Blind source extraction based on a direction-dependent a-priori SNR

Lukas Pfeifenberger[1], *Franz Pernkopf*[1]

[1] Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

lukas.pfeifenberger@alumni.tugraz.at, pernkopf@tugraz.at

Abstract

In many hands-free applications, we encounter a speaker located in the near-field embedded in diffuse far-field noise. In this paper, we contribute an algorithm to estimate the speech and noise power spectral density (PSD) based on a *direction-dependent SNR* (DD-SNR). The only prior knowledge needed is a model of the diffuse noise sound field. The enhanced speech signal is obtained by a parametric multi-channel Wiener filter (PMWF), which is constructed without any speech presence or absence probabilities, or smoothing in frequency. We achieve high speech quality and sufficient noise reduction by iteratively improving the speech PSD estimate using the output of the PMWF. The performance of our algorithm is demonstrated by using the PESQ and PEASS measures.

Index Terms: multi-channel speech enhancement, blind source extraction, noise PSD estimation

1. Introduction

Speech intelligibility is a paramount issue in many telecommunication devices. Especially in hands-free applications, background noise is the primary source of speech degradation. Great efforts have been made over the last decades to reduce this ambient noise. While single-channel speech enhancement algorithms require an inherent trade-off between noise reduction and speech quality, multi-channel algorithms also exploit the spatial information of the sound field and thereby achieve better results.

Recent algorithms try to estimate the noise *power spectral density* (PSD) from either the output of a beamformer, or directly from the microphone signals. Noise reduction is then achieved by using a single- or multi-channel Wiener filter. In [1], a *transient beam to reference ratio* (TBRR) is used to derive a speech absence probability, which is used to control a recursive noise floor averaging estimator. In [2], a *multi-channel speech presence probability* (MC-SPP) is derived as a generalization of the classical single-channel *a posteriori* speech presence probability. The MC-SPP is used as a soft-decision rule to estimate the noise PSD of the microphone signals. A similar approach can be found in [3], where a *direct to diffuse ratio* (DDR) is used instead of the MC-SPP.

In this paper we do not use a speech absence or presence probability, but estimate the noise PSD from differential signals which cancel out the speaker. First, we extend the signal-to-reverberant ratio proposed in [4] to the multi-channel case, in order to obtain an a-priori *direction-dependent SNR* (DD-SNR). This SNR is then used to estimate the *acoustic transfer functions* (ATFs) from the speaker to the microphones. The ATFs include reflections like acoustic echoes, and are therefore hard to estimate in general. However, in our case of a near-field speaker the ATFs mainly consist of a constant delay and unity gain. The ATFs are used to align the microphone inputs so that the speech signal is either constructively or destructively added, thereby allowing us to estimate the speech and noise PSDs. Finally, a *parametric multi-channel Wiener filter* (PMWF) is employed to obtain the enhanced speech signal. We show how the PMWF can be used in a second iteration to achieve considerable noise reduction and maintain a high speech quality at the same time. The noise reduction performance and speech quality of our approach is evaluated by using the PESQ score and the *Perceptual Evaluation Methods for Audio Source Separation* (PEASS) Toolkit [5, 6].

This paper is organized as follows: Section 2 introduces the signal model and the involved sound field. Section 3 consideres the estimation of the DD-SNR, and in section 4 the speech and noise PSDs are estimated. In section 5, we formulate the PMWF and show how to use its output in a second iteration. Section 6 summarizes the entire algorithm for clarity. Section 7 evaluates our approach using the PESQ and PEASS scores, and compares it against the TBRR and MC-SPP algorithms. Section 8 concludes the paper.

2. Problem formulation

In our setup, we assume the desired speech source to be in the near-field, and the interfering noise source to be located in the far-field of a linear microphone array of M sensors, with an inter-microphone distance of $d = 5$cm. Diffuse noise and a near-field speaker (i.e. 0.5m speaker distance) are found in many real-world scenarios, i.e. car interiors, subway stations or roadside emergency telephones.

