EURASIP Journal on Applied Signal Processing 2003:11, 1147–1156

Multichannel Direction-Independent Speech Enhancement Using Spectral Amplitude Estimation

Thomas Lotter
*Institute of Communication Systems and Data Processing, Aachen University (RWTH), Templergraben 55,
D-52056 Aachen, Germany
Email: lotter@ind.rwth-aachen.de*

Christian Benien
*Philips Research Center, Aachen, Weißhausstraße 2, D-52066 Aachen, Germany
Email: christian.benien@philips.com*

Peter Vary
*Institute of Communication Systems and Data Processing, Aachen University (RWTH), Templergraben 55,
D-52056 Aachen, Germany
Email: vary@ind.rwth-aachen.de*

Received 25 November 2002 and in revised form 12 March 2003

This paper introduces two short-time spectral amplitude estimators for speech enhancement with multiple microphones. Based on joint Gaussian models of speech and noise Fourier coefficients, the clean speech amplitudes are estimated with respect to the MMSE or the MAP criterion. The estimators outperform single microphone minimum mean square amplitude estimators when the speech components are highly correlated and the noise components are sufficiently uncorrelated. Whereas the first MMSE estimator also requires knowledge of the direction of arrival, the second MAP estimator performs a direction-independent noise reduction. The estimators are generalizations of the well-known single channel MMSE estimator derived by Ephraim and Malah (1984) and the MAP estimator derived by Wolfe and Godsill (2001), respectively.

Keywords and phrases: speech enhancement, microphone arrays, spectral amplitude estimation.

1. INTRODUCTION

Speech communication appliances such as voice-controlled devices, hearing aids, and hands-free telephones often suffer from poor speech quality due to background noise and room reverberation. Multiple microphone techniques such as beamformers can improve the speech quality and intelligibility by exploiting the spatial diversity of speech and noise sources. Upon these techniques, one can differentiate between fixed and adaptive beamformers.

A fixed beamformer combines the noisy signals by a time-invariant filter-and-sum operation. The filters can be designed to achieve constructive superposition towards a desired direction (delay-and-sum beamformer) or in order to maximize the SNR improvement (superdirective beamformer) [1, 2, 3].

Adaptive beamformers commonly consist of a fixed beamformer towards a fixed desired direction and an adaptive null steering towards moving interfering sources [4, 5].

All beamformer techniques assume the target direction of arrival (DOA) to be known a priori or assume that it can be estimated sufficiently enough. Usually the performance of such a beamforming system decreases dramatically if the DOA knowledge is erroneous. To estimate the DOA during runtime, time difference of arrival (TDOA)-based locators evaluate the maximum of a weighted cross correlation [6, 7]. Subspace methods have the ability to detect multiple sources by decomposing the spatial covariance matrix into a signal and a noise subspace. However, the performance of all DOA estimation algorithms suffers severely from reverberation and directional or diffuse background noise.

Single microphone speech enhancement frequency domain algorithms are comparably robust against reverberation and multiple sources. However, they can achieve high noise reduction only at the expense of moderate speech distortion. Usually, such an algorithm consists of two parts. Firstly, a noise power spectral density estimator based on the assumption that the noise is stationary to a much higher

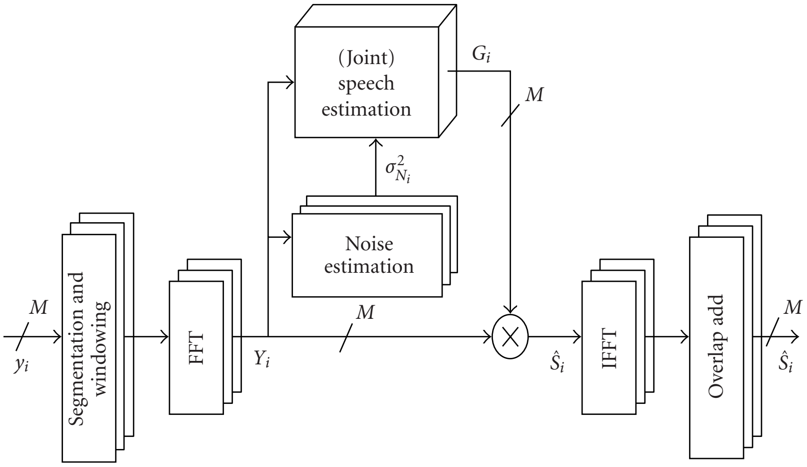


Figure 1: Multichannel noise reduction system.

degree than the speech. The noise power spectral density can be estimated by averaging discrete Fourier transform (DFT) periodograms in speech pauses using a voice activity detection or by tracking minima over a sliding time window [8]. Secondly, an estimator for the speech component of the noisy signal with respect to an error criterion. Commonly, a Wiener filter, the minimum mean square error (MMSE) estimator of the speech DFT amplitudes [9], or its logarithmic extension [10] are applied.

In this paper, we propose the extensions of two single channel speech spectral amplitude estimators for the use in microphone array noise reduction. Clearly, multiple noisy signals offer a higher-estimation accuracy possibility when the desired signals are highly correlated and the noise components are uncorrelated to a certain degree. The main contribution will be a joint speech estimator that exploits the benefits of multiple observations but achieves a DOA-independent speech enhancement.

Figure 1 shows an overview of the multichannel noise reduction system with the proposed speech estimators. The noisy time signals $y_i(k)$, $i \in \{1,\dots,M\}$, from M microphones are transformed into the frequency domain. This is done by applying a window $h(\mu)$, for example, a Hann window, to a frame of K consecutive samples and by computing the DFT on the windowed data. Before the next DFT computation, the window is shifted by Q samples. The resulting complex DFT values $Y_i(\lambda, j)$ are given by

$$Y_i(\lambda,k) = \sum_{\mu=0}^{K-1} y_i(\lambda Q+\mu)h(\mu)e^{-j2\pi k\mu/L}. \qquad (1)$$

Here, k denotes the DFT bin and λ the subsampled time index. For the sake of brevity, k and λ are omitted in the following.

The noisy DFT coefficient Y_i consists of complex speech $S_i = A_i e^{j\alpha_i}$ and noise N_i components:

$$Y_i = R_i e^{j\vartheta_i} = A_i e^{j\alpha_i} + N_i, \quad i \in \{1,\dots,M\}. \qquad (2)$$

The noise variances $\sigma^2_{N_i}$ are estimated separately for each channel and are fed into a speech estimator. If $M = 1$, the minimum mean square short-time spectral amplitude (MMS-STSA) estimator [9], its logarithmic extension [10], or less complex maximum a posteriori (MAP) estimators [11] can be applied to calculate real spectral weights G_1 for each frequency. If $M > 1$, a joint estimator can exploit information from all M channels using a joint statistical model of the DFT coefficients after IFFT and overlap-add M noise-reduced signals are synthesized. Since the phases are not modified, a beamformer could be applied additionally after synthesis.

