Multi-Channel Estimation of Power Spectral Density Matrix Using Inter-Frame and Inter-Band Information

Raziyeh Ranjbaryan
Electrical Engineering Dept.
Yazd University, Yazd, Iran
Email: ranjbaryan@stu.yazd.ac.ir

Hamid Reza Abutalebi
Electrical Engineering Dept.
Yazd University, Yazd, Iran
Email: habutalebi@yazd.ac.ir

***Abstract*—In this paper, we address the estimation of power spectral density (PSD) matrix. The accurate estimation of PSD matrix plays an important role in many speech enhancement methods. In traditional PSD estimation methods, only the information of previous frames is employed through a forgetting factor. In the current research, we consider the correlation of inter-band components and incorporate their information to compute the PSD matrix more accurately. The simulation results are presented to confirm the efficiency of this method. They show that the performance of the speech enhancement method is substantially improved by using the proposed PSD estimation technique.**

I. Introduction

Accurate estimation of power spectral density (PSD) matrix has been of considerable interest in several signal processing applications, such as radar, sonar, speech and so on. For example, the performance of many speech enhancement methods such as Wiener filter, the minimum variance distortionless response (MVDR) or linearly constraint minimum variance (LCMV) beamformers are highly sensitive to the accurate estimation of PSD matrix; the more accurate PSD matrix the more effective performance can be achieved.

The effect of speech PSD matrix error in multichannel Wiener filter (MWF) was examined in [1] in the noise reduction application. In [2], a recursive smoothing method, based on the linear combination of PSD matrix at previous frames was presented. This technique is commonly used in speech enhancement algorithms. In this method, a forgetting factor controls the effect of consecutive frames. The proper choice of this parameter has been the subject of many researches; in [3], the forgetting factor is was tuned based on speech presence probability (SPP) in each time-frequency unit. In [4], a recursive smoothing method for noise PSD matrix estimation was introduced which uses current, previous and close subsequent noisy speech frames; the forgetting factor is iteratively updated based on the overall signal to noise ratio (SNR) in all microphone signals.

In recent years, the problem of inter-frame and inter-band correlation has received much attention. In [5], a multi-frame approach was proposed for noise reduction in the short time Fourier transform (STFT) domain. The authors took inter-frame correlation into account and proposed several optimal filters which improve the SNR.

The inherent harmonics of voiced speech and the windowing process which is commonly used in the STFT processing introduce some correlation between neighbor frequency components [6]. Accordingly, the effect of inter-band and inter-frame correlation has been utilized in many speech-related applications. In [7], a multi-dimensional short time spectral amplitude (STSA) estimator was proposed for speech enhancement that considers the correlation between frequency components. In [8], the inter-band correlation was considered to propose a single channel noise reduction filter in the STFT domain. Also, in [9] the authors proposed a single channel SPP estimation method based on both inter-frame and inter-band correlations to increase the detection accuracy. In each time-frequency unit, they used a vector containing the components of the adjacent frames and frequencies to compute the SPP similar to the one presented in [10].

Previous works for PSD matrix estimation consider only the information of inter-frame correlations. To estimate the PSD matrix more accurately, in the current work, we propose and examine the effect of both inter-band and inter-frame correlations. To this end, the PSD matrix is computed in three steps; first, for each time-frequency unit, we compute an initial matrix using the common method presented in [2]. In the second stage, we apply the effect of inter-band information by applying a mapping matrix [11]; the resulted matrix is called transformed matrix. Eventually, the final PSD matrix is computed by a linear combination of the initial and the transformed matrix.

Furthermore, we utilize the modified (improved) PSD matrix to improve the accuracy of SPP (as introduced in [10]). Then, we incorporate the SPP to improve the performance of Parametric multichannel Wiener filter (PMWF) [12]. In traditional PMWF, which employs a fixed parameter, noise reduction comes at the cost of signal distortion. In this paper, we propose an adaptive parameter based on the SPP to achieve a better balance between noise reduction and speech distortion.

This paper is organized as follows. After formulating the

ISBN 978-0-9928626-7-1

problem in section II, the proposed method is introduced in section III. In section IV, we review the employed method for the estimation of mapping matrix. Section V explains the effect of the proposed PSD matrix estimation in the computation of SPP and in the speech enhancement system. Sections VI and VII consist simulation results and conclusions, respectively.