In the frequency domain, we define the signal at the i^{th} microphone as $Z_i(k,l) = A_i(k,l)S(k,l) + N_i(k,l)$, using the wave number $k = \frac{2\pi f}{c}$ and the frame index l, where f and c denote the frequency and the speed of sound, respectively. The unknown speech signal is denoted by $S(k,l)$, and $N_i(k,l)$ expresses the ambient noise signal at the i^{th} channel. $A_i(k,l)$ is the unknown ATF from the speaker to the i^{th} microphone. Covering all M microphones, the signal model can be written in compact vector notation as

$$\boldsymbol{Z}(k,l) = \boldsymbol{A}(k)S(k,l) + \boldsymbol{N}(k,l). \tag{1}$$

The spatial correlation matrix [7] for all microphone signals is defined as expectation of $\boldsymbol{Z}(k,l)\boldsymbol{Z}^{\boldsymbol{H}}(k,l)$, i.e.:

$$\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k) \triangleq E\{\boldsymbol{Z}(k,l)\boldsymbol{Z}^{\boldsymbol{H}}(k,l)\}. \tag{2}$$

Usually, $\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k)$ can be estimated by recursive averaging using $\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l) = \boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l-1)\alpha + (1-\alpha)\boldsymbol{Z}(k,l)\boldsymbol{Z}^{\boldsymbol{H}}(k,l)$. By assuming uncorrelated speech and noise signals, Eqn. (2)

can also be stated as

$$\begin{aligned}\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l) &= \boldsymbol{\Phi}_{\boldsymbol{SS}}(k,l) + \boldsymbol{\Phi}_{\boldsymbol{NN}}(k,l)\\ &= \boldsymbol{A}(k)\boldsymbol{A}^{\boldsymbol{H}}(k)\Phi_S(k,l) + \boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)\Phi_N(k,l),\end{aligned} \tag{3}$$

where $\Phi_S(k,l)$ and $\Phi_N(k,l)$ denote the PSDs of the unknown speech and noise sources, and $\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)$ is the spatial correlation matrix of the diffuse or isotropic noise sound field. It can be thought of as the summation of infinitely many plane waves impinging from all directions at equal strength [8]. Its elements are given as $\Gamma_{N_iN_j}(k) = \frac{\sin(kd_{ij})}{kd_{ij}}$, where d_{ij} is the distance between the i^{th} and the j^{th} microphone. With this setup, our aim is to estimate the speech source $S(k,l)$. However, we do not intend to perform blind dereverberation, it is sufficient to estimate the speech signal at the first (the reference) microphone $A_1(k)S(k,l)$.

3. Direction-dependent a-priori SNR

By simplifying the ATFs to single monochromatic plane waves [8], it becomes possible to detect the presence of the speech signal in the mixed sound field $\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k)$ without prior knowledge [4], i.e.:

$$A_i(k,l) \approx \tilde{A}_i(k,l) = e^{jkd_i \sin\Theta}, \tag{4}$$

where Θ is the impinging angle of the sound wave towards the array, and d_i is the distance between the i^{th} microphone and an arbitrary reference point. This reference point is chosen to be the first microphone [9], so that $\tilde{A}_1(k,l) \triangleq 1$. Using this simplified model, the SNR between the speech and noise PSDs $\xi_\Theta(k,l) = \frac{\Phi_S(k,l)}{\Phi_N(k,l)}$ can be estimated from Eqn. (3). We will use this *direction-dependent a-priori SNR* (DD-SNR) as a sensitive and robust voice activity detector. It can be derived using the spatial coherence matrix for all M microphone signals [7]:

$$\boldsymbol{\Gamma}_{\boldsymbol{ZZ}}(k,l) \triangleq \boldsymbol{E}(k,l)\boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l)\boldsymbol{E}(k,l), \tag{5}$$

