MULTICHANNEL SPEECH ENHANCEMENT USING BAYESIAN SPECTRAL AMPLITUDE ESTIMATION

Thomas Lotter, Christian Benien, and Peter Vary

Institute of Communication Systems and Data Processing (ind)
Aachen University (RWTH), Templergraben 55, D-52056 Aachen, Germany
E-mail: {lotter | vary}@ind.rwth-aachen.de

ABSTRACT

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 is highly correlated and the noise is sufficiently uncorrelated. Whereas the first MMSE estimator also requires the desired signals to be in phase, 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 and the MAP estimator derived by Wolfe and Godsill respectively.

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. Single microphone speech enhancement algorithms, e.g. the Minimum Mean Square Error (MMSE) estimator of the speech Discrete Fourier Transform (DFT) amplitudes [1], can achieve high noise reduction at the expense of moderate speech distortion. With multiple microphones spatial information can be exploited, e.g. by beamforming, to reduce noise and reverberation causing only very little speech distortion. However, if the Direction Of Arrival (DOA) can not be estimated with sufficient accuracy, the performance of the beamforming system degrades.
In this contribution we propose two estimators for speech DFT amplitudes that exploit the benefits of multiple microphones. Whereas the first estimator requires the desired signal components to be in phase, the second estimators delivers DOA independent noise reduction.

Figure 1 shows an overview of the multichannel noise reduction system with the proposed speech estimators. The time signals $y_i(k)$, $i \in \{1 \dots M\}$ from M microphones are segmented and multiplied by half overlapping Hann windows. The resulting blocks are transformed via FFT. $Y_i(\lambda)$ denotes the complex value of signal i in DFT bin λ. For the sake of brevity the frequency index λ is omitted.

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

Y_i consists of a speech component $S_i = A_i e^{j\alpha_i}$ and noise N_i. A_i denotes the spectral amplitude of speech and α_i the corresponding phase.
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 (MMSE-STSA) estimator [1], its logarithmic extension [2], or less complex MAP estimators [3] 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. 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 maximum a posteriori (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 real environments is discussed.

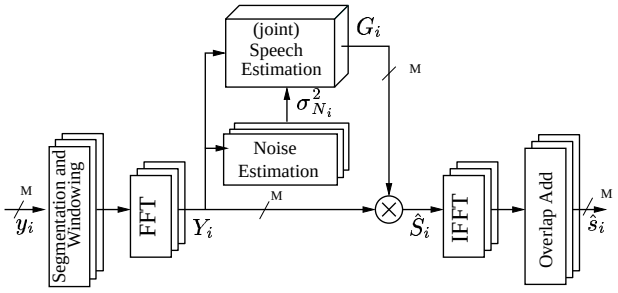

Figure 1: Multichannel Noise Reduction System

2. STATISTICAL MODELS

Motivated by the central limit theorem, real and imaginary part of the DFT coefficients are usually modelled as zero mean Gaussian [4].This leads to the following statistical model for a DFT bin of the i-th signal. (cf. [1],[5]):

$$p(A_i, \alpha_i) = \frac{A_i}{\pi\sigma^2_{S_i}} \exp\left(-\frac{A_i^2}{\sigma^2_{S_i}}\right) \qquad (2)$$

$$p(R_i|A_i) = \frac{2R_i}{\sigma^2_{N_i}} \exp\left\{-\frac{R_i^2 + A_i^2}{\sigma^2_{N_i}}\right\} I_0\left(\frac{2A_iR_i}{\sigma^2_{N_i}}\right) \qquad (3)$$

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

Here $\sigma^2_{S_i}$ describes the variance of the speech in channel i and 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, e.g. inside a room or a car. A desired signal s arrives at a microphone

