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

Radu Balan, Justinian Rosca

Multimedia and Video Technologies Department
Siemens Corporate Research
755 College Road E, Princeton, NJ 08540
radu.balan@scr.siemens.com , 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 [1] 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]. The approach has been the focus of improvements in recent literature [2].

Next section formulates the generalized signal enhancement problem for the case when the number of microphones is $N > 1$. Section 3 presents solutions to these multi-channel estimation problems. Section 4 discusses implementation issues and experimental results with the two-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

Let us consider a system of N sensors (microphones) as in Figure 1. The mixing model has the following form:

$$\begin{aligned} x_1(t) &= k_1 \circ s(t) + n_1(t) \\ &\cdots \\ x_N(t) &= k_N \circ s(t) + n_N(t) \end{aligned} \tag{1}$$

where $s(t)$ is the source signal (voice), $n_1(t)$,...,$n_N(t)$ are the microphone noises, $x_1(t)$,...,$x_N(t)$ the measured signals, and k_1, ..., k_N 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_N and assuming the following:

1. The source and noise signals are short-time stationary and gaussian distributed with zero average;
2. The source signal is independent of the noises.

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

$$\begin{aligned} X_1(\omega) &= K_1(\omega)S(\omega) + N_1(\omega) \\ &\cdots \\ X_N(\omega) &= K_N(\omega)S(\omega) + N_N(\omega) \end{aligned} \tag{2}$$

where the capital letters are the short-time Fourier transforms of the lower-case signals. Note the mixing model can be compactly rewritten:

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

where $X = [X_1 \ \cdots \ X_N]^T$ and $K = [K_1 \ \cdots \ K_N]$ are complex N-vectors. In frequency domain, the hypotheses made before turn into:

1. The source signal $S(\omega)$ is gaussian distributed with zero mean and spectral power R_s;
2. The noise signals $(N_1(\omega), \ldots, N_N(\omega))$ are gaussian distributed with zero mean and spectral covariance matrix R_n;
3. The source signal is independent of the noise signals, and each of them is independent of the other at different frequencies.

Now the problem we solve can be stated as follows: determine the minimum mean square error estimator for short-time spectral amplitude (STSA) $|S|$, log short-time spectral amplitude (log-STSA) $log(|S|)$, and 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 [9]):

$$|\hat{S}|_1(\omega) = \mathbf{E}[|S| \mid x_1(\cdot), \ldots, x_N(\cdot)] \qquad (4)$$

Since $X_1(\cdot), \ldots, X_N(\cdot)$ is an equivalent representation for the measurements, and furthermore $X_i(\omega_1)$ is independent from $X_j(\omega_2)$ for $\omega_1 \neq \omega_2$, the MMSE estimator turns into:

$$|\hat{S}|_1 = \mathbf{E}[|S| \mid X_1, \ldots, X_N] \qquad (5)$$

where we dropped the argument ω for convenience of notation.

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

$$|\hat{S}|_{log} = exp\{\mathbf{E}[\log(|S|) \mid X_1, \ldots, X_N]\} \qquad (6)$$

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

$$\hat{z} = \min_{|z|=1} \mathbf{E}[|z - \frac{S}{|S|}|^2] \qquad (7)$$

as in [4]. It turns out (see [4]) the solution to (7) is given by

$$\hat{z} = \frac{\mathbf{E}[\frac{S}{|S|}|X_1, \ldots, X_N]}{|\mathbf{E}[\frac{S}{|S|}|X_1, \ldots, X_N]|} \qquad (8)$$

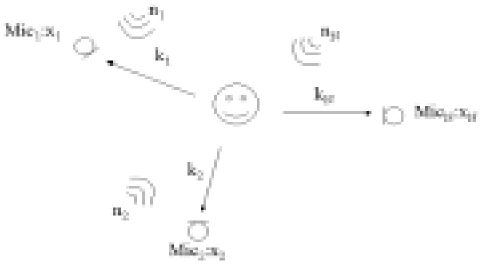

Figure 1: *The N-sensors mixing scheme.*

The final estimator that solves the problem is:

$$\hat{S} = |\hat{S}| \cdot \hat{z} \qquad (9)$$

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 (5,6,7).