with $\boldsymbol{E}(k,l) =$ $\operatorname{diag}\left(\frac{1}{\sqrt{\Phi_{Z_1Z_1}(k,l)}}, \frac{1}{\sqrt{\Phi_{Z_2Z_2}(k,l)}}, \cdots, \frac{1}{\sqrt{\Phi_{Z_MZ_M}(k,l)}}\right)$, where diag($\cdot$) denotes a diagonal matrix. The PSDs $\Phi_{Z_iZ_i}(k,l)$ are the main diagonal elements of the spatial correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{ZZ}}$(k,l). Especially with small microphone array apertures, we can assume equal signal energies among all microphones. Therefore $\boldsymbol{E}(k,l) = \frac{1}{\sqrt{\Phi_S(k,l)+\Phi_N(k,l)}}$ $\boldsymbol{I}_{\boldsymbol{M\times M}}$, and Eqn. (5) becomes

$$\boldsymbol{\Gamma}_{\boldsymbol{ZZ}}(k,l) = \boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l)\frac{1}{\Phi_S(k,l) + \Phi_N(k,l)}. \tag{6}$$

Substituting Eqn. (6) into (3) gives the DD-SNR:

$$\begin{aligned}\xi_\Theta(k,l) = \operatorname{Tr}([&\boldsymbol{\Gamma}_{\boldsymbol{ZZ}}(k,l) - \tilde{\boldsymbol{A}}(k)\tilde{\boldsymbol{A}}^{\boldsymbol{H}}(k)]^{-1}\\ &\cdot [\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k) - \boldsymbol{\Gamma}_{\boldsymbol{ZZ}}(k,l)]),\end{aligned} \tag{7}$$

which is similar to [3]. If the direction of arrival Θ is not known a-priori, it can be globally detected by searching over a small set of possible angles using $\Theta_{OPT} = \arg\max_{\Theta} \frac{1}{K}\sum_{k=0}^{K} \xi_\Theta(k,l)$. In [3], a similar measure to the DD-SNR is used to derive a noise reduction Wiener filter. However, we found $\xi_\Theta(k,l)$ to be too inaccurate especially for low frequencies, because in practice the ATFs won't be pure time delays, and the signal energies at the microphones won't be equal due to gain tolerances. But we can use $\xi_\Theta(k,l)$ to improve the model of the ATFs from simple plane waves to multi-path propagations, i.e. acoustic echos. With a good estimate $\hat{\boldsymbol{A}}(k) \approx \boldsymbol{A}(k)$ we can align the microphone signals to either constructively or destructively add the speech components, which are used to derive the speech PSD $\Phi_S(k,l)$ and the noise PSD $\Phi_N(k,l)$. From the mixture model in Eqn. (1), it can be seen that $A_i(k)$ is generally unobservable, since it is embedded in additive noise. However, by inserting $\xi_\Theta(k,l)$ into Eqn. (3) and using $\hat{A}_1(k,l) \triangleq 1$ for the reference microphone we can construct the following estimator:

$$\begin{aligned}\hat{A}_i(k,l+1) =&\hat{A}_i(k,l)\alpha_1(k,l) + (1-\alpha_1(k,l))\\ &\cdot \left[\frac{1+\xi_\Theta(k,l)}{\xi_\Theta(k,l)}\frac{\Phi_{Z_iZ_1}(k,l)}{\Phi_{Z_1Z_1}(k,l)} - \frac{\Gamma_{N_iN_1}(k)}{\xi_\Theta(k,l)}\right].\end{aligned} \tag{8}$$

To ensure this algorithm only adapts on frequency bins containing speech, we use the DD-SNR as voice activity detector:

$$\alpha_1(k,l) = \begin{cases} \alpha, & \text{if } \xi_\Theta(k,l) > \xi_0 \text{ and } \frac{1}{K}\sum_{k=0}^{K} \xi_\Theta(k,l) > \xi_0\\ 0, & \text{otherwise.}\end{cases} \tag{9}$$

Clearly, estimating the ATFs only works if the filters they represent have a finite impulse response which can be modeled within the duration of a FFT frame l. We chose 32ms as frame length, which is enough to model the acoustic path from the speaker's location to the microphone array within the speaker distance of 0.5m. Reasonable results are obtained by setting the threshold ξ_0 to 0dB.