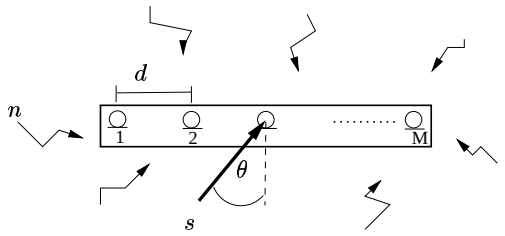

Figure 2: Speech and noise sources arriving at microphone array

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 between two omnidirectional microphones i and j of a diffuse noise field is given by

$$MSC_{ij}(f) = \frac{|\Phi_{ij}(f)|^2}{\Phi_{ii}(f)\Phi_{jj}(f)} = \mathrm{si}^2\left(\frac{2\pi f d_{ij}}{c}\right). \quad (5)$$

Therefore, above a critical frequency depending on the microphone distance, the MSC becomes very low and thus the noise components of the noisy spectra can be considered uncorrelated with

$$E\{N_i N_j^*\} = \begin{cases} \sigma_{N_i}^2 & ; \quad i = j \\ 0 & ; \quad i \neq j \end{cases} \quad (6)$$

Hence (3) and (4) can be extended to

$$p(R_1, \ldots, R_M | A_n) = \prod_{i=1}^{M} p(R_i | A_n) \quad (7)$$

$$p(Y_1, \ldots, Y_M | A_n, \alpha_n) = \prod_{i=1}^{M} p(Y_i | A_n, \alpha_n) \quad (8)$$

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 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 channel dependent factor c_i, i.e. $A_i = c_i \cdot A$ and $\sigma_{S_i}^2 = c_i^2 \sigma_S^2$.
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

$$\xi_i = \frac{\sigma_{S_i}^2}{\sigma_{N_i}^2} \; ; \; \gamma_i = \frac{R_i^2}{\sigma_{N_i}^2}, \quad (9)$$

where the a priori SNRs ξ_i are estimated by the decision directed approach [1].

3. M-D 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 straight forward multichannel extension of the well known MMSE-STSA by Ephraim and Malah [1] is derived. Second, a practically more useful MAP estimator for DOA independent noise reduction is introduced. All estimators output M spectral amplitudes and thus M enhanced signals are delivered by the noise reduction system.

3.1. Estimation conditioned on complex spectra

The single channel algorithm derived by Ephraim and Malah calculates the expectation of the speech spectral amplitude A conditioned on the observed complex Fourier coefficient Y_n, i.e. $E\{A_n | Y_n\}$. 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:

$$\hat{A}_n = E\{A_n | Y_1, \ldots, Y_M\} \quad (10)$$

$$= \int_0^\infty \int_0^{2\pi} A_n \, p(A_n, \alpha_n | Y_1, \ldots, Y_M) \, d\alpha_n \, dA_n. \quad (11)$$

This estimator can be expressed via Bayes' Rule and using (8) as

$$\hat{A}_n = \frac{\int_0^\infty A_n \int_0^{2\pi} p(A_n, \alpha_n) \prod_{i=1}^{M} p(Y_i | A_n, \alpha_n) \, d\alpha_n \, dA_n}{\int_0^\infty \int_0^{2\pi} p(A_n, \alpha_n) \prod_{i=1}^{M} p(Y_i | A_n, \alpha_n) \, d\alpha_n \, dA_n}. \quad (12)$$

To solve (12) we assume perfect DOA correction, i.e. $\alpha_i := \alpha$ for all $i \in \{1 \ldots M\}$. Inserting $A_i = \frac{c_i}{c_n} A_n$ in (8),(4) the integral over α in (12) becomes: ([6] eq. (3.339))

$$\int_0^{2\pi} \exp\left\{ -\sum_{i=1}^{M} \frac{|Y_i - \frac{c_i}{c_n} A_n e^{\alpha}|^2}{\sigma_{N_i}^2} \right\} d\alpha =$$

$$\exp\left\{ -\sum_{i=1}^{M} \frac{|Y_i|^2 + (\frac{c_i}{c_n} A_n)^2}{\sigma_{N_i}^2} \right\} I_0\left(2A_n \left| \sum_{i=1}^{M} \frac{\frac{c_i}{c_n} Y_i}{\sigma_{N_i}^2} \right| \right). \quad (13)$$