The remainder of the paper is organized as follows. Section 2 introduces the underlying statistical model of multichannel Fourier coefficients. In Section 3, two new multichannel spectral amplitude estimators are derived. First, a minimum mean square estimator that evaluates the expectation of the speech spectral amplitude conditioned on all noisy complex DFT coefficients is described. Secondly, a MAP estimator conditioned on the joint observation of all noisy amplitudes is proposed. Finally, in Section 4, the performance of the proposed estimators in ideal and realistic conditions is discussed.

2. STATISTICAL MODELS

Motivated by the central limit theorem, real and imaginary parts of both speech and noise DFT coefficients are usually modelled as zero-mean independent Gaussian [9, 12, 13] with equal variance. Recently, MMSE estimators of the complex DFT spectrum S have been developed with Laplacian or Gamma modelling of the real and imaginary parts of the speech DFT coefficients [14]. However, for MMSE or MAP estimation of the speech spectral amplitude, the Gaussian model facilitates the derivation of the estimators. Due to the unimportance of the phase, estimation of the speech spectral amplitude instead of the complex spectrum is more suitable from a perceptual point of view [15].

The Gaussian model leads to Rayleigh distributed speech amplitudes A_i, that is,

$$p(A_i, \alpha_i) = \frac{A_i}{\pi\sigma_{S_i}^2}\exp\left(-\frac{A_i^2}{\sigma_{S_i}^2}\right). \tag{3}$$

Here, $\sigma_{S_i}^2$ describes the variance of the speech in channel i. Moreover, the pdfs of the noisy spectrum Y_i and noisy amplitude R_i conditioned on the speech amplitude and phase are Gaussian and Ricians, respectively,

$$p(Y_i|A_i, \alpha_i) = \frac{1}{\pi\sigma_{N_i}^2}\exp\left(-\frac{\left|Y_i - A_i e^{j\alpha_i}\right|^2}{\sigma_{N_i}^2}\right), \tag{4}$$

$$p(R_i|A_i) = \frac{2R_i}{\sigma_{N_i}^2}\exp\left\{-\frac{R_i^2 + A_i^2}{\sigma_{N_i}^2}\right\} I_0\left(\frac{2A_iR_i}{\sigma_{N_i}^2}\right). \tag{5}$$

Here, I_0 denotes the modified Bessel function of the first kind and zeroth order. To extend this statistical model for multiple noisy signals, we consider the typical noise reduction scenario of Figure 2, for example, inside a room or a car. A desired signal s arrives at a microphone array from angle θ. Multiple noise sources arrive from various angles. The resulting diffuse noise field can be characterized by its coherence function. The magnitude squared coherence (MSC) between two omnidirectional microphones i and j of a diffuse noise field is given by

$$\mathrm{MSC}_{ij}(f) = \frac{\left|\Phi_{ij}(f)\right|^2}{\Phi_{ii}(f)\Phi_{jj}(f)} = \mathrm{si}^2\left(\frac{2\pi f d_{ij}}{c}\right). \tag{6}$$

Figure 3 plots the theoretical coherence of an ideal diffuse noise field and the measured coherence of the noise field inside a crowded cafeteria with a microphone distance of $d_{ij} = 12\,\mathrm{cm}$. For frequencies above $f_0 = c/2d_{ij}$, the MSC becomes very low and thus the noise components of the noisy spectra can be considered uncorrelated with

$$E\{N_i N_j^*\} = \begin{cases} \sigma_{N_i}^2, & i = j, \\ 0, & i \neq j. \end{cases} \tag{7}$$

Hence, (5) and (4) can be extended to

$$p(R_1, \ldots, R_M|A_n) = \prod_{i=1}^{M} p(R_i|A_n), \tag{8}$$

$$p(Y_1, \ldots, Y_M|A_n, \alpha_n) = \prod_{i=1}^{M} p(Y_i|A_n, \alpha_n), \tag{9}$$

for each $n \in \{1, \ldots, M\}$. We assume the time delay of the speech signals between the microphones to be small compared to the short-time stationarity of speech and thus assume the speech spectral amplitudes A_i to be highly correlated. However, due to near-field effects and different microphone amplifications, we allow a deviation of the speech amplitudes by a constant channel-dependent factor c_i, that is, $A_i = c_i \cdot A$ and $\sigma_{S_i}^2 = c_i^2\sigma_S^2$. Thus we can express $p(R_i|A_i = (c_i/c_n)A_n) = p(R_i|A_n)$. The joint pdf of all noisy amplitudes R_i given the speech amplitude of channel n can then be written as

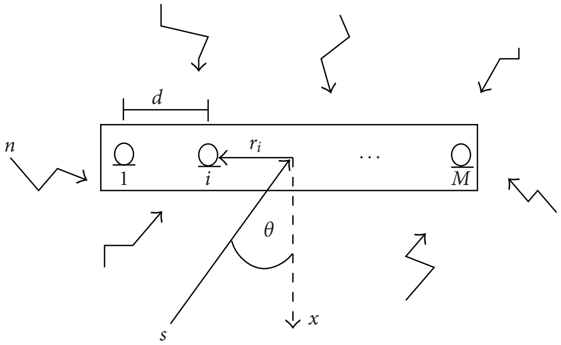


Figure 2: Speech and noise arriving at microphone array.

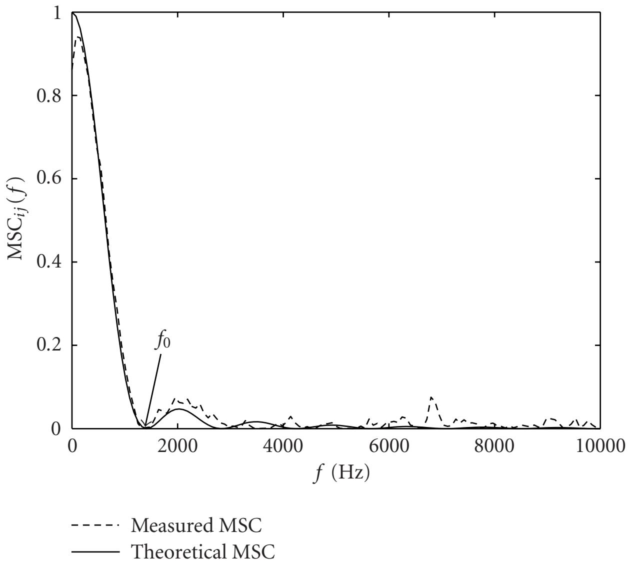


Figure 3: Theoretical MSC of a diffuse noise field and measured MSC inside a crowded cafeteria ($d_{ij} = 0.12\,\mathrm{m}$).

$$p(R_1, \ldots, R_M|A_n) = \exp\left\{-\sum_{i=1}^{M}\frac{R_i^2 + (c_i/c_n)^2 A_n^2}{\sigma_{N_i}^2}\right\} \cdot \prod_{i=1}^{M}\left[\frac{2R_i}{\sigma_{N_i}^2} I_0\left(\frac{2(c_i/c_n)A_nR_i}{\sigma_{N_i}^2}\right)\right], \tag{10}$$

where the c_i's are fixed parameters of the joint pdf. Similarly, the pdf of all noisy spectra Y_i conditioned on the clean speech amplitude and phase is

$$p(Y_1, \ldots, Y_M|A_n, \alpha_n) = \prod_{i=1}^{M}\frac{1}{\pi\sigma_{N_i}^2} \cdot \exp\left(-\sum_{i=1}^{M}\frac{\left|Y_i - (c_i/c_n)A_n e^{j\alpha_i}\right|^2}{\sigma_{N_i}^2}\right). \tag{11}$$

The unknown phases α_i can be expressed by α_n, the DOA, and the DFT frequency.