3.1. Statistical Analysis

The statistical hypotheses we made allow us to write the following conditional pdf of the data:

$$p(X|S; R_s, R_n, K) = \frac{1}{\pi^N \det(R_n)} exp\{-(X - KS)^* R_n^{-1}(X - KS\} \qquad (10)$$

This expression can be factorized as:

$$p(X|S; R_s, R_n, K) = g(S, T(X))h(X) \qquad (11)$$

where g and h are some functions and

$$T(X) = \frac{K^* R_n^{-1} X}{K^* R_n^{-1} K} \qquad (12)$$

Using the well-known Fisher-Neyman Factorization Theorem (see Proposition IV.C.1 in [9]) we deduce that $T(X)$ is *sufficient statistics (in the classical sense)* for S. 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 [11]), 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 $T(X)$:

$$p(\rho(S)|X) = p(\rho(S)|T(X)) \qquad (13)$$

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

$$\mathbf{E}[\rho(S)|X] = \mathbf{E}[\rho(S)|T(X)] \qquad (14)$$

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 (14) for $\rho(S) = |S|$, the MMSE STSA estimator (5) becomes:

$$|\hat{S}|_1 = \mathbf{E}[|S| \mid Y = T(X)] \qquad (15)$$

Note that

$$Y = S + \frac{K^* R_n^{-1} N}{K^* R_n^{-1} K} = S + N_r \qquad (16)$$

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

$$\begin{aligned} R_s^{eff} &= R_s && (17)\\ R_n^{eff} &= 1/(K^* R_n^{-1} K) && (18)\\ \xi &= \frac{R_s^{eff}}{R_n^{eff}} = R_s K^* R_n^{-1} K && (19) \end{aligned}$$

Note also that the mixing in (16) 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 of (15):

$$\mathbf{E}[|S| \mid Y] = \frac{\sqrt{\pi}}{2} \frac{R_n^{eff}\sqrt{v}}{b} exp(-\frac{v}{2})[(1+v)I_0(\frac{v}{2}) + vI_1(\frac{v}{2})] \qquad (20)$$

where:

$$b = |Y| \quad (21)$$

$$v = \frac{\xi}{1+\xi} b^2 K^* R_n^{-1} K \quad (22)$$

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 (6) becomes:

$$|\hat{S}|_{log} = \exp\{\mathbf{E}[\log(|S|)|Y = T(X)]\} \quad (23)$$

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

$$|\hat{S}|_{log} = \frac{\xi}{1+\xi} \exp\{\frac{1}{2}\int_v^\infty \frac{e^{-t}}{t} dt\} b \quad (24)$$

3.4. The MMSE STCE Estimator

The norm constrained phase estimator (8) is obtained by first computing

$$\mathbf{E}[\frac{S}{|S|}|X] = \mathbf{E}[\frac{S}{|S|}|Y = T(X)] \quad (25)$$

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:

$$\hat{z} = \frac{Y}{|Y|} \quad (26)$$

3.5. Overall STFT estimator

Putting together (20), (24) and (26), the estimate of S is:

$$\hat{S} = H(v, b, \xi) Y \quad (27)$$

with:

$$H(v, b, \xi) = \frac{\sqrt{\pi}}{2} \frac{R_n^{eff}\sqrt{v}}{b^2} exp(-\frac{v}{2})[(1+v)I_0(\frac{v}{2}) + vI_1(\frac{v}{2})] \quad (28)$$

in the MMSE STSA case, or

$$H(v, b, \xi) = \frac{\xi}{1+\xi} \exp(\frac{1}{2}\int_v^\infty \frac{e^{-t}}{t} dt) b \quad (29)$$

in the MMSE log-STSA case, and ξ, b, v given in (19,21,22).