4. Estimation of the speech and noise PSD

We estimate the speech and noise PSDs by using a summation signal and a differential signal, which are both obtained from the ATF estimate $\hat{\boldsymbol{A}}(k,l)$. The summation signal is given by

$$Y(k,l) = \boldsymbol{F}^{\boldsymbol{H}}(k,l)\boldsymbol{Z}(k,l), \tag{10}$$

which constructively adds the speech components in $\boldsymbol{Z}(k,l)$. This is done via the matrix $\boldsymbol{F}(k,l) = \frac{\hat{\boldsymbol{A}}(k,l)}{||\hat{\boldsymbol{A}}(k,l)||_2^2}$ [10, 11]. The differential signal is obtained with

$$\boldsymbol{U}(k,l) = \boldsymbol{B}^{\boldsymbol{H}}(k,l)\boldsymbol{Z}(k,l) \approx \boldsymbol{B}^{\boldsymbol{H}}(k,l)\boldsymbol{N}(k,l), \tag{11}$$

which destructively adds the speech components, such that $\boldsymbol{B}^{\boldsymbol{H}}(k,l)\hat{\boldsymbol{A}}(k,l) \triangleq \boldsymbol{0}$. In effect, the speech signal is canceled out. $\boldsymbol{B}(k,l)$ can be identified as a blocking matrix [10, 12] which forms a spatial zero towards the speech source [9]. A very straightforward blocking matrix is given by $\boldsymbol{B}(k,l) = \boldsymbol{I}_{\boldsymbol{M\times M}} - \hat{\boldsymbol{A}}(k,l)\boldsymbol{F}^{\boldsymbol{H}}(k,l)$. We used a more efficient sparse blocking matrix, which is discussed in detail in [13]. The spatial correlation matrix of the noise reference is given by

$$\begin{aligned}\boldsymbol{\Phi}_{\boldsymbol{UU}}(k,l) &\triangleq E\{\boldsymbol{U}(k,l)\boldsymbol{U}^{\boldsymbol{H}}(k,l)\}\\ &\approx \boldsymbol{B}^{\boldsymbol{H}}(k,l)\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)\boldsymbol{B}(k,l)\Phi_N(k,l),\end{aligned} \tag{12}$$

using Eqn. (3) and the assumptions above. An estimate of the noise PSD $\Phi_N(k,l)$ is then obtained by:

$$\hat{\Phi}_N(k,l) = \frac{\operatorname{Tr}\left(\boldsymbol{\Phi}_{\boldsymbol{UU}}(k,l)\right)}{\operatorname{Tr}\left(\boldsymbol{B}^{\boldsymbol{H}}(k,l)\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)\boldsymbol{B}(k,l)\right)}. \tag{13}$$

In a similar fashion, the PSD of the summation signal is given by

$$\begin{aligned}\Phi_{YY}(k,l) &\triangleq E\{Y(k,l)Y^*(k,l)\}\\ &= \boldsymbol{F}^{\boldsymbol{H}}\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}}\boldsymbol{F}\Phi_S + \boldsymbol{F}^{\boldsymbol{H}}\boldsymbol{\Gamma}_{\boldsymbol{NN}}\boldsymbol{F}\Phi_N,\end{aligned} \tag{14}$$

where we omitted the frequency and frame indices for brevity. Solving for $\Phi_S(k,l)$ gives an estimate of the speech PSD

$$\hat{\Phi}_S(k,l) = \max\left(\Phi_{YY} - \boldsymbol{F}^{\boldsymbol{H}}\boldsymbol{\Gamma}_{\boldsymbol{NN}}\boldsymbol{F}\hat{\Phi}_N, 0\right), \tag{15}$$

with $\boldsymbol{F}^{\boldsymbol{H}}\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}}\boldsymbol{F} \approx 1$. Following the signal model in Eqn. (3), the estimated spatial correlation matrices for the noise and speech signals are obtained by:

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_{\boldsymbol{NN}}(k,l) &= \boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)\hat{\Phi}_N(k,l)\\ \hat{\boldsymbol{\Phi}}_{\boldsymbol{SS}}(k,l) &= \hat{\boldsymbol{A}}(k,l)\hat{\boldsymbol{A}}^{\boldsymbol{H}}(k,l)\hat{\Phi}_S(k,l).\end{aligned} \tag{16}$$

Many algorithms do not estimate the speech PSD $\hat{\Phi}_S(k,l)$ separately, since $\boldsymbol{\Phi}_{\boldsymbol{SS}}(k,l) = \boldsymbol{\Phi}_{\boldsymbol{ZZ}}(k,l) - \boldsymbol{\Phi}_{\boldsymbol{NN}}(k,l)$. However, in practice this will cause over-subtraction in the PMWF. As a consequence, musical artifacts may appear in the output signal.

5. Parametric multi-channel Wiener filter

Having an estimate of both the noise and speech PSD matrices, a parametric multi-channel noise reduction Wiener filter [7] can be formulated as:

$$\boldsymbol{h}_{\boldsymbol{PMWF}}(k,l) = \frac{\hat{\boldsymbol{\Phi}}^{-1}_{\boldsymbol{NN}}(k,l)\hat{\boldsymbol{\Phi}}_{\boldsymbol{SS}}(k,l)\boldsymbol{F}(k,l)}{\zeta(k,l)+\mu(k,l)}, \tag{17}$$

where $\zeta(k,l) = \text{Tr}\left(\hat{\boldsymbol{\Phi}}^{-1}_{\boldsymbol{NN}}(k,l)\hat{\boldsymbol{\Phi}}_{\boldsymbol{SS}}(k,l)\right)$ can be identified as the multi-channel SNR [7]. Heuristically, we chose the trade-off parameter to be $\mu(k,l) = \frac{1}{\zeta(k,l)}$. For low $\mu(k,l)$, the PMWF is close to the MVDR filter [9,14]. And for high $\mu(k,l)$ there is a sufficient amount of noise reduction. Finally, a MMSE estimate of the clean speech signal at the first microphone is obtained by $X(k,l) = \boldsymbol{h}^{\boldsymbol{H}}_{\boldsymbol{PMWF}}(k,l)\boldsymbol{Z}(k,l)$. Clearly, estimation errors in the ATFs $\hat{\boldsymbol{A}}(k,l)$, and the possible oversimplification of the noise sound field $\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k)$ will degrade the overall performance. We found this degradation to be mainly caused by residual noise, and not by missing speech components. By inserting Eqn. (16) into (17), and using $\boldsymbol{F}(k,l) = \frac{\hat{\boldsymbol{A}}(k,l)}{||\hat{\boldsymbol{A}}(k,l)||_2^2}$ it can be seen that:

$$\boldsymbol{h}_{\boldsymbol{PMWF}}(k,l) = \frac{\boldsymbol{\Gamma}^{-1}_{\boldsymbol{NN}}(k)\hat{\boldsymbol{A}}(k,l)}{\hat{\boldsymbol{A}}^{\boldsymbol{H}}(k,l)\boldsymbol{\Gamma}^{-1}_{\boldsymbol{NN}}(k)\hat{\boldsymbol{A}}(k,l)}\frac{\zeta(k,l)}{\zeta(k,l)+\mu(k,l)}. \tag{18}$$

This result can be identified as the MVDR filter [9] multiplied by a SNR-dependent gain function. Given that the MVDR filter does not distort signals defined by the ATFs, the PMWF output $X(k,l)$ contains the same speech components as the summation signal $Y(k,l)$ [9, 10]. We could greatly enhance the noise reduction performance by updating the speech PSD $\hat{\Phi}_S(k,l)$ from Eqn. (15). For this update, $\Phi_{XX}(k,l) \triangleq E\{X(k,l)X^*(k,l)\}$ is used instead of $\Phi_{YY}(k,l)$, so that Eqn. (15) turns into:

$$\hat{\Phi}'_S(k,l) = \max\left(\Phi_{XX} - \boldsymbol{F}^{\boldsymbol{H}}\boldsymbol{\Gamma}_{\boldsymbol{NN}}\boldsymbol{F}\hat{\Phi}_N, 0\right). \tag{19}$$

The updated speech PSD is then used to iterate Eqn. (16) and (17) a second time, which removes almost all residual noise components and preserves the speech components identified in the first run.