In analogy to the single channel MMSE estimator of the speech spectral amplitudes, the resulting joint estimators will be formulated in terms of a priori and a posteriori SNRs

$$\xi_i = \frac{\sigma_{S_i}^2}{\sigma_{N_i}^2}, \qquad \gamma_i = \frac{R_i^2}{\sigma_{N_i}^2}, \tag{12}$$

whereas the a posteriori SNRs γ_i can be directly computed, the a priori SNRs ξ_i are recursively estimated using the estimated speech amplitude $\hat{A}_i$ of the previous frame [9]:

$$\begin{aligned} \hat{\xi}_i(\lambda) &= \alpha \frac{\hat{A}_i^2(\lambda - 1)}{\sigma_{N_i}^2} + (1 - \alpha) P(\gamma_i(\lambda) - 1) \\ &\text{with } P(x) = \begin{cases} x, & x > 0, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{13}$$

The smoothing factor α controls the trade-off between speech quality and noise reduction [16].

3. MULTICHANNEL SPECTRAL AMPLITUDE ESTIMATORS

We derive Bayesian estimators of the speech spectral amplitudes A_n, $n \in \{1, \ldots, M\}$, using information from all M channels. First, a straightforward multichannel extension of the well-known MMSESTSA by Ephraim and Malah [9] is derived. Second, a practically more useful MAP estimator for DOA-independent noise reduction is introduced. All estimators output M spectral amplitudes A_n and thus M-enhanced signals are delivered by the noise reduction system.

3.1. Estimation conditioned on complex spectra

The single channel algorithm for channel number n derived by Ephraim and Malah calculates the expectation of the speech spectral amplitude A conditioned on the observed complex Fourier coefficient Y_n, that is, $E\{A_n | Y_n\}$. In the multichannel case, we can condition the expectation of each of the speech spectral amplitudes A_n on the joint observation of all M noisy spectra Y_i. To estimate the desired spectral amplitude of channel n, we have to calculate

$$\begin{aligned} \hat{A}_n &= E\{A_n | Y_1, \ldots, Y_M\} \\ &= \int_0^\infty \int_0^{2\pi} A_n p(A_n, \alpha_n | Y_1, \ldots, Y_M) \, d\alpha_n \, dA_n. \end{aligned} \tag{14}$$

This estimator can be expressed via Bayesian rule as

$$\hat{A}_n = \frac{\int_0^\infty A_n \int_0^{2\pi} p(A_n, \alpha_n) p(Y_1, \ldots, Y_M | A_n, \alpha_n) \, d\alpha_n \, dA_n}{\int_0^\infty \int_0^{2\pi} p(A_n, \alpha_n) p(Y_1, \ldots, Y_M | A_n, \alpha_n) \, d\alpha_n \, dA_n}. \tag{15}$$

To solve (15), we assume perfect DOA correction, that is, $\alpha_i := \alpha$ for all $i \in \{1, \ldots, M\}$. Inserting $A_i = (c_i/c_n) A_n$ in (9) and (4), the integral over α in (15) becomes

$$\begin{aligned} I &= \int_0^{2\pi} \exp\left\{ -\sum_{i=1}^{M} \frac{|Y_i - (c_i/c_n) A_n e^{\alpha_n}|^2}{\sigma_{N_i}^2} \right\} d\alpha \\ &= \exp\left\{ -\sum_{i=1}^{M} \frac{|Y_i|^2 + ((c_i/c_n) A_n)^2}{\sigma_{N_i}^2} \right\} \\ &\quad \times \int_0^{2\pi} \exp\{p \cos\alpha + q \sin\alpha\} \, d\alpha \end{aligned} \tag{16}$$

with

$$\begin{aligned} p &= \sum_{i=1}^{M} \frac{2 c_i A_n}{c_n \sigma_{N_i}^2} \operatorname{Re}\{Y_i\}, \\ q &= \sum_{i=1}^{M} \frac{2 c_i A_n}{c_n \sigma_{N_i}^2} \operatorname{Im}\{Y_i\}. \end{aligned} \tag{17}$$

The sum of sine and cosine is a cosine with different amplitude and phase:

$$p \cos\alpha + q \sin\alpha = \sqrt{p^2 + q^2} \cos\left(\alpha - \arctan\left(\frac{p}{q} \right) \right). \tag{18}$$

Since we integrate from 0 to 2π, the phase shift is meaningless. With

$$\sqrt{p^2 + q^2} = \left| \sum_{i=1}^{M} \frac{(c_i/c_n) Y_i}{\sigma_{N_i}^2} \right| \tag{19}$$

and $\int_0^\pi \exp\{z \cos x\} dx = \pi I_0(z)$, the integral becomes

$$\begin{aligned} I &= 2\pi \exp\left\{ -\sum_{i=1}^{M} \frac{|Y_i|^2 + ((c_i/c_n) A_n)^2}{\sigma_{N_i}^2} \right\} \\ &\quad \times I_0\left(2 A_n \left| \sum_{i=1}^{M} \frac{(c_i/c_n) Y_i}{\sigma_{N_i}^2} \right| \right). \end{aligned} \tag{20}$$

The remaining integrals over A_n can be solved using [17, equation (6.631.1)]. After some straightforward calculations, the gain factor for channel n is expressed as

$$\begin{aligned} G_n = \frac{\hat{A}_n}{|Y_n|} &= 1.5\gamma \cdot \sqrt{\frac{\xi_n}{\gamma_n \left(1 + \sum_{i=1}^{M} \xi_i \right)}} \\ &\quad \cdot F_1\left(-0.5, 1, \frac{\left| \sum_{i=1}^{M} \sqrt{\gamma_i \xi_i} e^{j\vartheta_i} \right|^2}{1 + \sum_{i=1}^{M} \xi_i} \right), \end{aligned} \tag{21}$$

where F_1 denotes the confluent hypergeometric series and Γ the Gamma function. The argument of F_1 contains a sum of a priori and a posteriori SNRs with respect to the noisy phases ϑ_i, $i \in \{1, \ldots, M\}$. The confluent hypergeometric series F_1 has to be evaluated only once since the argument is independent of n. Note that in case of $M = 1$, (21) is the single channel MMSE estimator derived by Ephraim and Malah. In a practical real-time implementation, the confluent hypergeometric series is stored in a table.