II. Problem Formulation

Consider an N-element microphone array which captures the source signal in a noisy field. We assume that the received signal is corrupted by uncorrelated additive noise. In the STFT domain, the received signal can be expressed as

$$\begin{aligned} Y_n(m,k) &= G_n(k)S(m,k) + V_n(m,k) \\ &= \quad X_n(m,k) \quad + V_n(m,k), \end{aligned} \tag{1}$$

where $Y_n(m,k)$, $G_n(k)$, $S(m,k)$, $V_n(m,k)$ and $X_n(m,k)$ are the nth microphone signal, the impulse response from the source to the nth microphone, the source signal, the additive noise and the clean source signal at the nth microphone at time-frame m and discrete-frequency k, respectively. We assume that the noise and source signals are zero mean random processes.

The output of microphone array can be written as

$$\mathbf{y}(m,k) = \mathbf{g}(k)S(m,k) + \mathbf{v}(m,k), \tag{2}$$

where

$$\begin{aligned} \mathbf{y}(m,k) &= [Y_1(m,k),\ Y_2(m,k),\ ...,\ Y_N(m,k)]^T \\ \mathbf{g}(k) &= [G_1(k),\ G_2(k),\ ...,\ G_N(k)]^T \\ \mathbf{v}(m,k) &= [V_1(m,k),\ V_2(m,k),\ ...,\ V_N(m,k)]^T \\ \mathbf{x}(m,k) &= [X_1(m,k),\ X_2(m,k),\ ...,\ X_N(m,k)]^T \end{aligned}$$

and the superscript ^{T} denotes the transpose operation.

In this research, our goal is to estimate the output PSD matrix at each time-frequency unit, i.e.,

$$\begin{aligned} \mathbf{R}_{yy}(m,k) &= E\left\{\mathbf{y}(m,k)\mathbf{y}^H(m,k)\right\} \\ &= \mathbf{g}(k)R_{ss}(m,k)\mathbf{g}^H(k) + \mathbf{R}_{vv}(m,k) \\ &= \mathbf{R}_{xx}(m,k) + \mathbf{R}_{vv}(m,k) \end{aligned} \tag{3}$$

where $R_{ss}(m,k)$ is the variance of the source signal and $\mathbf{R}_{vv}(m,k)$ and $\mathbf{R}_{xx}(m,k)$ are the PSD matrices of the noisy and clean signals, respectively. The superscript ^{H} represents transpose-conjugate operation.

In the state-of-the-art methods [2], the output PSD matrix is computed recursively by using the information of previous frames and a forgetting factor as

$$\mathbf{R}_{yy}(m,k) = \lambda\mathbf{R}_{yy}(m-1,k) + (1-\lambda)\mathbf{y}(m,k)\mathbf{y}^H(m,k) \tag{4}$$

where λ is the forgetting factor.

Assuming the pseudo-stationarity of noise PSD, we can update the noise PSD matrix by applying (4) on silent frames. Consider that noise and source signals are uncorrelated, the PSD matrix of clean signal can be computed as

$$\mathbf{R}_{xx}(m,k) = \mathbf{R}_{yy}(m,k) - \mathbf{R}_{vv}(m,k). \tag{5}$$

III. Proposed Method For The Estimation Of Power Spectral Density Matrix

In this section, we present a new method for estimating the PSD matrix that considers the correlation of both inter-frame and inter-band components. In the proposed method, an initial PSD matrix is computed using (4); then we consider the frequency correlation between the previous band and current one. This process can be explained as follows by considering Fig. 1.

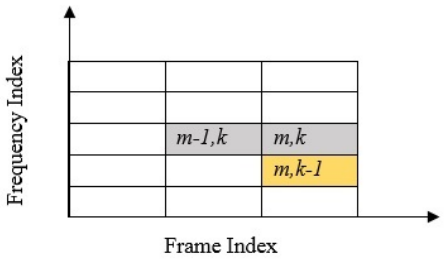

Fig. 1: Time and frequency illustration for computing PSD matrix

Since the signal subspace at each frequency is different from that at another frequency, it is not possible to simply add two PSD matrices of different frequencies directly. We use a mapping matrix to transform the PSD matrix from previous band to the current one; the result is called transformed matrix. After this transformation, we can combine the initial and the transformed matrix linearly to compute the final matrix.