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

1. First a projection, of the measured signal along $K^* R_n^{-1}$:

$$Y = \frac{K^* R_n^{-1} X}{K^* R_n^{-1} K} \quad (30)$$

so that the reduced system becomes (16) with effective spectral powers R_s^{eff}, R_n^{eff} given by (17,18), effective à priori SNR (in terminology of [8]) ξ given by (19), and effective à posteriori SNR (the same terminology) given by:

$$\gamma = \frac{|Y|^2}{R_n^{eff}} = b^2 K^* R_n^{-1} K \quad (31)$$

2. Second, a Ephraim-Malah estimator ([4],[5]) of S is applied on Y given by (27).

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 $SNR = \frac{R_s |A^* K|^2}{A^* R_n A}$ over A.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

4.1. Parameters Estimation

An inspection of (19,21,22) shows that the required quantities to estimate S are: the channel transfer functions $K_1, \ldots, K_N$, the noise spectral covariance matrix R_n, and the *effective à priori SNR* ξ.

As long as we do not want a channel equalization (i.e. dereverberation), we can replace the source s by the signal received by sensor 1 (for instance) in the absence of any noise. Then the constants K_j's are the relative transfer functions (i.e. ratios of the actual transfer functions). We present two adaptive estimators of K's, and the estimation of R_n and ξ below.

4.1.1. An Adaptive Non-Parametric Estimator of K

Assume the measurements $x_1, \ldots, x_D$ contain signal and noise via (2). Assume we have an estimate of the noise spectral power R_n, an estimate of signal spectral power R_s (obtained through spectral subtraction from $R_{x;1,1}$ and $R_{n;1,1}$), and an estimate K' that we want to update. The measured signal (short-time) spectral power $R_x(k, \omega)$ is now

$$R_x(k, \omega) = R_s(k, \omega) K K^* + R_n(k, \omega) \quad (32)$$

We want to update K to $K' = K + \Delta K$ constrained by $\| \Delta K \|$ small, and $\Delta K = [0 \ \Lambda]^T$, where $\Lambda = [\Delta K_2 \ldots \Delta K_D]$, which best fits (32) in some norm. We choose the Frobenius norm, $\| A \|_F^2 = trace\{AA^*\}$. Then the criterion to minimized becomes:

$$J(X) = trace\{(R_x - R_n - R_s(K + [0 \ \Lambda]^T)(K + [0 \ \Lambda]^T)^*)^2\} \quad (33)$$

The gradient at $\Lambda = 0$ is:

$$\frac{\partial J}{\partial \Lambda}|_0 = -2R_s(K^* E)_r \quad (34)$$

where the index r truncates the vector by cutting out the first component: for $v = [v_1 \ v_2 \ \ldots \ v_D]$, $v_r = [v_2 \ \cdots \ v_D]$, and $E = R_x - R_n - R_s K K^*$. Thus the gradient algorithm for K gives the following adaptation rule:

$$K' = K + [\ 0 \quad \Lambda\ ]^T \ , \ \Lambda = \alpha R_s (K^* E)_r \quad (35)$$

where $0 < \alpha < 1$ is the learning rate.

4.1.2. An Adaptive Model-based Estimator of K

A second adaptive estimator of K makes use of a particular mixing model, thus reducing the number of parameters. The simplest but fairly efficient model is the *direct path* model

$$K_l(\omega) = a_l e^{i\omega\delta_l} \ , \ l \geq 2 \quad (36)$$

In this case, a similar criterion to (33) is to be minimized. The actual criterion is:

$$I(a_2,\dots,a_D,\delta_2,\dots,\delta_D) = \sum_\omega trace\{(R_x - R_n - R_s K K^*)^2\} \tag{37}$$

Note the summation across the frequencies because the same parameters $(a_l, \delta_l)_{2 \le l \le D}$ have to explain all the frequencies. The gradient of I evaluated on the current estimate $(a_l, \delta_l)_{2 \le l \le D}$ is

$$\frac{\partial I}{\partial a_l} = -4 \sum_\omega \rho_s \cdot \mathrm{real}(K^* E v_l) \tag{38}$$

$$\frac{\partial I}{\partial \delta_l} = -2 a_l \sum_\omega \omega R_s \cdot \mathrm{imag}(K^* E v_l) \tag{39}$$

where $E = R_x - R_n - R_s K K^*$ and v_l the D-vector of zeros everywhere except on the l^{th} entry where it is $e^{i\omega\delta_l}$, $v_l = [0 \cdots 0 \;\; e^{i\omega\delta_l} \;\; 0 \cdots 0]^T$. Then the updating rule is given by:

$$a_l' = a_l - \alpha \frac{\partial I}{\partial a_l} \tag{40}$$

$$\delta_l' = \delta_l - \alpha \frac{\partial I}{\partial \delta_l} \tag{41}$$

with $0 \le \alpha \le 1$ the learning rate.

