used in a conventional noise reduction system that is operating in the frequency domain.

Instrumental measurements show a consistent improvement in terms of noise/speech attenuation and STOI for the proposed system.

References

[1] BENESTY, J., S. MAKINO and J. CHEN (eds.): *Speech Enhancement*. Springer, Berlin, 2005.

[2] EPHRAIM, Y. and D. MALAH: *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*. IEEE Transactions on Acoustic, Speech and Signal Processing, 32(6):1109–1121, 1984.

[3] ESCH, T., M. RÜNGELER, F. HEESE, and P. VARY: *Combined reduction of time varying harmonic and stationary noise using frequency warping*. In *Conference Record of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, Nov. 2010. IEEE.

[4] ESCH, T., M. RÜNGELER, F. HEESE, and P. VARY: *A modified minimum statistics algorithm for reducing time varying harmonic noise*. In *ITG-Fachtagung Sprachkommunikation*. VDE Verlag GmbH, Oct. 2010.

[5] HENDRIKS, R., R. HEUSDENS, and J. JENSEN: *Mmse based noise psd tracking with low complexity*. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4266–4269. IEEE.

[6] LIM, J. S. and A. V. OPPENHEIM: *Enhancement and Bandwidth Compression of Noisy Speech*. Proceedings of the IEEE, 1979.

[7] LOTTER, T. and P. VARY: *Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model*. EURASIP Journal on Applied Signal Proc., 2005.

[8] MARTIN, R.: *Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics*. IEEE Transactions on Speech and Audio Processing, 9(5):501–512, 2001.

[9] TAAL, C., R. HENDRIKS, R. HEUSDENS, and J. JENSEN: *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4214–4217. IEEE, 2010.

[10] WIDROW, B., J. GLOVER JR, J. MCCOOL, J. KAUNITZ, C. WILLIAMS, R. HEARN, J. ZEIDLER, E. DONG JR, and R. GOODLIN: *Adaptive noise cancelling: Principles and applications*. Proceedings of the IEEE, 63(12):1692–1716, 1975.