GENERALIZED STOCHASTIC PRINCIPLE FOR MICROPHONE ARRAY SPEECH ENHANCEMENT AND APPLICATIONS TO CAR ENVIRONMENTS

Radu Balan, Justinian Rosca

Siemens Corporate Research,
755 College Road East, Princeton, NJ 08540
{rvbalan,rosca}@scr.siemens.com

Christophe Beaugeant, Virginie Gilg, Tim Fingscheidt

Siemens AG, ICM MP
Haidenauplatz 1, 81667 München Germany
{christophe.beaugeant, virginie.gilg,tim.fingscheidt}@siemens.com

ABSTRACT

In this paper we present novel solutions for microphone array speech enhancement systems that intelligently use the multipath environment to enhance signal coming from a desired location. We obtain a statistical principle that explains previously known factorization results of optimal beamformers, and proves a similar factorization holds for other new optimal estimators. Our solution requires a low computational load, and can be deployed on most of the platforms. We present speech recognition rates on real data, and compare a stereo versus a mono solution on this database.

1. INTRODUCTION

Recent research in microphone-array systems indicate the promise of such techniques in speech enhancement and hands-free communication applications. Of particular interest are techniques using small arrays of microphones, e.g. two-four, in designs of several centimeters in diameter whose goal is to offer a few dB improvement when compared with mono techniques [1] in the case of real-world environments.

Beamforming techniques and in general approaches grounded in the array processing literature indicate tiny SNR improvements in the case of a small number of microphones. Rather than directly considering such approaches, we explore extensions of successful classical mono noise reduction technique to multiple channels. In particular, in this paper we discuss a multi-channel factorization result, which turns out to generalize the well-known single channel statistical estimators of Ephraim-Malah [2, 3], as well as the psychoacoustically motivated speech signal estimator of Gustafsson [4], and sheds more light on the optimal beamformer construction as described in e.g. [5].

Next section formulates the generalized estimation problem for $D > 1$ sensors and solves it under fairly general stochastic hypotheses. Section 3 presents implementation details where some of the formulae are modified to produce less artifacts based on subjective testings. In Section 4 we present numerical results in terms of Speech Recognition Rate for a car environment. We compare the SR rates obtained with the 2-microphone solution with results from mono solution.

2. THE SPEECH ENHANCEMENT ALGORITHM

2.1 Mixing Model and Signal Assumptions

The mixing model we consider is as follows. We assume D microphone signals $x_1(t), \ldots, x_D(t)$ record a source $s(t)$ and noise $n_1(t), \ldots, n_D(t)$ signals,

$$x_l(t) = \sum_{k=0}^{L_l} a_k^l s(t - \tau_k^l) + n_l(t) \tag{1}$$

where (a_k^l, τ_k^l) are the attenuation and delay on the k^{th} path to microphone l. In frequency domain, the convolutions become multiplications. Furthermore, since we are not interested in balancing the channels, we redefine the source so that the first channel becomes unity:

$$\begin{aligned} X_1(k,\omega) &= S(k,\omega) + N_1(k,\omega) \\ X_2(k,\omega) &= M_2(\omega)S(k,\omega) + N_2(k,\omega) \\ &\cdots \\ X_D(k,\omega) &= M_D(\omega)S(k,\omega) + N_D(k,\omega) \end{aligned} \tag{2}$$

where k is the frame index, and ω the frequency index. More compactly, this model can be rewritten as

$$X = MS + N \tag{3}$$

where X, M, S, N are D-complex vectors. Regarding this model, we make the following assumptions:

1. $S(\omega)$ are zero-mean stochastic processes with spectral power $\rho_s(\omega) = \mathbf{E}[|S|^2]$;
2. $(N_1, N_2, \ldots, N_D)$ is a zero-mean stochastic signal with spectral covariance matrix R_n.
3. s is independent of n.

The problem is to estimate s (or S) given the measurements $X = (X_1, \ldots, X_D)^T$.

