Blind Adaptive Principal Eigenvector Beamforming for Acoustical Source Separation

Ernst Warsitz, Reinhold Haeb-Umbach, Dang Hai Tran Vu

University of Paderborn, Dept. of Communications Engineering
33098 Paderborn, Germany
{warsitz,haeb,tran}@nt.uni-paderborn.de

Abstract

For separating multiple speech signals given a convolutive mixture, time-frequency sparseness of the speech sources can be exploited. In this paper we present a multi-channel source separation method based on the concept of approximate disjoint orthogonality of speech signals. Unlike binary masking of single-channel signals as e.g. applied in the DUET algorithm we use a likelihood mask to control the adaptation of blind principal eigenvector beamformers. Furthermore orthogonal projection of the adapted beamformer filters leads to mutually orthogonal filter coefficients thus enhancing the demixing performance. Experimental results in terms of the achievable signal-to-interference ratio (SIR) and a perceptual speech quality measure are given for the proposed method and are compared to the DUET algorithm.

Index Terms: BSS, DUET, PCA, Beamforming

1. Introducion

For hands-free multi-channel telecommunication or auditory scene analysis in a scenario of simultaneously active sources it can be interesting to estimate not only one source signal, as is usually done in adaptive beamforming (ABF) [1], but all source signals from their mixtures at the sensors. ABF is a form of spatial filtering enhancing signals from the look direction and attenuating signals from other directions. Its performance depends heavily on the estimation of the direction-of-arrival (DoA), and the adaptation to an individual source is rather difficult if sources are simultaneously active. On the other hand, blind source separation (BSS) methods need no information of the array geometry and speaker direction, and demixing is possible even when the sources are simultaneously active.

In the last decade BSS has focused on methods based on independent component analysis (ICA). They involve higher order statistics and non-linear cost functions and are usually computationally quite expensive. To deal with convolutive sound mixtures ICA can be employed in the frequency domain for each frequency bin separately [2], however at the expense of permutation ambiguity. To overcome the problems of BSS and ABF, combined methods have been proposed [3, 4]: geometrically constrained BSS algorithms solve the frequency permutation problem of BSS and they are less sensitive to DoA estimation errors than ABF. It should be noted, that geometrically constrained separation is not blind anymore.

From a physical point of view the separation of two sources by frequency-domain BSS can be shown to be equivalent to null-beamforming by two frequency-domain ABFs. In both cases the algorithms attempt to reduce the interference signal by forming a spatial null in the interference direction [5]. The performance of BSS is limited by that of perfectly adapted beamformers [6].

Here we propose to separate multiple speech signals by exploiting the time-frequency sparseness (disjoint orthogonality) of the speech sources [7] to control the adaptation of blind principal eigenvector beamformers. The dominant eigenvector or principal component generates a beam towards the direction of maximum power which can also be interpreted as matched-filtering for the unknown mixing system [8, 9]. Adaptation is only carried out, when the source of interest is dominant. We derive an adaptation control mechanism for the case of two sources by use of a so-called likelihood mask. In opposition to the hard decision of binary masking applied in the DUET algorithm [7, 10] we use some kind of soft decision to control the principal eigenvector adaptation.

In eigenvector beamforming the filter coefficients are optimized to form a spatial pattern with a maximum in the direction of the dominant source, however without spatial minima at the interferer directions. Therefore we additionally apply mutual orthogonal projection (MOP) of the eigenvector filters to enforce orthogonality and to place implicitly a null or at least a minimum in the interferer direction.

2. Principal Eigenvector Beamforming

We are given an array of M microphones. The DFTs of the microphone signals are denoted by $X_i(k,m)$, $i = 1,\ldots,M$. Here, m and k denote the frame index and frequency bin, respectively. The output of a first Filter-and-Sum beamformer is

$$\begin{aligned} Y_1(k,m) &= \sum_{i=1}^{M} F_{1,i}^*(k,m)\cdot X_i(k,m) \\ &= \mathbf{F}_1^H(k,m)\cdot\mathbf{X}(k,m), \end{aligned} \tag{1}$$

