MICROPHONE ARRAY SPEECH ENHANCEMENT BY BAYESIAN ESTIMATION OF SPECTRAL AMPLITUDE AND PHASE

Radu Balan

Multimedia and Video Technologies Department
Siemens Corporate Research
755 College Road E, Princeton, NJ 08540
radu.balan@scr.siemens.com

Justinian Rosca

Multimedia and Video Technologies Department
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540
justinian.rosca@scr.siemens.com

ABSTRACT

Microphone arrays provide new opportunities for noise reduction and speech enhancement. This paper presents a novel decomposition of the estimation problems for short-time spectral amplitude (STSA), log STSA, and phase in the Bayesian estimation framework. The decomposition is based on the notion of sufficient statistics for the microphone array case. It nicely generalizes the well-known single-channel Ephraim-Malah estimators [4, 5] to the microphone array case. We also compare noise reduction obtained in the single channel with the two- and four-channel cases on real data.

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 in the case of real-world environments. Theoretically, multi-channel techniques offer more information about the acoustic environment, therefore should indeed offer possibility for improvement especially in the case of reverberant environments due to multi-path effects and severe noise conditions. These conditions are known to affect the performance of state-of-the-art single channel techniques. However the effectiveness of multiple channel techniques for just a few microphones is yet to be proven.

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 novel multi-channel approach, which turns out to be a generalization of the well-known Ephraim-Malah signal estimators [4, 5].

Below, section 2 formulates the generalized multi channel signal enhancement problem. Section 3 presents solutions to these multi-channel estimation problems. Section 4 discusses implementation issues and experimental results with the two and four-channel implementation. We compare the segmental SNR results obtained here with results from an implementation of the Epraim-Malah estimators. We conclude with a qualitative analysis of the multiple channel approach.

2. MULTI-CHANNEL ESTIMATION PROBLEMS

The D sensor (microphone) mixing model we consider has the following form:

$$x_i(t) \quad = \quad k_i \circ s(t) + n_i(t) \qquad (1)$$

where $i = 1, ..., D$, $s(t)$ is the source signal, $n_1(t), ..., n_D(t)$ are the microphone noises, $x_1(t)$,. . .,$x_D(t)$ the measured signals, and k_1, . . ., k_D the channel impulse responses, for $1 \leq t \leq T$ and $\circ$ denotes convolution. The problem is to estimate the source $s(t)$ in an optimal sense that would be made more precise below, given the measurements x_1,. . .,x_D and assuming the following:

1. The impulse responses $k_1, \dots, k_D$ are known;
2. The source and noise signals are short-time stationary and gaussian distributed with zero average and known covariance matrix;
3. The source signal is independent of the noises.

These assumptions imply that the short-time Fourier transforms of $s, n_1, \dots, n_D$ decouple into independent variables at different frequencies (see [4]). The frequency-domain representation of(1) is:

$$X_i(\omega) \quad = \quad K_i(\omega)S(\omega) + N_i(\omega) \qquad (2)$$

where $i = 1, ..., D$, $s(t)$ and the capital letters are the short-time Fourier transforms of the lower-case signals. The mixing model can be compactly rewritten:

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

where $X = [X_1 \ \cdots \ X_D]^T$ and $K = [K_1 \ \cdots \ K_D]$ are complex D-vectors.

Now the problem can be stated as follows: determine the minimum mean square error estimator for short-time spectral amplitude (STSA) $|S|$, or the log short-time spectral amplitude (log-STSA) $log(|S|)$, and the short-time complex exponential (STCE), $S/|S|$.