6. Summary of the DD-SNR algorithm

The proposed DD-SNR algorithm consists of three main parts: The calculation of the DD-SNR, the estimation of the ATFs, and the calculation of the PMWF. It can be summarized as follows:

1. Calculate the spatial coherence matrix $\boldsymbol{\Gamma}_{\boldsymbol{ZZ}}(k,l)$ using Eqn. (5) and (2).
2. Define a range for Θ and maximize $\xi_\Theta(k,l)$ using Eqn. (4) and (7), and $\Theta_{OPT} = \arg\max_{\Theta} \frac{1}{K}\sum_{k=0}^{K}\xi_\Theta(k,l)$.
3. Recursively update the ATFs $\hat{\boldsymbol{A}}(k,l)$ using Eqn. (8).
4. Calculate the speech and noise PSDs using Eqn. (15) and (13).
5. Obtain the speech estimate $X(k,l) = \boldsymbol{h}^{\boldsymbol{H}}_{\boldsymbol{PMWF}}(k,l)\boldsymbol{Z}(k,l)$ using the PMWF in Eqn. (17).
6. Update the speech PSD $\hat{\Phi}'_S(k,l)$ using Eqn. (19), and iterate Eqn. (16) and (17) a second time to obtain the final result.

7. Performance evaluation

7.1. Experimental Setup

We used 2 microphones with a distance of $d = 5cm$ in an approximately 8x5m wide hall to record multi-channel speech and noise tracks. The noise source is located $5m$ apart from the array, and the speaker source $0.5m$, to get the desired far and near sound fields. The speech and noise tracks have been mixed together with a signal-to-interference ratio (SIR) ranging from -20dB to +20dB. To test the algorithm against a significant amount of speech data, the TIMIT [15], KCORS [16] and KCOSS [17] speech corpora are employed. For the noise data, recordings from various sources, i.e. traffic noise, industry parks, subway stations and the NOIZEUS database have been used. In total, 60 minutes of test material has been generated.

For comparison to other approaches, we implemented the aforementioned TBRR [1] and the MC-SPP [2]. To get the theoretical maximum performance, the ground truth (i.e., the true speech and noise PSD matrices $\boldsymbol{\Phi}_{\boldsymbol{SS}}(k)$ and $\boldsymbol{\Phi}_{\boldsymbol{NN}}(k)$) has been used to construct a PMWF in an additional implementation. A sampling frequency of $f_s = 16$ kHz and a FFT size of 512 bins with an overlap of 50% is used for each algorithm. To evaluate the performance of the algorithms in terms of perceptual speech quality, the PESQ score and the PEASS toolkit [5,6] are used. The latter explicitly aims at the psycho-acoustically motivated quality assessment of audio source separation algorithms. It delivers four scores: The *Target Perceptual Score* (TPS) measures the perceptual quality of the desired speech signal contained in the enhanced output. The *Interference Perceptual Score* (IPS) measures the influence of the residual noise components. The *Artifact Perceptual Score* (APS) measures the influence of artificial artifacts like musical noise, and the *Overall Perceptual Score* (OPS) provides a global measure of the perceptual quality. Each score ranges from 0 to 100 and large values indicate better performance.

7.2. Performance results

In figure 1, a comparison in terms of the PESQ score is given. The ground truth marks the theoretical limit for the algorithms. It can be seen that the DD-SNR algorithm provides a significant improvement over the TBRR and MC-SPP algorithms, especially when the second iteration is included. For a SIR of +5dB, the DD-SNR achieves an improvement in PESQ of about 1.6 points.