where $\mathbf{F}_1(k,m) = (F_{1,1}(k,m),..,F_{1,M}(k,m))^T$ is the filter coefficient vector and $\mathbf{X}(k,m) = (X_1(k,m),..,X_M(k,m))^T$ is the vector of input signals. Let us assume for the moment that a single speech source signal $U_1(k,m)$ is present, which is connected to the i-th microphone by the transfer function $H_{1,i}(k)$. The vector of microphone signals is then given by

$$\mathbf{X}(k,m) = \mathbf{H}_1(k)U_1(k,m) + \mathbf{N}(k,m) \tag{2}$$

where $\mathbf{H}_1(k) = (H_{1,1}(k),..,H_{1,M}(k))^T$, and $\mathbf{N}(k,m)$ is a vector of stationary noise terms which are assumed to be spatially white. Our goal is to determine a vector of filter coefficients $\mathbf{F}_1(k,m)$ such that the signal-to-noise ratio of the output $Y_1(k,m)$ is maximized. The frequency dependent SNR is maximized by the eigenvector $\mathbf{v}_{max_1}(k)$ corresponding to the

largest eigenvalue of the cross power spectral density matrix of the microphone signals, which turns out to be

$$\mathbf{v}_{max_1}(k) = \zeta_1(k) \cdot \mathbf{H}_1(k), \tag{3}$$

where $\zeta_1(k) \neq 0$ is an arbitrary complex scalar. We can compute $\mathbf{F}_1(k,m)$ iteratively to be the desired eigenvector (3), by employing a two-step stochastic gradient ascent algorithm (see also [9]):

$$\begin{aligned}\tilde{\mathbf{F}}_1(k,m+1) &= \mathbf{F}_1(k,m) \\ &+\mu_1(k,m)Y_1^*(k,m)\left[\mathbf{X}(k,m) - \mathbf{F}_1(k,m)Y_1(k,m)\right] \quad (4)\\ \mathbf{F}_1(k,m+1) &= \tilde{\mathbf{F}}_1(k,m+1)\frac{1+\tilde{\mathbf{F}}_1^H(k,m+1)\tilde{\mathbf{F}}_1(k,m+1)}{2\tilde{\mathbf{F}}_1^H(k,m+1)\tilde{\mathbf{F}}_1(k,m+1)}.\end{aligned}$$

Due to the block frequency nature of the algorithm a frequency-dependent step size $\mu_1(k,m)$, which is inversely proportional to the power levels in the DFT frequency bins, can be used. Since (4) performs a principal component analysis (PCA) the resulting multi-channel filtering can be called PCA-beamforming.

3. Time-Frequency Masking

Now we extend the problem to a multi-speaker scenario with P sources $\mathbf{U}(k,m) = (U_1(k,m),..,U_P(k,m))^T$, $M \geq P$ microphones and P beamformers. The relation between the source signals and the microphone signals is given by

$$\mathbf{X}(k,m) = \sum_{l=1}^{P} \mathbf{H}_l(k)U_l(k,m) + \mathbf{N}(k,m). \tag{5}$$

Denoting the mixing system as

$$\mathbf{H}(k) = [\mathbf{H}_1(k), \mathbf{H}_2(k), .., \mathbf{H}_P(k)] \tag{6}$$

(5) can be written as

$$\mathbf{X}(k,m) = \mathbf{H}(k)\mathbf{U}(k,m) + \mathbf{N}(k,m). \tag{7}$$

The concept of disjoint orthogonality to estimate the source signals $U_l(k,m)$ giving the mixtures $\mathbf{X}(k,m)$ is based on the time-frequency sparseness of the speech sources [7]. The underlying assumption is that

$$U_l(k,m)U_\lambda(k,m) = 0, \quad \forall k,m \quad l \neq \lambda, \tag{8}$$

i.e. at no time-frequency bin two sources are simultaneously active.

3.1. Binary Masking (BM)

Based on the assumption (8) a binary mask for approximate disjoint orthogonality is introduced:

$$\alpha_l^{(\mathrm{BM})}(k,m) = \begin{cases} 1, & \text{if} \quad U_l(k,m) > \beta U_\lambda(k,m) \\ 0, & \text{else} \end{cases}, \tag{9}$$

where β is a heuristic parameter. The demixed output is then obtained as

$$\hat{U}_l(k,m) = \alpha_l^{(\mathrm{BM})}(k,m)X_1(k,m). \tag{10}$$

In the DUET algorithm the mask is estimated by exploiting phase and amplitude information of two sensors [7]. Unfortunately, reliable estimate of $\alpha_l^{(\mathrm{BM})}(k,m)$ are hard to obtain in a reverberant environment. Hence, the demixed speech sources will suffer from serious signal distortion.