The *Minimum Mean Square Error Short-Time Spectral Amplitude Estimator (MMSE STSA)* is given at each frequency ω by (see [4]):

$$|\hat{S}|_1() = \mathbf{E}[|S| \mid X_1, \ldots, X_D] \tag{4}$$

The *MMSE log-STSA* estimator is then given by:

$$|\hat{S}|_{log} = exp\{\mathbf{E}[\log(|S|) \mid X_1, \ldots, X_D]\} \tag{5}$$

The *MMSE Short-Time Complex Exponential (MMSE STCE)* is given as the constrained MMSE over the unit complex numbers:

$$\hat{z} = \min_{|z|=1} \mathbf{E}[|z - \frac{S}{|S|}|^2] \tag{6}$$

as in [4]. It turns out (see [4]) the solution to (6) is given by

$$\hat{z} = \frac{\mathbf{E}[\frac{S}{|S|}|X_1, \ldots, X_D]}{|\mathbf{E}[\frac{S}{|S|}|X_1, \ldots, X_D]|} \tag{7}$$

The final estimator that solves the problem is:

$$\hat{S} = |\hat{S}| \cdot \hat{z} \tag{8}$$

where $|\hat{S}|$ is one of $|\hat{S}|_1$ or $|\hat{S}|_{log}$.

3. BAYESIAN APPROACH TO THE ESTIMATION PROBLEMS

In this section we present explicit solutions for the estimation problems (4,5,6).

3.1. Problem Decomposition using Sufficient Statistics

The statistical hypotheses we made allow us to factor the conditional pdf:

$$p(X|S; R_s, R_n, K) = \frac{1}{\pi^D \det(R_n)} exp\{-(X-KS)^* R_n^{-1}(X-KS\} \tag{9}$$

into

$$p(X|S; R_s, R_n, K) = g(S, T(X))h(X) \tag{10}$$

where

$$T(X) = \frac{K^* R_n^{-1} X}{K^* R_n^{-1} K} \tag{11}$$

Using the Fisher-Neyman factorization theorem (see [10]) we deduce that $T(X)$ is *sufficient statistics* for S *in the classical sense*, i.e. when the parameters to be estimated are deterministic. At the same time, a simple statistics exercise shows that $T(X)$ is sufficient statistics for any function $\rho(S)$ (such are for instance $|S|$ and $arg(S)$).

On the other hand, $T(X)$ is *sufficient statistics in the Bayes sense* for our stochastic model (see Theorem 2.14 in [10]), that is for any prior probability of S, the posterior probability of S (or $\rho(S)$) conditioned by the observation X is the same as the conditional with respect to $T(X)$:

$$p(\rho(S)|X) = p(\rho(S)|T(X)) \tag{12}$$

Consequently, the conditional expectation of $\rho(S)$ with respect to X becomes simply:

$$\mathbf{E}[\rho(S)|X] = \mathbf{E}[\rho(S)|T(X)] \tag{13}$$

Note that $T(X)$ is a one-dimensional random variable. Equation 13 shows that the multi channel problem $\mathbf{E}[\rho(S)|X]$ can be decomposed into the projection of the data given by $T(X)$ and a one-dimensional estimation problem. The decomposition considerably simplifies the original problem and enables us to rely on previous single-channel algorithms such as Ephraim-Malah [4].

In the following we shall particularize this result for the cases: $\rho(S) = |S|$, $\rho(S) = \log(|S|)$, and $\rho(S) = S/|S|$.

3.2. The MMSE STSA Estimator

Using (13) for $\rho(S) = |S|$, the MMSE STSA estimator (4) becomes:

$$|\hat{S}|_1 = \mathbf{E}[|S| \mid Y = T(X)] \tag{14}$$

Note that

$$Y = S + \frac{K^* R_n^{-1} N}{K^* R_n^{-1} K} = S + N_r \tag{15}$$

is a single-channel signal containing both source and noise signals, and the effective spectral power of each and effective SNR ξ are:

$$R_s^{eff} = R_s \tag{16}$$

$$R_n^{eff} = 1/(K^* R_n^{-1} K) \tag{17}$$

$$\xi = \frac{R_s^{eff}}{R_n^{eff}} = R_s K^* R_n^{-1} K \tag{18}$$

Note also that the mixing in (15) preserves the Gaussianity of the signals, i.e. N_r is Gaussian.

Now we can immediately apply the Ephraim-Malah estimator [4] to our estimation problem, to obtain in a closed form the expectation in (14):

$$\mathbf{E}[|S|\,|\,Y] = \frac{\sqrt{\pi}}{2}\frac{R_n^{eff}\sqrt{v}}{b}exp(-\frac{v}{2})[(1{+}v)I_0(\frac{v}{2}){+}vI_1(\frac{v}{2})] \quad (19)$$

where:

$$b = |Y| \quad (20)$$

$$v = \frac{\xi}{1+\xi}b^2K^*R_n^{-1}K \quad (21)$$

and I_0, I_1 are the modified Bessel function of the first kind and order 0, respectively 1 (see [6], (8.431-4,5)).