2.2 The Factorization and the Optimal Filter

Since there are many ways one can estimate the signal S, we consider here several approaches and show they all admit a similar solution.

2.2.1 Statistic Estimators

We are interested in a principled statistical way of estimating S. In [5], Chapter 6, the author briliantly summarizes several optimal beamformers and show them all admit a factorization through so-called *Capone beamformer*. For some of them the noise needs to be assumed Gaussian, for others this is not a necessary hypothesis. More specifically he addresses the following estimators: the BLUE (Best Linear Unbiased Estimator) that minimizes the output noise variance, the MVUE (Minimum Variance Unbiased Estimator)

also known as MVDR, the ML estimator, the MMSE estimator, the Max-SNR estimator, and the MPDR (minimum Power Distortionless Response) filter. Independently of [5], in [6], Simmer and collaborators obtained the factorization of the MMSE estimator of S. In both works, the factorization of the solution comes à posteriori, once the actual solution has been found. The authors merely verify the overall filter (which is linear in all these cases) factors as mentioned before. Our approach and the purpose of this section is different. We point out a statistical principle that explains à priori why the (possibly nonlinear) estimator has to factor into the MVDR beamformer followed by a scalar estimator. To obtain this we need to make the stochastic assumption mentioned above:

4. $N = (N_1, \ldots, N_D)^T$ is a Gaussian random variable.

With this assumption in place we obtain, as in [7], that the linear functional

$$T(X) = \frac{M^* R_n^{-1} X}{M^* R_n^{-1} M} \tag{4}$$

is a sufficient statistics both in classical (i.e. Fisher) and Bayes sense for S, and for any function of S. Hence any MMSE or MAP estimator of S, or a function of S (like $|S|$, $log|S|$, $S/|S|$) factors through $T(X)$. More specifically, the MMSE or MAP estimator of $\varphi(S)$, where φ is a scalar function of S (identity, or modulus, or logarithm of modulus, or the complex phase) is given by

$$\hat{\varphi(S)}_{MMSE} = \mathbf{E}[\varphi(S) \mid X] \tag{5}$$

$$\hat{\varphi(S)}_{MAP} = argmax_S \; p(\varphi(S)|X) \tag{6}$$

and factors as:

$$\hat{\varphi(S)}_{MMSE} = \mathbf{E}[\varphi(S) \mid T(X)] \tag{7}$$

$$\hat{\varphi(S)}_{MAP} = argmax_S \; p(\varphi(S)|T(X)) \tag{8}$$

Hence the optimal (in MMSE or MAP sense) estimator of $\varphi(S)$ is obtained by solving a *single-channel* optimization problem, and thus reduces to previously known estimators. In particular the multichannel MMSE estimator of short-time spectral amplitude (i.e. $|S|$), known in the mono case as the Ephraim-Malah filter [2], as well as the MMSE estimator of log-STSA (see [3] for the mono case), or several multichannel MAP estimators as in [8] can all be easily derived. They all factor as the linear beamformer (4) followed by a scalar signal estimator from:

$$Z = T(X) = S + N_{eff} \quad , \quad N_{eff} = T(N) \tag{9}$$

To obtain Z an estimate of the noise spectral covariance matrix R_n is required. However this is a cumbersome task involving the use of a Voice Activity Detector, and hence prone to errors. Note the Minimum Statistics approach as proposed by R.Martin in [9] cannot be used here, since we deal with nonpositive estimates (the crosscovariance terms). Instead we propose a different way of computing Z that does not use R_n. This is based on the Matrix Inversion Lemma and has also been noticed in e.g. [5]. Since $R_x = R_n + R_s M M^*$ a direct computation shows

$$Z = \frac{1}{M^* R_x^{-1} M} M^* R_x^{-1} X \tag{10}$$

which is implemented by computing R_x instead of R_n. This is a much easier task, since R_x corresponds to the measured signal spectral covariance matrix, easily available.

2.2.2 Psychoacoustically Motivated Estimators

.