3.2. Likelihood-based Masking (LM)

As opposed to (9) we propose to use a soft-valued mask based on some kind of source activity probability:

$$\alpha_l^{(\mathrm{LM})}(k,m) \approx p(|U_l(k,m)| \gg |U_\lambda(k,m)| \,|\, \mathbf{X}),\ \forall \lambda; l \neq \lambda. \tag{11}$$

This so-called likelihood mask can assume any value between Zero and One and can be used to control the adaptation of the corresponding l-th beamformer by replacing the step size $\mu_l(k,m)$ of the gradient update rule by:

$$\tilde{\mu}_l(k,m) = \mu_l(k,m)\alpha_l^{(\mathrm{LM})}(k,m). \tag{12}$$

Note that the adaptation stops, if sources other than the target source are too strong ($\alpha_l^{(\mathrm{LM})}(k,m) \rightarrow 0$) or if the beamformer has adapted towards its dominant source ($\mathbf{X}(k,m) - \mathbf{F}_l(k,m)Y_l(k,m) \rightarrow 0$). As the optimal beamformer is a matched-filter, see eq. (3), the mixing system (6) is thus estimated column-wise (by adapting to the currently dominant source) in every time-frequency bin.

3.3. Symmetric Adaptive Decorrelation (SAD)

For the case of two sources $P = 2$ and $l \in \{1,2\}$, a simple but effective likelihood estimator $\alpha_l^{(\mathrm{LM})}(k,m)$ can be obtained as follows. Motivated by [11] symmetrical adaptive cross filters $D_{i,j}(k,m)$ between two neighboring microphones i and j, where $i = 1,..,M-1, \quad j = i+1$ for decorrelating two neighboring microphone signals are used:

$$\begin{aligned} Z_{j,i}(k,m) &= X_i(k,m) - D_{j,i}(k,m)X_j(k,m) \quad (13)\\ Z_{i,j}(k,m) &= X_j(k,m) - D_{i,j}(k,m)X_i(k,m). \quad (14)\end{aligned}$$

The adaptation of (13) and (14) can be performed by normalized least mean square (NLMS) algorithms

$$D_{j,i}(k,m+1) = D_{j,i}(k,m) + \mu_{j,i}(k,m)Z_{j,i}(k,m)Z_{i,j}^*(k,m), \tag{15}$$

from which it can be seen that adaptation stops if $Z_{j,i}$ and $Z_{i,j}$ are decorrelated. The ratio

$$\alpha_{i,j}(k,m) = \frac{|Z_{i,j}(k,m)|^2}{|Z_{i,j}(k,m)|^2 + |Z_{j,i}(k,m)|^2} \tag{16}$$

serves as estimate for the dominance of one speech source over the other based on the microphone pair i and $j = i+1$. These estimates are combined to the "likelihood" mask

$$\alpha_l^{(\mathrm{LM})}(k,m) = \prod_{i=1}^{M-1} \alpha_{i,i+1}(k,m), \tag{17}$$

where $l \in \{1,2\}$ is the one of the two sources which is spatially closer to the i-th than to the $(i+1)$-st microphone. The more microphone pairs are used, the more reliable likelihood estimations are obtained. In the limit of infinitely many microphones (17) will become a binary mask.