3.2. Estimation conditioned on spectral amplitudes

The assumption $\alpha_i := \alpha$, $i \in \{1, \dots, M\}$, introduces a DOA dependency since this is only given for speech from $\theta = 0°$ or after perfect DOA correction. For a DOA-independent speech enhancement, we condition the expectation of A_n on the joint observation of all noisy amplitudes R_i, that is, $\hat{A}_n = E\{A_n | R_1, \dots, R_M\}$.

When the time delay of the desired signal s in Figure 2 between the microphones is small compared to the short-time stationarity of speech, the noisy amplitudes R_i are independent of the DOA θ. Unfortunately, after using (10), we have to integrate over a product of Bessel functions, which leads to extremely complicated expressions even for the simple case $M = 2$.

Therefore, searching for a closed-form estimator, we investigate a MAP solution which has been characterized in [11] as a simple but effective alternative to the mean square estimator in the single channel application.

We search for the speech spectral amplitude $\hat{A}_n$ that maximizes the pdf of A_n conditioned on the joint observation of all R_i, $i \in \{1, \dots, M\}$:

$$\begin{aligned}\hat{A}_n &= \arg\max_{A_n} p(A_n | R_1, \dots, R_M) \\ &= \arg\max_{A_n} \frac{p(R_1, \dots, R_M | A_n)\, p(A_n)}{p(R_1, \dots, R_M)}.\end{aligned} \tag{22}$$

We need to maximize only $L = p(R_1, \dots, R_M | A_n) \cdot p(A_n)$ since $p(R_1, \dots, R_M)$ is independent of A_n. It is however easier to maximize $\log(L)$, without effecting the result, because the natural logarithm is a monotonically increasing function. Using (10) and (3), we get

$$\begin{aligned}\log L = {} & \log\left(\frac{A_n}{\pi\sigma_{S_n}^2}\right) - \frac{A_n^2}{\sigma_{S_n}^2} \\ & + \sum_{i=1}^{M}\left[\log\left(\frac{2R_i}{\sigma_{N_i}^2}\right) - \frac{R_i^2 + (c_i/c_n)^2 A_n^2}{\sigma_{N_i}^2}\right. \\ & \left. + \log\left(I_0\left(2\frac{(c_i/c_n)A_n R_i}{\sigma_{N_i}^2}\right)\right)\right].\end{aligned} \tag{23}$$

A closed-form solution can be found if the modified Bessel function I_0 is considered asymptotically with

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}} e^x. \tag{24}$$

Figure 4 shows that the approximation is reasonable for larger arguments and becomes erroneous only for very low SNRs.

Thus the term in the likelihood function containing the Bessel function is simplified to

$$\begin{aligned}& \log\left(I_0\left(2\frac{(c_i/c_n)A_n R_i}{\sigma_{N_i}^2}\right)\right) \\ & \approx \frac{2(c_i/c_n)A_n R_i}{\sigma_{N_i}^2} - \frac{1}{2}\log\left(4\pi\frac{(c_i/c_n)A_n R_i}{\sigma_{N_i}^2}\right).\end{aligned} \tag{25}$$

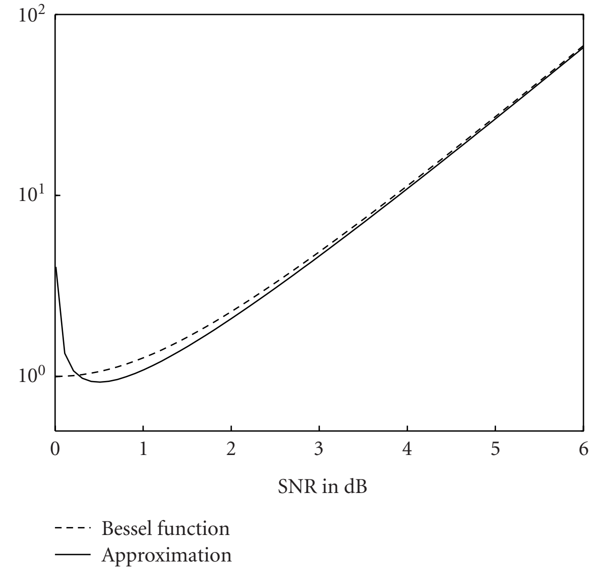


Figure 4: Bessel function and its approximation, $2(c_i/c_n A_n R_i)/\sigma_{N_i}^2 \approx 2\sqrt{\xi_i\gamma_i}$.

Differentiation of $\log L$ and multiplication with the amplitude A_n results in $A_n(\partial(\log L)/\partial A_n) = 0$:

$$A_n^2\left(-\frac{1}{\sigma_{S_n}^2} - \sum_{i=1}^{M}\frac{(c_i/c_n)^2}{\sigma_{N_i}^2}\right) + A_n\sum_{i=1}^{M}\frac{(c_i/c_n)R_i}{\sigma_{N_i}^2} + \frac{2-M}{4} = 0. \tag{26}$$

This quadratic expression can have two zeros; for $M > 2$, it is also possible that no zero is found. In this case, the apex of the parabolic curve in (26) is used as an approximation identical to the real part of the complex solution. The resulting gain factor of channel n is given as

$$\begin{aligned}G_n &= \frac{\hat{A}_n}{|Y_n|} \\ &= \frac{\sqrt{\xi_n/\gamma_n}}{2 + 2\sum_{i=1}^{M}\xi_i} \\ &\quad \cdot \operatorname{Re}\left\{\sum_{i=1}^{M}\sqrt{\gamma_i\xi_i} + \sqrt{\left(\sum_{i=1}^{M}\sqrt{\gamma_i\xi_i}\right)^2 + (2-M)\left(1 + \sum_{i=1}^{M}\xi_i\right)}\right\}.\end{aligned} \tag{27}$$

For the calculation of the gain factors, no exotic function needs to be evaluated any more. Also, $\operatorname{Re}\{\cdot\}$ has to be calculated only once since the argument is independent of n. Again, if $M = 1$, we have the single channel MAP estimator as given in [11].

4. EXPERIMENTAL RESULTS

In this section, we compare the performance of the joint speech spectral amplitude estimators with the well-known single channel Ephraim and Malah algorithm. Both M single channel estimators and the joint estimators output M-enhanced signals. In all experiments, we do not apply additional (commonly used) soft weighting techniques [9, 13] in order to isolate the benefits of the joint speech estimators compared to the single channel MMSE estimator.