3.3. The MMSE log-STSA Estimator

Using $\rho(S) = \log(|S|)$ the MMSE log-STSA estimator (5) becomes:

$$|\hat{S}|_{log} = \exp\{\mathbf{E}[\log(|S|)|Y = T(X)]\} \quad (22)$$

Using the same argument as before, the above conditional expectation can be taken from [5] as:

$$|\hat{S}|_{log} = \frac{\xi}{1+\xi}\exp\{\frac{1}{2}\int_v^\infty \frac{e^{-t}}{t}dt\}b \quad (23)$$

3.4. The MMSE STCE Estimator

The norm constrained phase estimator (7) is obtained by first computing

$$\mathbf{E}[\frac{S}{|S|}|X] = \mathbf{E}[\frac{S}{|S|}|Y = T(X)] \quad (24)$$

But the right hand side has been computed in [4], and it turns out to be proportional, up to a real constant, to Y. Thus, the MMSE STSP estimator becomes:

$$\hat{z} = \frac{Y}{|Y|} \quad (25)$$

3.5. Overall STFT estimator

Putting together (19), (23) and (25), the estimate of S is:

$$\hat{S} = H(v,b,\xi)Y \quad (26)$$

with:

$$H(v,b,\xi) = \frac{\sqrt{\pi}}{2}\frac{R_n^{eff}\sqrt{v}}{b^2}exp(-\frac{v}{2})[(1{+}v)I_0(\frac{v}{2}){+}vI_1(\frac{v}{2})] \quad (27)$$

in the MMSE STSA case, or

$$H(v,b,\xi) = \frac{\xi}{1+\xi}\exp(\frac{1}{2}\int_v^\infty \frac{e^{-t}}{t}dt)b \quad (28)$$

in the MMSE log-STSA case, and ξ, b, v given in (18,20,21).

The overall estimator can be interpreted as a two-step procedure:

1. First a projection, of the measured signal along $K^*R_n^{-1}$, $Y = \frac{K^*R_n^{-1}X}{K^*R_n^{-1}K}$ so that the reduced system becomes (15).

2. Second, an Ephraim-Malah estimator ([4],[5]) of S applied on Y, given by (26).

It is worth mentioning an optimality property of the filter at step 1: it maximizes the output SNR over all linear filters, that is it maximizes $SNR = \frac{R_s|A^*K|^2}{A^*R_nA}$ over A.

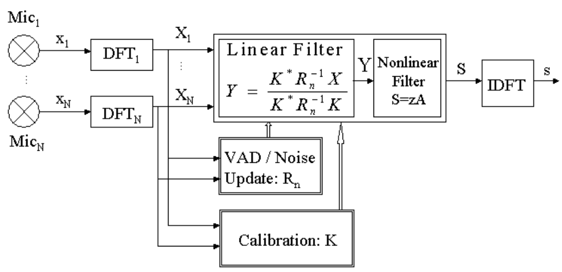

Figure 1: *The D-sensors enhancement scheme.*

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Implementation details, results and conclusions are defered to the full paper. The implementation section highlights the quantities to be estimated: channel transfer functions, noise covariance matrix, effective a priori SNR ξ. We tested the resulting algorithm on two-channel data recorded in a car environment with microphones placed at twenty cm apart and contrasted results to mono results obtained with an implementation of the Ephraim-Malah algorithm.

5. REFERENCES

[4] Ephraim, Y., and Malah, D., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. ASSP 32: 1109-1121, 1984.

[5] Ephraim, Y., and Malah, D., "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. ASSP 33: 443-445, 1985.

[6] Gradshteyn, I.S., and Ryzhik, I.M., Table of Integrals, Series, and Products, Academic Press, 1980.

[10] Schervish, M.J., "Theory of Statistics", Springer-Verlag 1995.