In [10] we looked for a linear filter $A = [A_1, \ldots, A_D]$ applied to X that minimizes the variance

$$R_e = \mathbf{E}[|AX - (S + \zeta_1 N_1 + \ldots + \zeta_D N_D)|^2] \tag{11}$$

subject to $(A - \zeta) R_n (A^* - \zeta^T) = R_T$ and $|AM| \leq 1$, where $\zeta = [\zeta_1, \cdots, \zeta_D]$ is the $1 \times D$ vector of desired levels of noise in the estimate, and R_T is a psychoacoustically motivated threshold so that any noise with spectral power below R_T becomes unnoticeable.

The solution in [10] for this constrained optimization problem turns into:

$$A_o = \zeta + \frac{1 - \zeta M}{|1 - \zeta M|} \sqrt{\frac{R_T}{M^* R_n^{-1} M}} M^* R_n^{-1} \tag{12}$$

when the right-hand side satisfies $|A_o M| \leq 1$, and

$$A_o = [1, 0, \ldots, 0] \tag{13}$$

otherwise. Furthermore, if we define

$$Z = \frac{1}{M^* R_n^{-1} M} M^* R_n^{-1} X \tag{14}$$

then $Z = S + N_{eff}$ where the effective noise spectral power is $R_N^{eff} = \frac{1}{K^* R_n^{-1} K}$, and the previously known Gustafsson's psychoacoustically motivated mono filter ([4]) applied on Z becomes:

$$H_G = \zeta + \sqrt{\frac{R_T}{R_N^{eff}}} \tag{15}$$

which yields virtually the same output as A_o applied on X, when we properly define ζ above. As proved before, one would implement (10) rather than (14) since it is more robust to errors.

The conclusion of these two approaches is that, from an algorithmic point of view, the estimation problem decouples (or, factors) into two components: first a *generalized beamformer* given by (14), and then a mono optimization problem of the signal (or a function of the signal) based on Z only.

3. IMPLEMENTATION

3.1 Implementation Details

1. Beamformer. The theory developed so far showed that for a large class of signal estimators, the implementation factors into two steps: a generalized beamformer that linearly filter the multidimensional input into a mono signal, followed by a single-channel signal enhancement block that is optimized for a mono mixing setup

$$Z = S + N_{eff}$$

However, listening tests showed that for high input SNR, the preprocessing (10) introduces some distortions due to poor conditioning of R_x. Instead we use a different filter, namely

$$Z_{highSNR} = \frac{M^* X}{\|M\|^2} \tag{16}$$

which is a plain beamformer adapted to the mixing environment though. For a range of estimated input SNR, we linearly interpolate between these two filters:

$$Z_{actual} = (1 - a(SNR))Z + a(SNR)Z_{highSNR} \quad (17)$$

where $0 \leq a(SNR) \leq 1$, and

$$a(SNR) = \begin{cases} 0 & for \quad SNR < SNR_1 \\ \frac{SNR - SNR_1}{SNR_2 - SNR_1} & for \quad SNR_1 \leq SNR \leq SNR_2 \\ 1 & for \quad SNR > SNR_2 \end{cases} \quad (18)$$

The SNR is estimated in the mono block and the two thresholds SNR_1, SNR_2 are estimated experimentally.

2. Estimation of M. The mixing vector (or the *stirring vector* as used in beamforming literature, see [11]) is estimated here by calibration. This means that is low-noise condition, a voice signal is measured by the device. This is done at the begining of use, and fixed thereafter. Assume the measuring noise is independent on the D channels, and has estimated SNR_i on channel i, the spectral autocovariance of channel i, and the spectral cross-covariance between channel i and channel 1 become

$$R_{x;ii} = |M_i|^2 R_s + R_{n;i} \quad (19)$$

$$R_{x;1i} = M_1 \bar{M}_i R_s \quad (20)$$

An "easy", but biased, estimator of M is given by $M_{i;biased} = \frac{R_{x;ii}}{R_{x;1i}}$. Solving for M_i in (19,20) we obtain:

$$M_i = \frac{M_{i;biased}}{1 + \frac{SNR_i}{(SNR_1 - 1)^2 |M_{i;biased}|^2}} \quad (21)$$