All estimators were embedded in the DFT-based noise reduction system in Figure 1. The system operates at a sampling frequency of $f_s = 20\,\text{kHz}$ using half-overlapping Hann windowed frames. Both noise power spectral density $\sigma^2_{N_i}$ and variance of speech $\sigma^2_{S_i}$ were estimated separately for each channel. For the noise estimation task, we applied an elaborated version of *minimum statistics* [8] with adaptive recursive smoothing of the periodograms and adaptive bias compensation that is capable of tracking nonstationary noise even during speech activity.

To measure the performance, the noise reduction filter was applied to speech signals with added noise for different SNRs. The resulting filter was then utilized to process speech and noise separately [18]. Instead of only considering the segmental SNR improvement obtained by the noise reduction algorithm, this methods allows separate tracking of speech quality and noise reduction amount. The tradeoff between speech quality and noise reduction amount can be regulated by, for example, changing the smoothing factor for the decision-directed speech power spectral density estimation (13). The speech quality of the noise-reduced signal was measured by averaging the segmental speech SNR between original and processed speech over all M channels. On the other hand, the amount of noise reduction was measured by averaging segmental input noise power divided by output noise power. Although the results presented here were produced with offline processing of generated or recorded signals, the system is well suited for real-time implementation.

The computational power needed is approximately M times that of one single channel Ephraim-Malah algorithm since for each microphone signal, an FFT, an IFFT, and an identical noise estimation algorithm are needed. The calculation of the a posteriori and a priori SNR (12) and (13) is also done independently for each channel. The joint estimators following (21) and (27) hardly increase the computational load, especially because $\text{Re}(\cdot)$ and $F_1(\cdot)$ need to be calculated only once per frame and frequency bin.

4.1. Performance in artificial noise

To study the performance in ideal conditions, we first utilize the estimators on identical speech signals disturbed by spatially uncorrelated white noise. Figures 5 and 6 plot noise reduction and speech quality of the noise-reduced signal averaged over all M microphones for different number of microphones. While in Figure 5 the multichannel MMSE estimators according to (21) were applied, Figure 6 shows the performance of the multichannel MAP estimators according to (27). All joint estimators provide a significant higher speech quality and noise attenuation than the single channel MMSE estimator. The performance gain increases with the number of used microphones. The MAP estimators conditioned on the noisy amplitudes deliver a higher noise reduc-

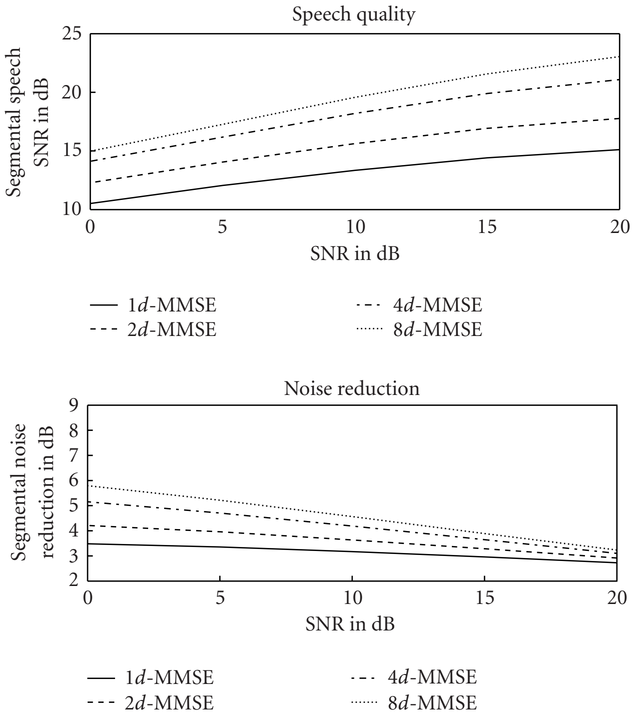


Figure 5: Speech quality and noise reduction of $1d$-MMSE estimators (reference) and Md-MMSE estimators with $M \in \{2, 4, 8\}$ for noisy signals containing identical speech and uncorrelated white noise.

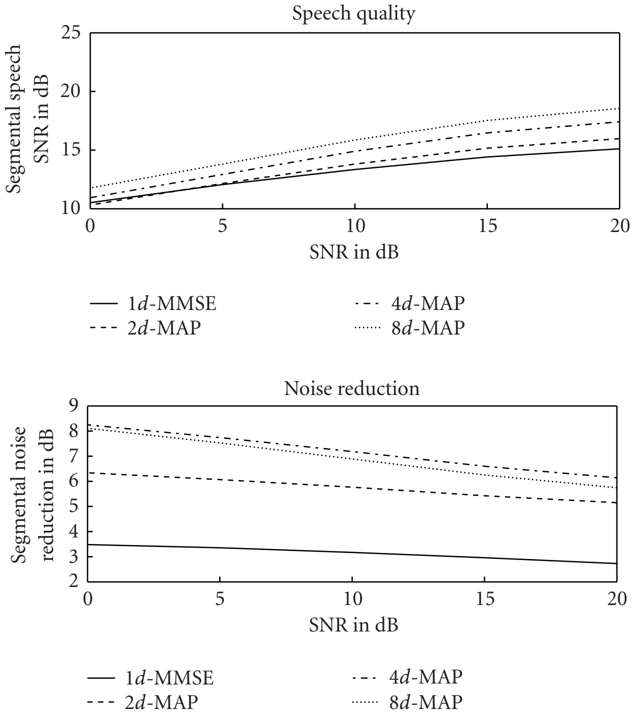


Figure 6: Speech quality and noise reduction of $1d$-MMSE estimators (reference) and Md-MAP with $M \in \{2, 4, 8\}$ for noisy signals containing identical speech and uncorrelated white noise.

tion than the multichannel MMSE estimator conditioned on the complex spectra at a lower speech quality. The gain in terms of noise reduction can be exchanged for a gain in terms of speech quality by different parameters.

4.2. Performance in realistic noise

Instead of uncorrelated white noise, we now mix the speech signal with noise recorded with a linear microphone array inside a crowded cafeteria. The coherence function of the approximately diffuse noise field is shown in Figure 3. Figure 7 plots the performance of the estimators using $M = 4$ microphones with an interelement spacing of $d = 12\,\mathrm{cm}$. Figure 8 shows the performance when using recordings with half the microphone distance, that is, $d = 6\,\mathrm{cm}$ interelement spacing. The $4d$-MAP estimator provides both higher speech quality and higher noise reduction amount than the Ephraim-Malah estimator. In both cases, the multichannel MMSE estimator delivers a much higher speech quality at an equal or lower noise reduction. According to (6), the noise correlation increases with decreasing microphone distance. Thus, the performance gain of the multichannel estimators decreases. However, Figures 7 and 8 illustrate that significant performance gains are found at reasonable microphone distances.

Clearly, if the noise is spatially coherent, no performance gain can be expected by the multichannel spectral amplitude estimators. Compared to the $1d$-MMSE, the Md-MMSE and Md-MAP deliver a lower noise reduction amount at a higher speech quality when applied to speech disturbed by coherent noise.