The PSD matrix at time-frame m and discrete-frequency k is estimated in the following steps:

1) The initial PSD matrix is computed using (4)

$$\begin{aligned} &\mathbf{R}_{yy,init}(m,k) = \\ &\quad \lambda\mathbf{R}_{yy,final}(m-1,k) + (1-\lambda)\mathbf{y}(m,k)\mathbf{y}^H(m,k) \end{aligned} \tag{6}$$

where $\mathbf{R}_{yy,final}(m-1,k)$ is the final computed PSD matrix at time-frame $m-1$ and discrete-frequency k.

2) We apply mapping matrix $\mathbf{T}$ (explained more in next section) to transform the final computed matrix at the time-frame m and discrete-frequency $k-1$ to the current time-frequency unit. The result is

$$\mathbf{R}_{yy,trans}(m,k) = \mathbf{T}\mathbf{R}_{yy,final}(m,k-1)\mathbf{T}^H. \tag{7}$$

3) In the last step, we compute the final PSD matrix at time-frame m and discrete-frequency k as a linear combination of the initial and transformed matrices using the relaxation term γ:

$$\begin{aligned} &\mathbf{R}_{yy,final}(m,k) = \\ &\quad \gamma\mathbf{R}_{yy,trans}(m,k) + (1-\gamma)R_{yy,init}(m,k). \end{aligned} \tag{8}$$

IV. Mapping Matrix Estimation

In order to transform the PSD matrix from one frequency band to another one, a mapping matrix is employed. We use rotational signal-subspace (RSS) mapping matrix [11] that has

ISBN 978-0-9928626-7-1 © EURASIP 2017

been presented in [13] for coherent signal subspace method (CSSM) in wideband direction of arrival (DOA) estimation. In a similar manner, in the current work, we apply the RSS matrix to transform the PSD matrix from the previous band to the current one as seen in (7). The mapping matrix is obtained as the solution of

$$\mathbf{T} = \arg\min_{\mathbf{T}} \|\mathbf{T}\mathbf{g}(k-1) - \mathbf{g}(k)\|_F \quad s.t. \quad \mathbf{T}^H\mathbf{T} = \mathbf{I}, \quad (9)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm, and $\mathbf{I}$ is the identity matrix. The solution has been given in [11] as

$$\mathbf{T} = \mathbf{V}\mathbf{U}^H. \quad (10)$$

Using the singular value decomposition, the unitary matrices $\mathbf{V}$ and $\mathbf{U}$ and diagonal matrix $\mathbf{\Sigma}$ are obtained such that

$$\mathbf{g}(k-1)\mathbf{g}^H(k) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H. \quad (11)$$

V. Adaptive Parametric Multichannel Wiener Filter

In this section, we firstly employ the proposed PSD estimation technique to increase the accuracy of SPP. We consider the SPP estimation method that introduced in [10] and expressed as

$$p = \{1 + \frac{q}{1-q}[1+\zeta]exp[-\frac{\beta}{1+\zeta}]\}^{-1}, \quad (12)$$

where

$$\zeta = trace\{\mathbf{R}_{nn}^{-1}(m,k)\mathbf{R}_{xx}(m,k)\}$$
$$\beta = \mathbf{y}^H(m,k)\mathbf{R}_{nn}^{-1}(m,k)\mathbf{R}_{xx}(m,k)\mathbf{R}_{nn}^{-1}(m,k)\mathbf{y}(m,k).$$

and q denotes the priori speech absence probability (SAP).

In the next step, we incorporate the above-mentioned SPP estimator in the PMWF for speech enhancement. The cost function of PMWF is defined as:

$$\begin{aligned}\varepsilon(m,k) = &[\mathbf{E} - \mathbf{h}(m,k)]^H\mathbf{R}_{xx}(m,k)[\mathbf{E} - \mathbf{h}(m,k)] \\ &+ \mu\mathbf{h}^H(m,k)\mathbf{R}_{nn}(m,k)\mathbf{h}(m,k).\end{aligned} \quad (13)$$

Considering the first microphone as a reference, $\mathbf{E} = [1, 0, \cdots, 0]^H$, $\mathbf{h}(m,k)$ is the filter coefficients and μ is the PMWF parameter. The first term of (13) is considered as a measure of the signal distortion and the second term as the noise reduction.