4. PCA-Beamforming Post-Processing

While the optimal filter coefficients $\mathbf{F}_l(k,m)$ form a spatial pattern with main lobe in the direction of the l-th source, no spatial selectivity is performed regarding the interferer. Thus the l-th output $Y_l(k,m) = \mathbf{F}_l^H(k,m)\mathbf{X}(k,m)$ has a good speech quality but contains also strong interfering signals from other

sources, especially at low frequencies. To gain in SIR we apply mutual orthogonal projection of the eigenvector filters by

$$\tilde{\mathbf{W}}_l(k,m) = \left(\prod_{\lambda;\lambda\neq l} [\mathbf{I} - \mathbf{F}_\lambda(k,m)\mathbf{F}_\lambda^H(k,m)] \right) \mathbf{F}_l(k,m), \quad (18)$$

with a following normalization step

$$\mathbf{W}_l(k,m) = \frac{\tilde{\mathbf{W}}_l(k,m)}{\| \tilde{\mathbf{W}}_l(k,m) \|}. \quad (19)$$

The beamformer outputs $\mathbf{Y}_{\mathrm{MOP}} = (Y_{l,\mathrm{MOP}}, \ldots, Y_{P,\mathrm{MOP}})^T$ are then given by

$$\mathbf{Y}_{\mathrm{MOP}}(k,m) = \mathbf{W}^H(k,m)\mathbf{X}(k,m). \quad (20)$$

where $\mathbf{W}(k,m) = [\mathbf{W}_1(k,m), .., \mathbf{W}_P(k,m)]$. Every output signal is thus determined by the subspace spanned by the remaining source signals ensured by the mutual orthogonal projection (note subscript MOP).

Beamforming approaches in general suffer from cross-talk in a multi-speaker scenario, although spatial minima are placed in the directions of the interferers. This is due to multipath signal propagation in a reverberant environment when reflected interference components are captured by the main lobe. As the interfering components can be interpreted as noise, the idea is to use a post-filter similar to the Wiener post-filter (WP) in beamforming [1] or speech enhancement. As estimate of the interferers' power present in the l-th beamformer output we use the squared outputs of the other beamformers. The post-filtering completes the proposed demixing process and the final source estimates are

$$\hat{U}_l(k,m) = \frac{|Y_{l,\mathrm{MOP}}(k,m)|^2}{\mathbf{Y}_{\mathrm{MOP}}^H(k,m)\mathbf{Y}_{\mathrm{MOP}}(k,m)} Y_{l,\mathrm{MOP}}(k,m). \quad (21)$$

Fig. 1 shows the overall structure of the proposed demixing system including the corresponding equation numbers.

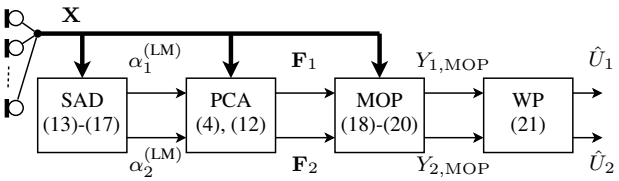

Figure 1: *Blind PCA-based source separation system.*

5. Simulation Results

In this section we experimentally evaluate the proposed blind eigenbeamformer source separation method for the case of two simultaneously active sources in a simulated reverberant enclosure of the size (6 m) x (5 m) x (3 m). 10 utterances from different speakers (5 male and 5 female) sampled at 16 kHz were used as source speech signals. Taking 2 out of the 10 speakers gives 45 mixing setups for each of 9 analyzed reverberation times T_{60} between 0 s and 0.5 s, totalling 405 audio data. One speech source was placed at 45° and the other source at -30° relative to broadside, each at a distance of 2 m. The power ratio of the two sources was about 0 dB. To every microphone speech signal white noise with a signal-to-noise ratio of about 25 dB was added.

For the PCA-based source separation method a linear array of $M = 8$ equally spaced microphones (distance of 4 cm) with 2 BFs and 7 SADs were used. The FIR filters had a length of 256 taps each, and the DFT length was set to 512 taps. All given simulation results are with steady state filter coefficients after convergence. For the online DUET algorithm [10] two microphones with a distance of 2.4 cm were employed.

Since we mix the signals artificially we know each of the reverberated microphone speech components and can compute a perfect binary mask. The optimum achievable performance obtained this way is analyzed in terms of the signal-to-interference ratio power (SIR) and the perceptual similarity measure (PSM) [12]. PSM has been shown to give comparable objective perceptual quality evaluation results as the well-known PESQ measure [13]. The PSM[1] and SIR values are shown in the upper and lower subplot of Fig. 2, respectively. The reference signal for PSM was the target speech signal at one microphone (without noise and interference) and the test signal to be evaluated was the processed target speech signal according to (9). Since we want to separate two sources two estimates $\hat{U}_1(k,m)$ and $\hat{U}_2(k,m)$ are available. For each utterance the better and the worse of the two (as measured by PSM) were collected in separate sets. In all figures "PSM-high" and "PSM-low" denote the averaged PSM on the output with the better and lower PSM scores, respectively. All results are averages over 45 utterances per T_{60}. "PSM-mean" is the mean of the two. The SIR scores were also evaluated separately for each of the two sets.