Figure 2 shows the PEASS scores of the algorithms. The OPS score for the TBRR and the MCSPP are very similar. A reason could be that they rely on the same Gaussian model for the speech presence probability. Our DD-SNR approach relies on the ATFs and the noise sound field model, and outperforms the other algorithms especially for high SIRs. The TBRR and DD-SNR show the highest TPS and IPS scores, which indicates a higher speech intelligibility and a higher amount of noise reduction. However, the TBRR seems to introduce the most artifacts, as the APS score indicates. A possible reason could be that this algorithm relies on recursive noise floor averaging, which is known to introduce artifacts for instationary noises.

For demonstration, the spectrograms of the signal at the first microphone $z_1(t)$, the output of the DD-SNR algorithm $x(t)$, and the SNR $\zeta(k,l)$ after the first and second iteration are shown in figures 3 through 6. In this experiment, 15s of KCOSS speech data have been mixed with instationary city traffic noise with a SIR of 0dB, using the setup described above. The benefit of the second iteration of the DD-SNR algorithm can be seen by comparing figures 5 and 6: The second iteration removes almost all residual noise while preserving the speech components.

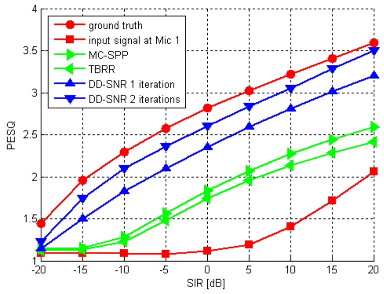

Figure 1: Comparison of the algorithms in terms of PESQ, using the MOS-LQO scale.

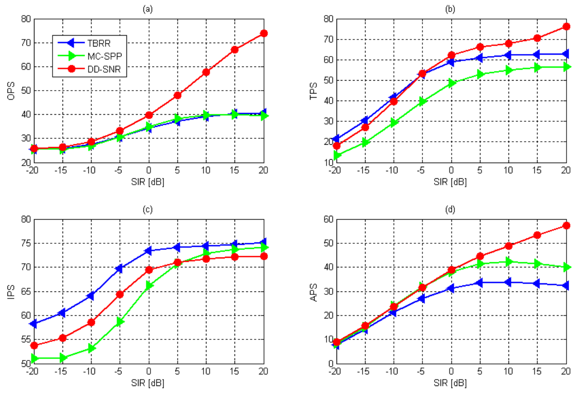

Figure 2: Comparison of the algorithms using PEASS measures; (a) OPS, (b) TPS, (c) IPS and (d) APS.

8. Conclusions

We proposed a multi-channel speech enhancement algorithm that blindly estimates the speech and noise PSDs based on a diffuse noise sound field and the DD-SNR. A PMWF is used to obtain a MMSE estimate of the desired clean speech signal. The overall performance is greatly increased by improving the estimate of the speech PSD using the output of the PMWF in a second iteration. We demonstrated that the algorithm outperforms similar approaches by using the PESQ and PEASS measures.

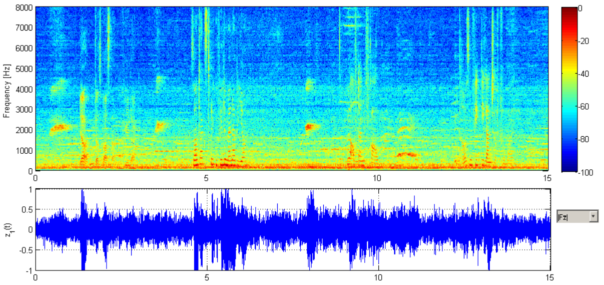

Figure 3: Received signal $z_1(t)$ at the first microphone, containing non-stationary city traffic noise with a SIR of 0dB in a setup with $M = 2$ microphones and $d_{12} = 5$cm.