4.3. DOA dependency

We examine the performance of the estimators when changing the DOA of the desired signal. We consider desired sources in both far and near field with respect to an array of $M = 4$ microphones with $d = 12\,\mathrm{cm}$.

4.3.1. Desired signal in far field

The far-field model assumes equal amplitudes and angle-dependent TDOAs:

$$s_i(t) = s\left(t - \tau_i(\theta)\right), \qquad \tau_i = d\sin\left(\frac{\theta}{c}\right). \tag{28}$$

Figures 9 and 10 show the performance of the $4d$-estimators with cafeteria noise when the speech arrives from $\theta = 0^\circ, 10^\circ, 20^\circ$, or 60° (see Figure 2). The performance of the MMSE estimator conditioned on the noisy spectra decreases with increasing angle of arrival. The speech quality decreases significantly, while the noise reduction amount is only slightly affected. This is because the phase assumption $\alpha_i = \alpha$, $i \in \{1, \ldots, M\}$ is not fulfilled.

On the other hand, the performance of the multichannel MAP estimator conditioned on the spectral amplitudes shows almost no dependency on the DOA.

4.3.2. Desired signal in near field

We investigate the performance when the source of the desired signal is located in the near field with distance ρ_i to

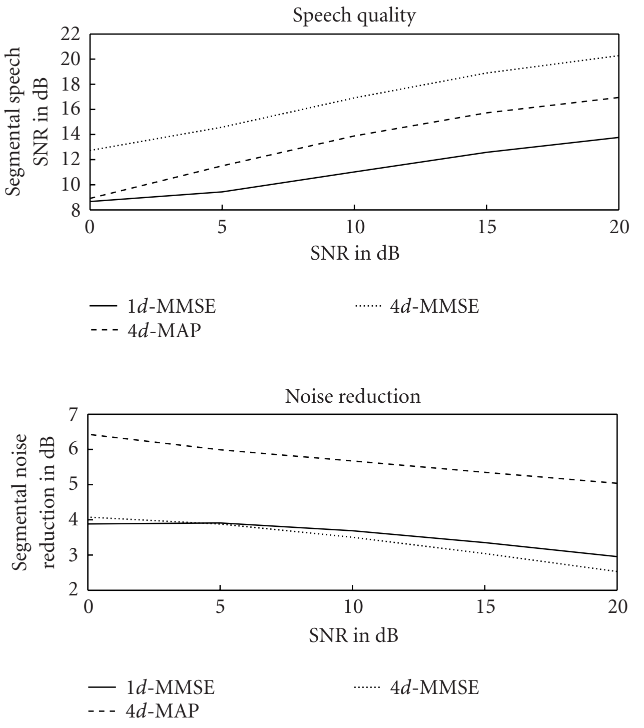


Figure 7: Speech quality and noise reduction of $1d/4d$-MMSE and $4d$-MAP for four signals containing identical speech and cafeteria noise (microphone distance $d = 12\,\mathrm{cm}$).

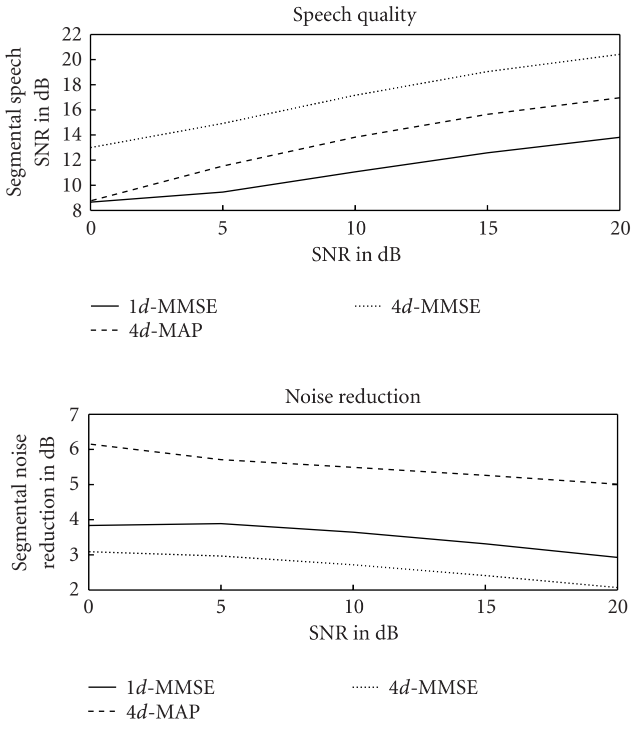


Figure 8: Speech quality and noise reduction of $1d/4d$-MMSE and $4d$-MAP for four signals containing identical speech and cafeteria noise (microphone distance $d = 6\,\mathrm{cm}$).

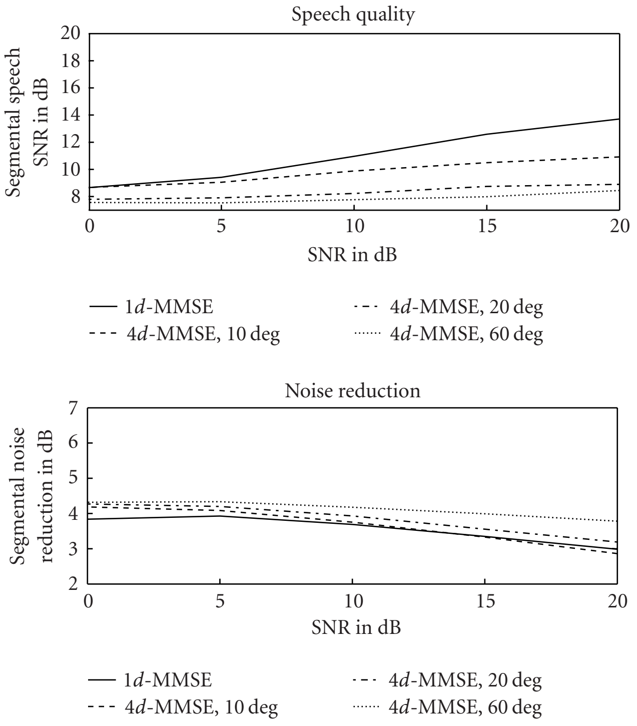


FIGURE 9: Speech quality and noise reduction of $4d$-MMSE compared to $1d$-MMSE for signals containing speech from $\theta = 10^\circ, 20^\circ$, and 60° and cafeteria noise (microphone distance $d = 12$ cm).