4.1.3. Noise Spectral Covariance

The noise spectral covariance matrix is estimated using nonspeech periods. Thus, given $R_n^{(t-1)}$ the current estimate of the noise spectral power, and knowing that frame t, $X^{(t)}$, contains only noise, the updated noise estimate is:

$$R_n^{(t)} = (1-\beta) R_n^{(t-1)} + \beta X^{(t)} X^{(t)*} \tag{42}$$

where the learning rate β is small. Typically $\beta = 0.2$. The voice detection has been performed by the Multichannel VAD described in [10].

4.1.4. Effective A Priori SNR

The effective à priori SNR ξ is estimated similar to the decision-directed method suggested in [4]: Given the previous estimate of the signal amplitude $|S|^{(t-1)}$, effective noise spectral power $(K^* R_n^{(t-1),-1} K)^{-1}$, and current effective amplitude $b^{(t)}$ and noise spectral power $(K^* R_n^{(t),-1} K)^{-1}$, the adaptation rule becomes:

$$\begin{aligned} \xi^{(t)} &= \alpha(|S|^{(t-1)})^2 K^* R_n^{(t-1),-1} K \\ &\quad + (1-\alpha) P[b^{(t),2} K^* R_n^{(t),-1} K - 1] \end{aligned} \tag{43}$$

where $P[x]$ is the positive part of x: $P[x] = x$, for $x \ge 0$, and $P[x] = 0$, for $x < 0$.

4.2. Experimental Results

Figure 2 describes our implementation of this estimation scheme. We used a four-microphone system. Microphones were placed at about twentynine centimeters apart on a tetrahedral frame. Recordings were made in a living room at a sampling frequency of 16kHz. The time-frequency analysis was performed by means of a Hamming window of size 512 samples and 50% overlapp. The reconstruction used the dual frame window as described in [12]. We used the 4-channel VAD described in [10] for learning the noise spectral covariance matrix R_n together. Furthermore, we assumed that the noise cross-spectral power is zero (i.e. $R_{nkj} = 0$, for $k \ne j$). The ratios K were estimated using the first K estimator presented before.

	1 Channel	*2 Channel*	*4 Channel*
MMSE AdaptK	0.48	0.78	1.23
Wiener	1.49	-0.25	-0.82

Table 1: Segmental SNRs for the STFT MMSE estimator (28) and Wiener filter for one, two and four channels.

	1 Channel	*2 Channel*	*4 Channel*
MMSE AdaptK	-2.29	-3.36	-4.62
Wiener	-8.24	11.62	40.27

Table 2: Distortions for the STFT MMSE estimator (28) and Wiener filter for one, two and four channels.

Next we present a comparison of our multi-channel STSA MMSE estimator (four and two channels) with the single channel STSA MMSE estimator. We also compare them to the two and four channel Wiener filters. The evaluation included the following criteria:

1. Segmental SNR, computed using the following formula:

$$segSNR = \frac{1}{N_f} \sum_{k=1}^{N_f} 10 \log_{10} \frac{\| S \|^2}{\| \hat{S} - S_1 \|^2} \tag{44}$$

   where N_f is the number of frames for which the instantaneous signal-to-distortion is between -10dB and +30dB (see [3]), and S_1 is the input signal on channel 1.

2. Signal distortion, computed as follows:

$$dist = 10 \log_{10} \frac{\| \hat{S}_S - S_1 \|^2}{\| S_1 \|^2} \tag{45}$$

3. Average SNR gain, given by:

$$again = \frac{1}{Nf} \sum_{k=1}^{Nf} \left(10 \log_{10} \frac{\| \hat{S}_S \|^2}{\| \hat{S}_N \|^2} - 10 \log_{10} \frac{\| \hat{S}_1 \|^2}{\| \hat{N}_1 \|^2}\right) \tag{46}$$

   where N_f is as before, and

$$\hat{S}_S = H(v,b,\xi) K^* R_n^{-1} S / (K^* R_n^{-1} K),$$

$$\hat{S}_N = H(v,b,\xi) K^* R_n^{-1} N / (K^* R_n^{-1} K)$$

   are the signal, respectively the noise component of the output, and S_1, N_1 are the input voice, respectively noise, signal on channel 1.