The remaining integrals over A_n can be solved using ([6] eq. (6.631.1)). After some straightforward calculations, the gain factor for channel n is obtained as

$$G_n = \frac{\hat{A}_n}{|Y_n|} = \Gamma(1.5) \cdot \sqrt{\frac{\xi_n}{\gamma_n \left(1 + \sum_{i=1}^{M} \xi_i\right)}} \cdot F_1\left(-0.5; 1; \frac{\left| \sum_{i=1}^{M} \sqrt{\gamma_i \xi_i} e^{j\vartheta_i} \right|^2}{1 + \sum_{i=1}^{M} \xi_i} \right). \quad (14)$$

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, $i \in \{1 \ldots M\}$. F_1 has only to be evaluated once, since the argument is independent of n. Note, that in case of $M = 1$ (14) is the single channel MMSE estimator derived by Ephraim and Malah.

3.2. Estimation conditioned on spectral amplitudes

The assumption $\alpha_i := \alpha \, , i \in \{1 \ldots M\}$ introduces a DOA dependency, since this is only given for speech from $\theta = 0°$. For a DOA independent speech enhancement we condition the expectation of $\hat{A}_n$ on the joint observation of all noisy amplitudes R_i, i.e. $\hat{A}_n = E\{A_n | R_1, \ldots, R_M\}$.
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 (3) and (7), 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 by [3] 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 R_i , $i \in \{1 \dots M\}$.

$$\hat{A}_n = \arg\max_{A_n} p(A_n|R_1, \dots, R_M) \quad (15)$$

$$= \arg\max_{A_n} \frac{p(R_1, \dots, R_M|A_n)p(A_n)}{p(R_1, \dots, R_M)}. \quad (16)$$

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 (7), (2) and (3) we get

$$\log L = \log\left(\frac{A_n}{\pi\sigma^2_{S_n}}\right) - \frac{A_n^2}{\sigma^2_{S_n}} + \sum_{i=1}^{M}\left[\log\left(\frac{2R_i}{\sigma^2_{N_i}}\right) - \frac{R_i^2 + (\frac{c_i}{c_n})^2 A_n^2}{\sigma^2_{N_i}} + \log\left(I_0\left(2\frac{\frac{c_i}{c_n}A_n R_i}{\sigma^2_{N_i}}\right)\right)\right]. \quad (17)$$

A closed form solution can be found if the modified Bessel function I_0 is considered asymptotically. For large arguments, the Bessel function can be approximated by

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}}\, e^x. \quad (18)$$

Here, the term in the likelihood function containing the Bessel function simplifies to:

$$\log\left(I_0\left(2\frac{\frac{c_i}{c_n}A_n R_i}{\sigma^2_{N_i}}\right)\right) \approx \frac{2\frac{c_i}{c_n}A_n R_i}{\sigma^2_{N_i}} - \frac{1}{2}\log\left(4\pi\frac{\frac{c_i}{c_n}A_n R_i}{\sigma^2_{N_i}}\right). \quad (19)$$

Differentiation of $\log L$ and multiplication with the amplitude A_n results in $A_n \frac{\partial \log L}{\partial A_n} = 0$

$$A_n^2\left(-\frac{1}{\sigma^2_{S_n}} - \sum_{i=1}^{M}\frac{(\frac{c_i}{c_n})^2}{\sigma^2_{N_i}}\right) + A_n\sum_{i=1}^{M}\frac{\frac{c_i}{c_n}R_i}{\sigma^2_{N_i}} + \frac{2-M}{4} = 0. \quad (20)$$

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 (20) is used as approximation, identical to the real part of the complex solution. The resulting gain factor of channel n is given as

$$G_n = \frac{\sqrt{\frac{\xi_n}{\gamma_n}}}{2 + 2\sum_{i=1}^{M}\xi_i} \cdot \mathrm{Re}\left\{\sum_{i=1}^{M}\sqrt{\gamma_i\xi_i} + \sqrt{\left(\sum_{i=1}^{M}\sqrt{\gamma_i\xi_i}\right)^2 + (2-M)\left(1 + \sum_{i=1}^{M}\xi_i\right)}\right\} \quad (21)$$

For the calculation of the gain factors, no exotic function needs to be evaluated any more. Also, $\mathrm{Re}\{\cdot\}$ has only to be calculated once, since the argument is independent of n. Again, if $M = 1$, we have the single channel MAP estimator as given in [3].