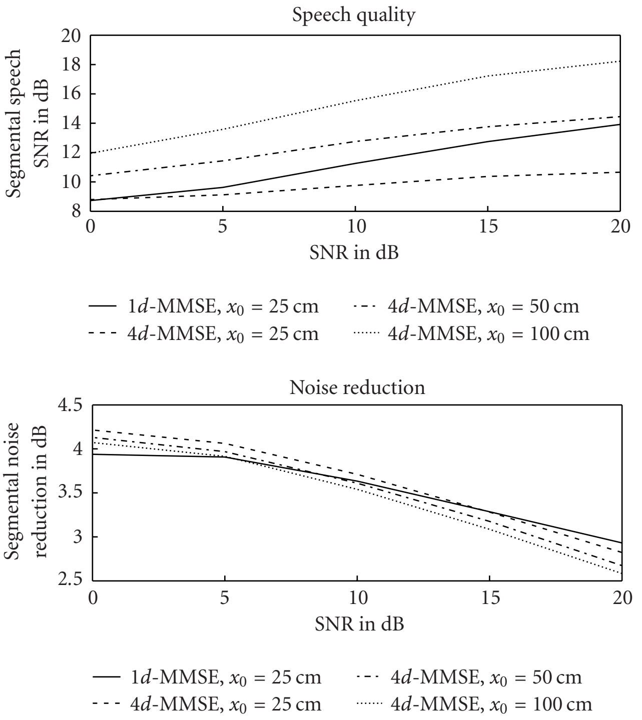


FIGURE 11: Speech quality and noise reduction of $4d$-MMSE compared to $1d$-MMSE for signals containing speech from $x_0 = 25$ cm, 50 cm, and 100 cm and cafeteria noise (microphone distance $d = 12$ cm).

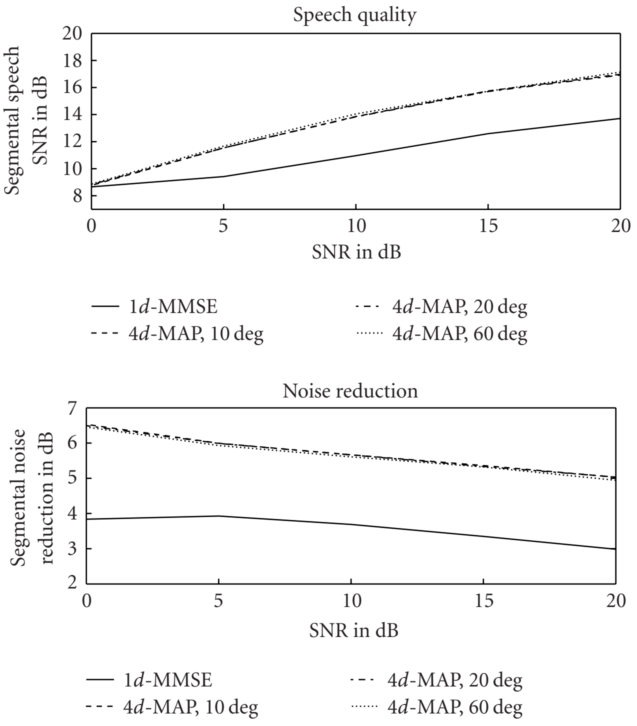


FIGURE 10: Speech quality and noise reduction of $4d$-MAP compared to $1d$-MMSE for signals containing speech from $\theta = 10^\circ, 20^\circ$ and 60° and cafeteria noise (microphone distance $d = 12$ cm).

microphone i. To simulate a near-field source, we use range-dependent amplifications and time differences:

$$s_i(t) = a_i s(t - \tau_i(\rho_i)), \tag{29}$$

where the amplitude factor for each channel decreases with the distance, $a_i \sim 1/\rho_i$. The source is located at different distances x_0 in front of the linear microphone array ($\theta = 0^\circ$) with $M = 4$ and $d = 12$ cm such that $\rho_i = \sqrt{x_0^2 + r_i^2}$, where r_i is defined in Figure 2.

Figures 11 and 12 show the performance of the $4d$-MMSE and $4d$-MAP estimators, respectively, when the source is located at $x_0 = 25$ cm, 50 cm, or 100 cm from the microphone array. The speech quality of the multichannel MMSE estimator decreases with decreasing distance. This is because at a higher distance from the microphone array, the time difference is smaller. Again, the multichannel MAP estimator conditioned on the noisy amplitudes shows nearly no dependency on the near-field position of the desired source.

4.4. *Reverberant desired signal*

Finally, we examine the performance of the estimators with a reverberant desired signal. Reverberation causes the spectral phases and amplitudes to become somewhat arbitrary, reducing the correlation of the desired signal. For the generation of reverberant speech signal, we simulate the acoustic situation depicted in Figure 13. The microphone array with

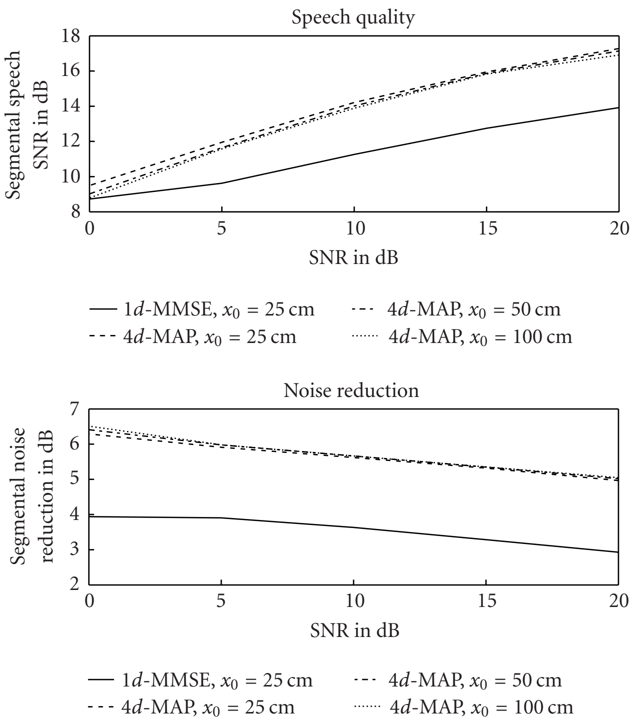


Figure 12: Speech quality and noise reduction of 4*d*-MAP compared to 1*d*-MMSE for signals containing speech from x_0 = 25 cm, 50 cm, and 100 cm and cafeteria noise (microphone distance d = 12 cm).

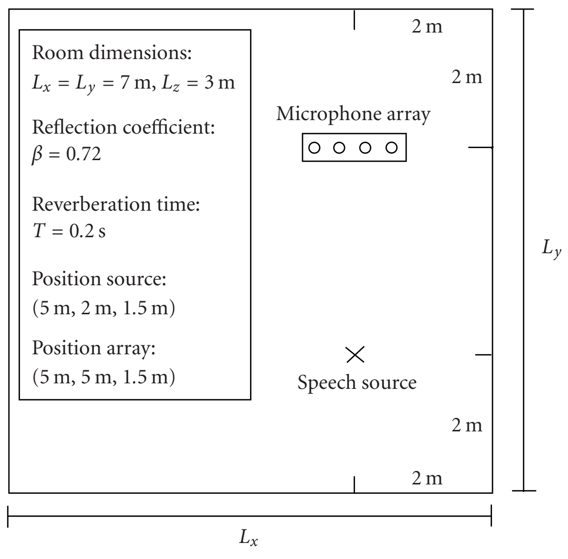


Figure 13: Speech source and microphone array inside a reverberant room.