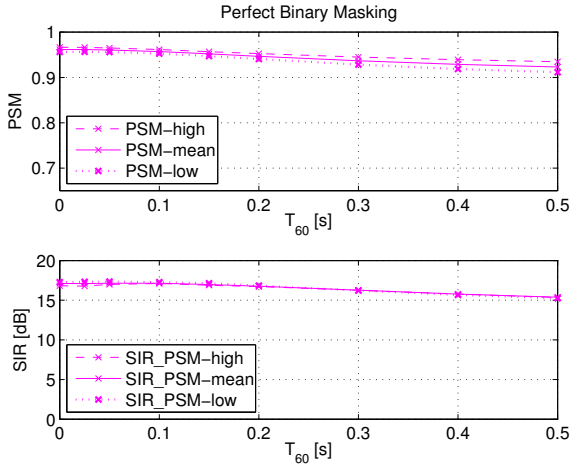

Figure 2: *Upper figure: Collected outputs with high and low PSM values separately and overall PSM results (PSM-mean). Lower figure: SIR separately for the outputs with high and low PSM and the overall SIR (SIR_PSM-mean).*

It is very remarkable that the performance of the perfect masking process works well for all simulated reverberation times and both outputs. This shows that disjoint orthogonality (8) is indeed a promising approach. However, the crucial issue is how to estimate the mask in a reverberant scenario.

Next, the performance degradation is examined when the binary mask is estimated online by the DUET algorithm [10], see Fig. 3. With increasing T_{60} the speech quality differs dramatically between the two outputs (PSM-high, PSM-low). This is due to the unequal demixing performance: the binary decision (9) is estimated to 1 for one output much more often than for the other output. Consequently the quality of target speech of that output, where the signal is passed through more often, is very good, however at the cost of reduced SIR compared to

[1]PSM results are generated with level alignment, lowpass filtering and without assimilation.

the other output. For that other one the speech quality is bad, because it was "switched off" very often. At the same time, however, the SIR is higher.

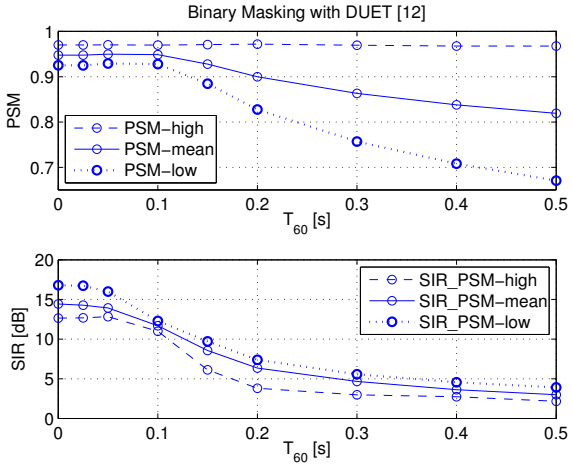

Figure 3: *Upper figure: Collected outputs with high and low PSM values separately and overall PSM results (PSM-mean). Lower figure: SIR separately for the outputs with high and low PSM and the overall SIR (SIR_PSM-mean).*

To overcome the unbalanced performance between the different outputs of the DUET algorithm the PCA-based source separation method is proposed, see Fig. 4. It can be seen, that the speech quality is high for both outputs, and the SIR of the two is also similar. To give an impression of the adaptation