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. All estimators were embedded in the DFT based noise reduction system in figure 1, where the noise power spectral density was estimated by means of *Minimum Statistics* [7].
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. The speech quality of the noise-reduced signal was measured by calculating the segmental speech SNR between original and processed speech. On the other hand, the amount of noise reduction was measured by dividing segmental input and output noise power. In all experiments we do not apply additional (commonly used) soft weighting techniques in order to isolate the benefits of the joint speech estimators compared to the single channel estimator. To study the performance in ideal conditions, we first utilize the estimators on $M = 4$ identical speech signals disturbed by uncorrelated white noise. Figure 3 plots noise reduction and speech quality of the noise reduced signal averaged over all four microphones. Both joint estimators provide a significant higher speech quality and noise attenuation than the single channel MMSE estimator. The MAP estimator conditioned on the noisy amplitudes specifically outperforms the MMSE estimator by a higher noise reduction. The MMSE estimator conditioned on the complex spectra delivers a much higher speech quality.

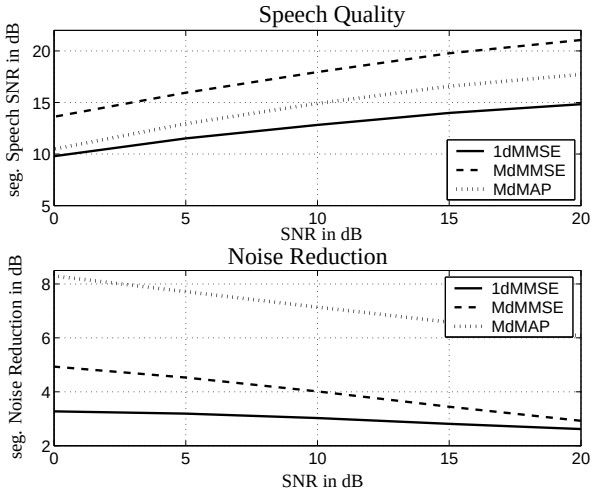

Figure 3: Speech quality and noise reduction of 1D/MD-MMSE and MD-Map for 4 signals containing identical speech and white uncorrelated noise

Instead of white uncorrelated noise, we now mix the speech signal from 0^o with noise recorded with a linear microphone array inside a crowded cafeteria. Figure 4 plots the performance of the estimators using $M = 4$ microphones with an interelement spacing of $d = 12cm$. Compared to figure 3 the gain in terms of speech quality prevails. The amount of additional noise reduction decreases. Figure 5 shows the performance when using noise recordings from inside the crowded cafeteria with half the microphone distance, i.e. $d = 6cm$ interelement spacing. The amount of noise reduction provided by the joint estimators decreases due to

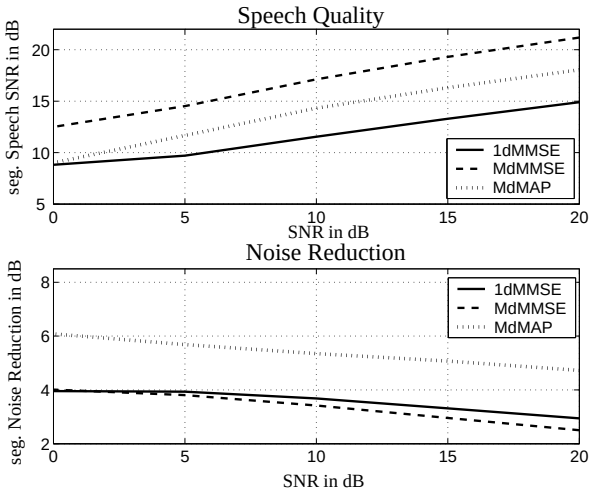

Figure 4: Speech quality and noise reduction of 1D/MD-MMSE and MdMap for 4 signals containing speech from 0^o and cafeteria noise (microphone distance: $d = 12cm$)

the increased correlation at low frequencies according to (5). However there is still a significant improvement left. The Md-MAP estimator still outperforms the single channel MMSE estimator in terms of both speech quality and noise attenuation.