$M = 4$ and an interelement spacing of $d = 12$ cm are positioned inside a reverberant room of size $L_x = 7$ m, $L_y = 7$ m, and $L_z = 3$ m. A speech source is located three meters in front of the array.

The acoustical transfer functions from the source to each microphone were simulated with the image method [19], which models the reflecting walls by several image sources. The intensity of the sound from an image source at the microphone array is determined by a frequency-independent reflection coefficient β and by the distance to the array.

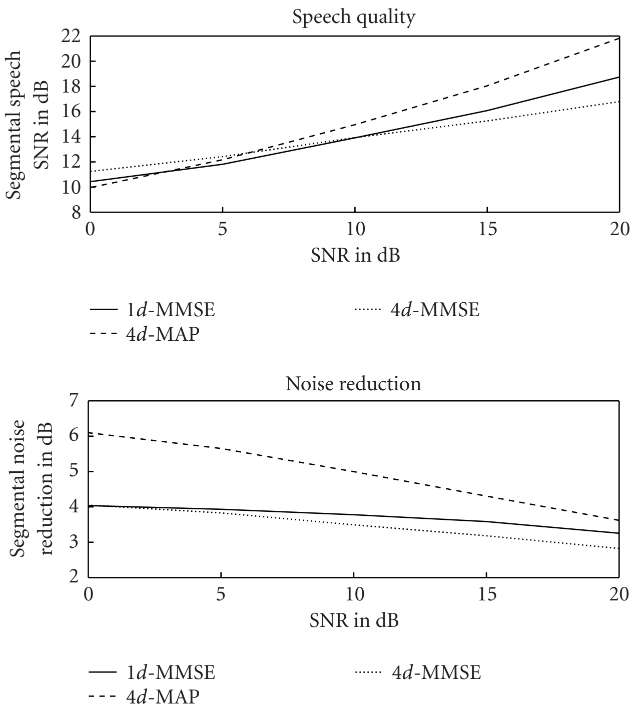


Figure 14: Speech quality and noise reduction of 1*d*/4*d*-MMSE and 4*d*-MAP for reverberant speech. (Figure 13) and cafeteria noise (microphone distance d = 12 cm).

In our experiment, the reverberation time was set to $T = 0.2$ second, which corresponds to a defection coefficient $\beta = 0.72$ according to Eyring's formula

$$\beta = \exp\left\{-\frac{13.82}{\left(c\left(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z}\right)T\right)}\right\}. \tag{30}$$

Figure 14 shows the performance of the estimators when the reverberant speech signal is mixed with cafeteria noise. As expected, the overall performance gain obtained by the multichannel estimators decreases. However, there is still a significant improvement by the multichannel MAP estimator conditioned on the spectral amplitudes left. The multichannel MMSE estimator conditioned on the complex spectra performs worse due to its sensitivity to phase errors caused by reverberation.

5. CONCLUSION

We have derived analytically a multichannel MMSE and a MAP estimator of the speech spectral amplitudes, which can be considered as generalizations of [9, 11] to the multichannel case. Both estimators provide a significant gain compared to the well-known Ephraim-Malah estimator when the highly correlated speech components are in phase and the noise components are sufficiently uncorrelated.

The MAP estimator conditioned on the noisy spectral amplitudes performs multichannel speech enhancement independent of the position of the desired source in the near or the far field and is only moderately susceptible to reverberation. The multichannel noise reduction system is well suited for real-time implementation. It outputs multiple enhanced signals which can be combined by a beamformer for additional speech enhancement.

ACKNOWLEDGMENT

The authors would like to thank Rainer Martin for many inspiring discussions.

REFERENCES

[1] E. Gilbert and S. Morgan, "Optimum design of directive antenna arrays subject to random variations," *Bell System Technical Journal*, vol. 34, pp. 637–663, May 1955.

[2] M. Dörbecker, *Multi-channel algorithms for the enhancement of noisy speech for hearing aids*, Ph.D. thesis, Aachener Beiträge zu digitalen Nachrichtensystemen, vol. 10, Wissenschaftsverlag Mainz, Aachen, Germany, 1998, Aachen University (RWTH), P. Vary, Ed.

[3] J. Bitzer and K. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 19–38, Springer-Verlag, Berlin, Germany, May 2001.

[4] L. Griffith and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[5] O. Hoshuyama and A. Sugiyama, "Robust adaptive beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 87–109, Springer-Verlag, Berlin, Germany, 2001.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 157–180, Springer-Verlag, Berlin, Germany, 2001.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[11] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Workshop on Statistical Signal Processing (SSP '01)*, pp. 496–499, Orchid Country Club, Singapore, August 2001.

[12] D. Brillinger, *Time Series, Data Analysis and Theory*, McGraw-Hill, New York, NY, USA, 1981.

[13] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[14] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, Orlando, Fla, USA, May 2002.

[15] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

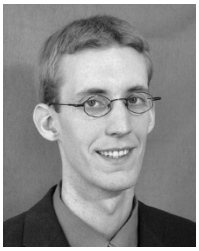
[16] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[17] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, Calif, USA, 1994.

[18] S. Gustafsson, R. Martin, and P. Vary, "On the optimization of speech enhancement systems using instrumental measures," in *Proc. Workshop on Quality Assessment in Speech, Audio, and Image Communication*, pp. 36–40, Darmstadt, Germany, March 1996.

[19] J. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

Thomas Lotter received the Diploma of Engineering degree in electrical engineering from Aachen University of Technology (RWTH), Germany, in 2000. He is now with the Institute of Communication Systems and Data Processing (IND), Aachen University of Technology, where he is currently pursuing the Ph.D. degree. His main research interests are in the areas of speech and audio processing, particularly in speech enhancement with single and multimicrophone techniques.



Christian Benien received the Diploma of Engineering degree in electrical engineering from Aachen University of Technology (RWTH), Germany, in 2002. He is now with Philips Research in Aachen. His main research interests are in the areas of speech enhancement, speech recognition, and the development of interactive dialogue systems.

Peter Vary received the Diploma of Engineering degree in electrical engineering in 1972 from the University of Darmstadt, Darmstadt, Germany. In 1978, he received the Ph.D. degree from the University of Erlangen-Nuremberg, Germany. In 1980, he joined Philips Communication Industries (PKI), Nuremberg, where he became Head of the Digital Signal Processing Group. Since 1988, he has been Professor at Aachen University of Technology, Aachen, Germany, and Head of the Institute of Communication Systems and Data Processing. His main research interests are speech coding, channel coding, error concealment, adaptive filtering for acoustic echo cancellation and noise reduction, and concepts of mobile radio transmission.