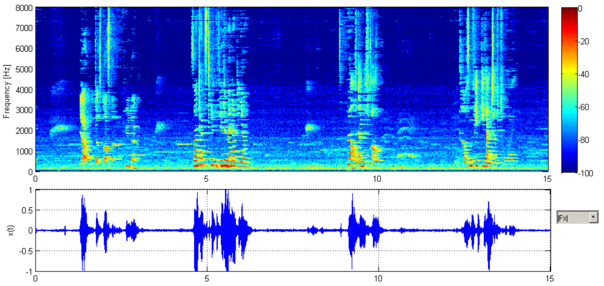

Figure 4: Output $x(t)$ of the DD-SNR algorithm after the second iteration. The gain in SNR is limited to 30dB.

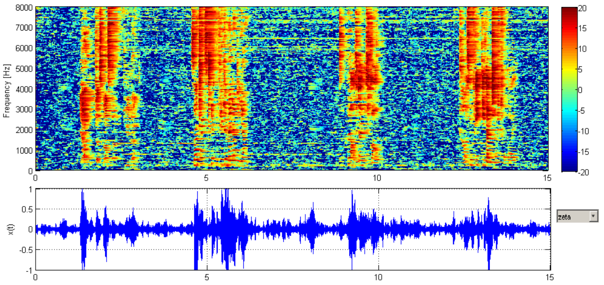

Figure 5: SNR $\zeta(k,l)$ and output $x(t)$ of the DD-SNR algorithm after the first iteration.

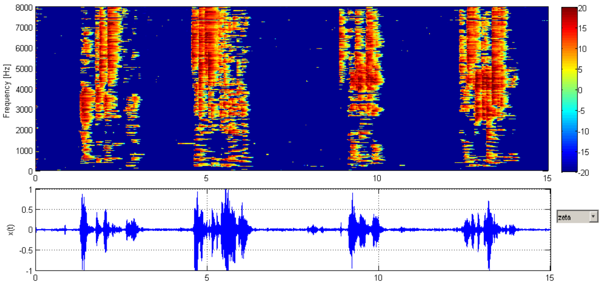

Figure 6: SNR $\zeta(k,l)$ and output $x(t)$ of the DD-SNR algorithm after the second iteration.

9. References

[1] S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function gsc and postfiltering,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, Nov. 2004.

[2] M. Souden, J. Chen, J. Benesty, and S. Affes, “An integrated solution for online multichannel noise tracking and reduction,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sep. 2011.

[3] M. Taseska and E. A. Habets, “Mmse-based blind source extraction in diffuse noise fields using a complex coherence-based a priori sap estimator,” *International Workshop on Acoustic Signal Enhancement*, Sep. 2012.

[4] O. Thiergart, G. D. Galdo, and E. A. P. Habets, “Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones,” Int. Audio Labs. Erlangen, Erlangen, Germany, Tech. Rep., 2012.

[5] E. Vincent, “Improved perceptual metrics for the evaluation of audio source separation,” INRIA, Centre de Rennes, Bretange Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France, Tech. Rep., 2011.

[6] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sep. 2011.

[7] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin–Heidelberg–New York: Springer, 2006.

[8] H. Kuttruff, *Room Acoustics*, 5th ed. London–New York: Spoon Press, 2009.

[9] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin–Heidelberg–New York: Springer, 2008.

[10] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin–Heidelberg–New York: Springer, 2008.

[11] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, “On microphone-array beamforming from a mimo acoustic signal processing perspective,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, Mar. 2007.

[12] I. Cohen, “Analysis of two-channel generalized sidelobe canceller (gsc) with post-filtering,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, Nov. 2003.

[13] M. G. Shmulik, S. Gannot, and I. Cohen, “A sparse blocking matrix for multiple constraints gsc beamformer,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012.

[14] P. Vary and R. Martin, *Digital Speech Transmission*. West Sussex: Wiley, 2006.

[15] “Timit acoustic-phonetic continuous speech corpus,” Website, available online at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1; visited on January 16th 2013.

[16] “The kiel corpus of read speech vol. 1,” Website, available online at http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html; visited on January 16th 2013.

[17] “The kiel corpus of spontaneous speech vol. 1-3,” Website, available online at http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html; visited on January 16th 2013.