Filter coefficients are calculated as below [12]:

$$\mathbf{h}(m,k) = \frac{\mathbf{R}_{nn}^{-1}(m,k)\mathbf{R}_{xx}(m,k)}{\mu + trace(\mathbf{R}_{nn}^{-1}(m,k)\mathbf{R}_{xx}(m,k))}. \quad (14)$$

In the special case that $\mu = 0$, PMWF is equivalent to MVDR; also, when $\mu = 1$, it is equivalent to multichannel Wiener filter. In order to better moderate the trade-off between signal distortion and noise reduction, an adaptive parameter based on the SPP is introduced. To this end, the cost function is reformulated as

$$\begin{aligned}&\varepsilon(m,k) = \mu(p(m,k))\mathbf{h}^H(m,k)\mathbf{R}_{nn}(m,k)\mathbf{h}(m,k) \\ &+[1-\mu(p(m,k))][\mathbf{E}-\mathbf{h}(m,k)]^H\mathbf{R}_{xx}(m,k)[\mathbf{E}-\mathbf{h}(m,k)],\end{aligned} \quad (15)$$

where $p(m,k)$ is the SPP parameter.

In each time-frequency unit where $p(m,k) = 0$, there exist only noise components; so, we set $\mu(p(m,k) = 0) = 1$. Remaining only the first term in (15), we can deal with noise reduction without worrying about signal distortion. On the other hand, where $p(m,k) = 1$, there exist speech components; it is more sensible to reduce signal distortion but, we prefer to keep a little noise to avoid musical noise. Hence, we set $\mu(p(m,k) = 1) \longrightarrow 0$ which leads to minimizing the signal distortion and increasing intelligibility. For this purpose, we need a smooth and decreasing function based on the SPP. Accordingly, an exponential function is suggested as below:

$$\mu(p(m,k)) = exp(-0.5 * p^2(m,k)) \quad (16)$$

VI. Simulation Results

In this section, we compare the performance of the proposed PSD estimation method with the traditional one which utilizes only the inter-frame correlation in (4). The effect of PSD estimation method in the performance of MVDR beamformer has been examined for different types of noise and echo conditions.

In this simulation, speech signal samples are taken from TIMIT database [14]. The sampling frequency is 16 kHz and the STFT is implemented using 32 ms Hamming window with 50% overlap. The room dimensions are considered $3m \times 4m \times 2.5m$ (width×length×height). A uniform linear array (ULA) of $N = 5$ microphones are placed on the axis ($x_n = x_{init} + (n-1)d,\ y = 2\ m,\ z = 1.3\ m),\ n = 1, \cdots, 5$, where $x_{init} = 1\ m$ and $d = 0.05\ m$. The source signal is located at ($x = 1.9\ m,\ y = 2.77\ m,\ z = 1.3\ m$). It is assumed that the microphone signals are corrupted by additive noise at different SNRs. The image method [15] was used to generate the impulse response from the source to the microphones.

Considering ideal voice activity detection and performing an empirical analysis, it was observed that the best performance is achieved by choosing $\lambda = 0.998$ and $\gamma = 0.96$.

We use a recursive smoothing method both in time and frequency domain which yields more accurate PSD matrix. It is noted that in the case of colored noise, we update only the PSD matrix of noisy data $\mathbf{R}_{yy}(m,k)$ according to (4), but in the presence of white Gaussian noise, the PSD matrices of both noisy data $\mathbf{R}_{yy}(m,k)$ and noise $\mathbf{R}_{nn}(m,k)$ are updated. Actually, the spectrum of white noise has the same amplitude over all frequencies, which justifies correlation between frequency components.

The performance of the proposed and traditional PSD matrix estimation are compared in speech enhancement applications in terms of two objective measures, namely PESQ and segmental SNR. In this experiment, we use speech samples from four male and four female speakers and report the average comparative results.

In Fig. 2 and Fig. 3, we depict the PESQ value at different input SNRs for different kinds of noise in anechoic and reverberant condition ($RT_{60} = 200\ ms$), respectively. It is seen that the proposed method outperforms the traditional one

ISBN 978-0-9928626-7-1 © EURASIP 2017

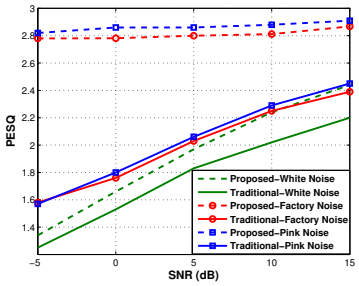

Fig. 2: PESQ at different input SNRs for MVDR beamformer, comparing the traditional and proposed method in anechoic condition.