3. Estimation of R_x

For R_x we used a first order learning rule with learning constant α:

$$R_x^{t+1} = (1 - \alpha) R_x^t + \alpha X X^* \quad (22)$$

Since the inverse of R_x is required, for large number of microphones it makes sense to apply the Matrix Inversion Lemma again and obtain a nice updating rule for R_x^{-1}:

$$(R_x^{t+1})^{-1} = \frac{1}{1 - \alpha}(R_x^t)^{-1} - \frac{\alpha}{(1 - \alpha)^2 + \alpha \|(R_x^t)^{-1} X\|^2} \cdot (R_x^t)^{-1} X X^* (R_x^t)^{-1} \quad (23)$$

whose computational complexity scales quadratically with the number of microphones (not cubically as the inverse may suggest).

4. Emphasis/Deemphasis Filters

To attenuate the high-frequency constribution to filtering and parameter estimation, we use an emphasis filter on the input signals:

$$y(t) = x(t) + ax(t - 1) \quad (24)$$

before computing FFT, but after windowing, and a deemphasis filter:

$$u(t) = -au(t - 1) + v(t) \quad (25)$$

after inverse FFT, but before overlap-add procedure.

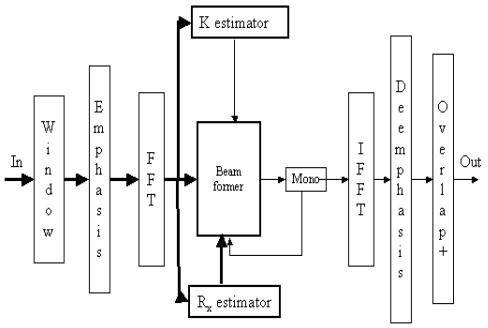

Figure 1: The Overall Architecture.

3.2 The Overall Architecture

The overall architecture is presented in Figure 1. The mono part is simply represented as a block in this architecture. Depending on the global criterion, one can use the Wiener filter, the Ephraim-Malah filter(s), or the Gustafsson filter for this block. The mono block has to feedback into the beamformer an estimate of its input SNR. The results reported in section 4 were obtained using a modified Wiener filter (as described in [12]), namely

$$H_W = \frac{R_Z}{R_Z + min(256 R_N^{eff}, 2R_Z)} \quad (26)$$

The noise spectral power R_N^{eff} was estimated using a minimum statistics filter similar to [9]. At the same time, this yields an estimate of SNR. Note there is no need of a Voice Activity Detector.

4. RESULTS ON REAL DATA

In this section we describe the experimental setup. We recorded data in a car environment. More specifically, in a noise-free garage we recorded in car several voices using 2 microphones. With the same microphones we recorded street noise while driving. Then we mixed together voice and noise signals at different input SNR levels. The calibration was done using only one voice signal per speaker. In total there were 20 speakers, 4 noise files, and 50 voice files per speaker. Therefore we had a database of 4000 files, each containing a 10-digit phone number. The output of our speech enhancement processor was feed into the Speech Recognition Engine described in [12]. For each frame the SRE front-end computes 12 MFCC coefficients plus the total frame energy, as well as their corresponding delta and acceleration values. The inclusion of delta and acceleration coefficients is known to improve robustness of the features against noise. The frames have 32 ms length and an overlap of 17 ms.