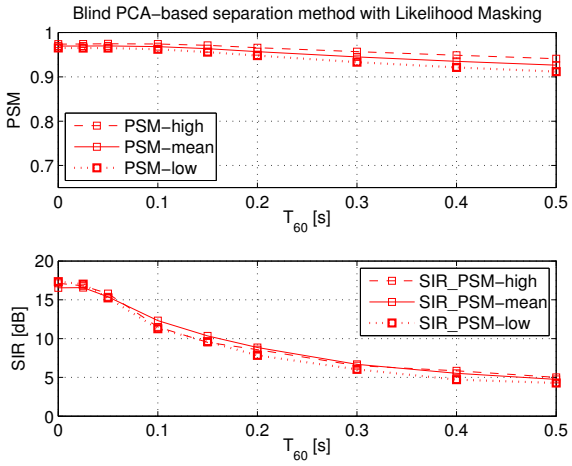

Figure 4: *Upper figure: Collected outputs with high and low PSM values separately and overall PSM results (PSM-mean). Lower figure: SIR separately for the outputs with high and low PSM and the overall SIR (SIR_PSM-mean).*

speed, in Fig. 5 the SIR for the proposed method is plotted over time. We chose the case of no reverberation ($T_{60} = 0$ s) to have the highest dynamic range of the SIR. The initial filter coefficients of the SAD were set to zero and the PCA coefficients were chosen randomly. Fig. 5 shows the fast adaptation speed of the proposed separating method.

6. Conclusions

A blind source separation method based on principal eigenvector beamforming exploiting the time-frequency sparseness of speech signals has been presented. The advantage of high

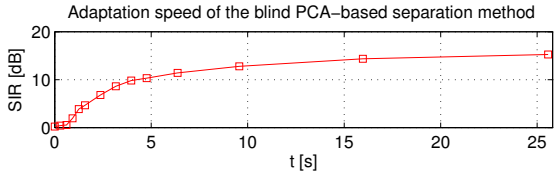

Figure 5: *Adaptation speed of the principal eigenvector separation, averaged over all runs for* $T_{60} = 0$ *s.*

speech quality obtained by adaptive beamforming is combined with blind source classification. Important properties of the proposed method are as follows: First, the beamformer makes no a priori assumptions on the acoustic enclosure and neither the source directions nor the array geometry need to be known. Second, the separation is carried out in two steps: a frequency bin-wise likelihood estimation of the dominant source, followed by principal eigenvector beamforming. This results in robust, fast adaptation behavior and a flexible structure. Third, the interference rejection is increased by mutually orthogonal projection of the filter coefficients and post-filtering.

7. Acknowledgement

This work was in part supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/4-1.

8. References

[1] H. L. Van Trees, *Optimum Array Processing*, John Wiley, New York 2001.

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-Domain Blind Source Separation", in *Speech Enhancement*, Springer, 2005.

[3] L. Parra and C.V. Alvino, "Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming", in *IEEE Trans. on Speech and Audio Processing* vol. 10, no. 6, Sep. 2002.

[4] M. Knaak, S. Araki and S. Makino, "Geometrically constrained Independent Component Analysis", *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 15, no. 2, pp 715-726, Feb. 2007.

[5] A. Shoko, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa and H. Saruwatari, "Equivalence between Frequency-Domain Blind Source Separation and Frequency-Domain Adaptive Beamforming for Convolutive Mixtures", *EURASIP Journal on Applied Signal Processing*, vol. 2003 (2003), issue 11, pp. 1157-1166.

[6] S. Makino, "Blind Source Separation of Convolutive Mixtures of Speech", in *Adaptive Signal Processing*, Springer, 2003.

[7] O. Yilmaz and S. Richard, "Blind Separation of Speech Mixtures via Time-Frequency Masking", in *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830-1847, July 2004.

[8] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech", in *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 425-437, Sep. 1997.

[9] E. Warsitz and R. Haeb-Umbach, "Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis", in *Proc. IEEE ICASSP*, Philadelphia, Mar. 2005.

[10] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation" in *Proc. ICA 2001*, San Diego, Dec. 2001

[11] B. Schölling, M. Heckmann, F. Joublin, and C. Goerick, "From Source Localization to Blind Source Separation: An Intuitive Route to Blind Source Separation", in *Proc. IWAENC*, Paris, Oct. 2006.

[12] R. Huber, "Objective assessment of audio quality using an auditory processing model", Ph.D. thesis, University of Oldenburg, Oldenburg, 2003.

[13] T. Rohdenburg, V. Hohmann and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes", in *Proc. IWAENC*, Eindhoven, Netherlands, Sep. 2005.