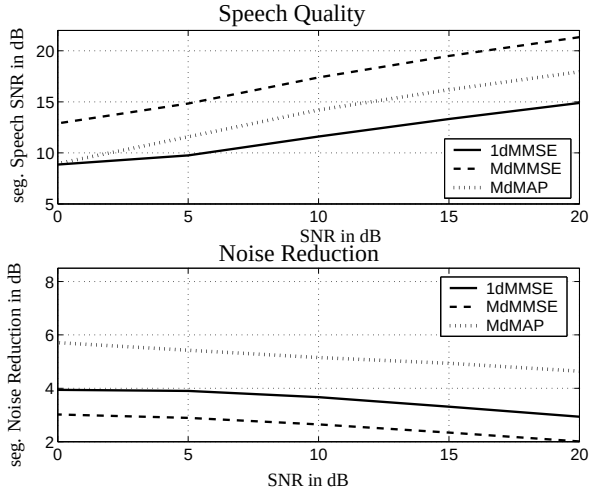

Figure 5: Speech quality and noise reduction of 1D/MD-MMSE and MdMap for 4 signals containing speech from 0^o and cafeteria noise (microphone distance: $d = 6cm$)

Finally we examine the important DOA dependency of the estimators. Figure 6 depicts the performance of the estimators, when the desired signal arrives from 60^o, i.e. the desired signals are not in phase any more.
It can be seen from comparison with figure 4, that the speech quality of the joint MMSE estimator decreases significantly. However, the change of speech DOA has no influence on the performance of the MAP estimator conditioned on the noisy amplitudes.

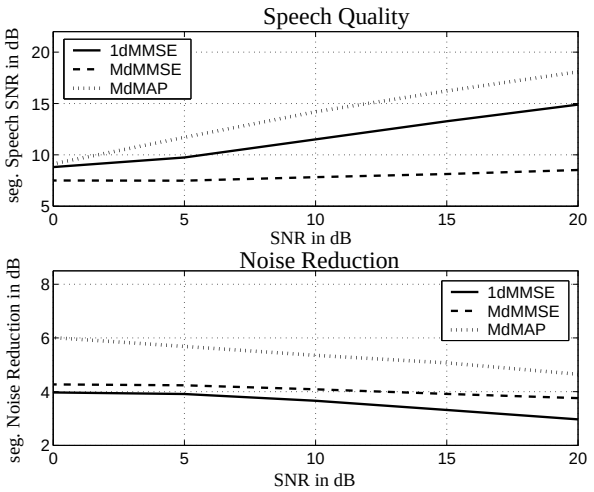

Figure 6: Speech quality and noise reduction of 1D/MD-MMSE and MdMap for 4 signals containing speech from 60^o and cafeteria noise (microphone distance $d = 12cm$)

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 [1] and [3] to the multichannel case. Both estimators provide a significant gain compared to the well known Ephraim and Malah estimator when the speech components are in phase. Moreover, the MAP estimator conditioned on the noisy spectral amplitudes performs a DOA independent speech enhancement. The multichannel noise reduction system using these estimators outputs multiple enhanced signals which can combined by a beamformer for additional speech enhancement.

REFERENCES

[1] 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, pp. 1109–1121, Dez. 1984.

[2] 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, April 1985.

[3] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim and Malah Suppression Rule for Speech Enhancement," *Proceedings of the 11th IEEE Workshop on Statistical Signal Processing*, pp. 496–499, August 2001.

[4] D. Brillinger, *Time Series, Data Analysis and Theory*. McGraw-Hill, 1981.

[5] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, pp. 137–145, April 1980.

[6] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, Inc., 1994.

[7] R.Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, July 2001.