The Speech Enhancement processor used the same frame length and overlap, for signals sampled at 8KHz. The thresholds in (18) where $SNR_1 = 0dB$, $SNR_2 = 6dB$. We performed speech recognition tests at four levels of input SNR, namely for $SNR = -6, 0, 6, 12$dB. We also applied the mono filter to one channel, and retained the best results among the two recognition rates. The rate measures the number of files correctly recognized. A file is correctly recognized when all 10 digits are correctly recognized. In

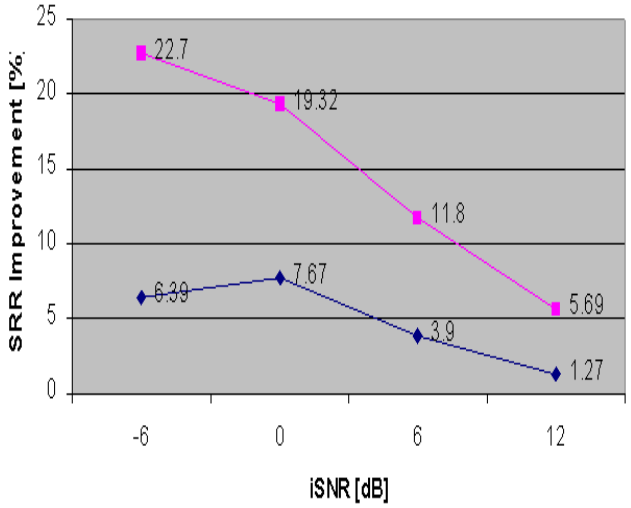

Figure 2: SRR Improvement for the two channel scheme (top curve) and mono scheme (bottom curve) with respect to the input mix.

iSNR[dB]	-6	0	6	12
Mix SRR[%]	8.93	40.23	64.10	77.13
Mono SRR[%]	15.32	47.9	68	78.4
Stereo SRR[%]	31.63	59.55	75.90	82.82
Voice SRR[%]	81.5	81.5	81.5	81.5

Table 1: Speech Recognition Rates.

terms of recognition rates, the results are presented in Table 1. For four input SNRs we present the speech recognition rate (SRR) for the input mix, the output of the mono solution only, the output of the 2-channel system, and the clean voice signal (obviously, independent of input SNR). A clearer visualisation of the 2-channel processing gain versus mono processing gain is rendered in Figure 2. There we plot the SRR improvement of mono and the 2-channel scheme versus the input mix at the four SNRs.

This plot shows the relative improvement of the stereo solution versus mono solution. At low SNR the relative improvement is as high as 100%, i.e. the SR rate doubles. As higher the SNR gets as lower the SRR improvement is obtained. Yet, even at 12dB, there is still an improvement of about 4.5% in absolute recognition rate terms.

5. CONCLUSIONS

In this paper we presented a multi-channel speech enhancement scheme and we validated its performance on real data in a 2-microphone car environment setup where the speech recognition rate was used as criterion. Based on general statistics principle we showed an optimal signal estimator factors into a linear generalized beamformer followed by a (usually) nonlinear mono filter that solves a mono estimation problem. This factorization scheme holds true for several MMSE and MAP estimators, as well as for psychoacoustically motivated multi-channel speech enhancement systems. Numerical experiments were performed where we compared a 2-microphone scheme with the mono solution on real car environment data. The 2-microphone system improved the recognition rate by up to 16% in absolute terms, where a successful recognition was considered when an entire set of 10 digits was correctly recognized. The algorithm involves little computational load and scales nicely with the number of microphones.

REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

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

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

[4] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *ICASSP*, 1998, pp. 397–400.

[5] H.L.van Trees, *Optimum Array Processing*, Wiley, 2002.

[6] K.U. Simmer, J. Bitzer, and C. Marro, *Microphone Arrays*, chapter Post-filtering techniques, pp. 39–60, Springer, 2001.

[7] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Proceedings of SAM 2002, Rosslyn VA*, 2002.

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

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

[10] J. Rosca, R. Balan, and C. Beaugeant, "Multichannel psychoacoustically motivated speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003), Hong-Kong, China*, April 2003.

[11] G.W Elko, *Microphone Arrays*, chapter Spatial Coherence Functions for Differential Microphones in Isotropic Noise Fields, pp. 61–86, Springer, 2001.

[12] S. Aalburg, C. Beaugeant, S. Stan, T. Fingscheidt, R. Balan, and J. Rosca, "Single- and two-channel noise reduction for robust speech recognition in car," in *ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, June 2002.