Ideally, $segSNR = \infty$, $dist = -\infty$, and $again = \infty$.

Algorithms were tested on four living-room voice and noise recordings with an input SNR of about -0.5dB. The experimental results are summarized in Tables 1,2 and 3.

Results for the four channel STFT MMSE show an absolute improvement of about 1.5dB segmental SNR at a level of about -4dB total distortion and a modest improvement over the Ephraim-Malah solution. The single-channel Wiener filter has a low distortion, but comparatively poor segmental and average SNR gains.

	1 Channel	*2 Channel*	*4 Channel*
MMSE AdaptK	2.4	2.85	3.23
Wiener	0.74	-0.89	-1.17

Table 3: Average SNR gains for the STFT MMSE estimator (28) and Wiener filter for one, two and four channels.

An informal listening of the results showed an audible improvement, from one to two and two to four channels. In all cases the signal estimate has some musical artifacts. Importantly, artefacts decrease with an increase in the number of microphones.

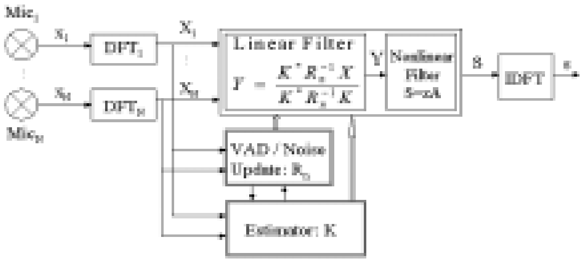

Figure 2: *The N-sensors enhancement scheme.*

5. CONCLUSIONS

This paper presents a formal extension of single channel STSA and log-STSA MMSE estimators to multichannel systems using Bayes sufficient statistics framework. The optimal scheme is decomposable into a linear beamforming-like filter followed by the single-channel Ephraim-Malah estimation scheme as proposed in [4, 5]. Our solution implicitely takes into account multi-path effects by using transfer function ratios.

We have applied this estimation technique to a four-microphone recording scheme. Results show a decrease in artefacts and an absolute improvement of about 1-2 dB segmental SNR over the single channel Ephraim-Malah estimate. Although the difference appears small, the relative improvement is significant particularly in the case of severe noise conditions. The multi-channel technique results in less artifacts compared to the single channel estimate, especialy when voice signals vary from strong to soft during the same utterance.

Although present work focused little on the choice of parameters of the multi-channel implementations, future work will adress the optimization of parameters of interest in order to fully exploit the potential of the microphone array schemes indicated by these results.

6. REFERENCES

[1] M. Brandstein, D. Ward, Microphone Arrays, Springer-Verlag, 2001.

[2] I. Cohen, “On Speech Enhancement Undert Signal Presence Uncertainty”, Proc. ICASSP 2001.

[3] J.R. Deller, J.H.L. Hansen, J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press , New York, 2000

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

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

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

[7] R. Martin, “Spectral Subtraction based on Minimum Statistics”, Signal Processing VII: Theories and Applications, 1182–1185, EASP , 1994

[8] R.J. McAulay, M.L. Malpass, ”Speech Enhancement Using a Soft-Decision Noise Suppression Filter”, IEEE Trans. ASSP 28: 137–145, 1980.

[9] H.V. Poor, An Introduction to Signal Detection and Estimation, Springer Verlag, New York, 1994.

[10] J. Rosca, R. Balan, N.P. Fan, C.Beaugeant, V. Gilg, “Multichannel Voice Detection in Adverse Environments”, EUSIPCO 2002.

[11] M.J. Schervish, “Theory of Statistics”, Springer-Verlag 1995.

[12] M. Zibulski, Y.Y. Zeevi ”Frame Analysis of the Discrete Gabor-Scheme”. IEEE Trans. SP 42: 942–945, 1994