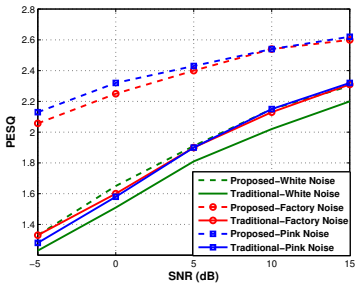

Fig. 3: PESQ at different input SNRs for MVDR beamformer, comparing the traditional and proposed method in reverberant condition ($RT_{60} = 200\ ms$).

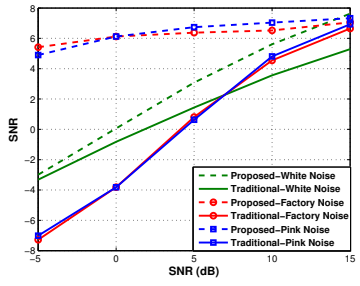

Fig. 4: Segmental SNR at different input SNRs for MVDR beamformer, comparing the traditional and proposed method in anechoic condition.

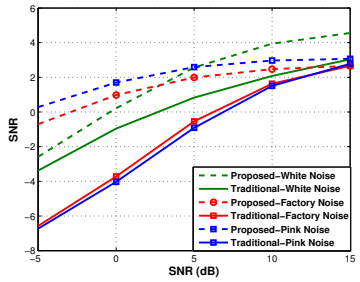

Fig. 5: Segmental SNR at different input SNRs for MVDR beamformer, comparing the traditional and proposed methods in reverberant ($RT_{60} = 200\ ms$).

in all input SNR values. The proposed method has drastically improved the PESQ. This can be justified by considering the effect of inter-band correlation in PSD matrix estimation.

The superiority of the proposed method is much more evident in the case of colored noises (compared to that in the case of white noise). This can be justified by considering that both the speech and the colored noise (e.g. pink noise) have similar lowpass spectra; so, the traditional PSD matrix estimators fail to perform correctly, especially in low SNRs.

Also, Fig. 4 and Fig. 5 shows the segmental SNR value at different input SNRs for different kinds of noise in anechoic and reverberant condition ($RT_{60} = 200\ ms$), respectively. Results demonstrate the performance improvement obtained by the usage of proposed method for PSD estimation.

Fig. 6 and Fig. 7 illustrate the spectrograms of sample clean data at the first microphone, noisy data at $SNR = 10\ dB$, the enhanced signal (output of MVDR beamformer) with the traditional PSD matrix estimation method, and with the proposed PSD matrix estimation method, in the case of white Gaussian noise and factory noise respectively. It is seen that incorporating the inter-band correlations, the proposed PSD estimation method considerably improves the performance of speech enhancement. The noise is significantly reduced while the speech is not substantially distorted.

Furthermore, we utilized the proposed PSD estimation technique to increase the accuracy of SPP in (12). We suppose that priori SAP is a fixed parameter (e.g., $q = 0.6$ as in [10]). We incorporate the SPP to improve the trade-off between noise reduction and speech distortion in PMWF. In Fig. 8, the PESQ values at different input SNRs for fixed parameter-based PMWF and adaptive parameter based on (16) in the presence of white Gaussian noise in anechoic condition are depicted. In contrast to the fixed parameter-based PMWF, where noise reduction introduces speech distortion, an adaptive parameter-based PMWF leads to a better trade-off and consequently more improvement in intellegibility and PESQ, especiallly in high SNRs.

VII. Conclusion

In this paper, the inter-band correlation was considered in the estimation of PSD matrix. We took this correlation into account and applied a mapping technique that had already been used for coherent signal subspace methods. We utilized this method to transform PSD matrix into different frequencies. The final PSD matrix was computed as the linear combination of the transformed and initial matrices. We examined the effect of PSD estimation technique in the speech enhancement application using MVDR beamformer. It was shown that the PESQ and segmental SNR parameters are improved by exploiting both inter-band and inter-frame correlations. Simulation results confirm the superiority of the proposed method over the traditional method for PSD matrix estimation as well.

ISBN 978-0-9928626-7-1 © EURASIP 2017

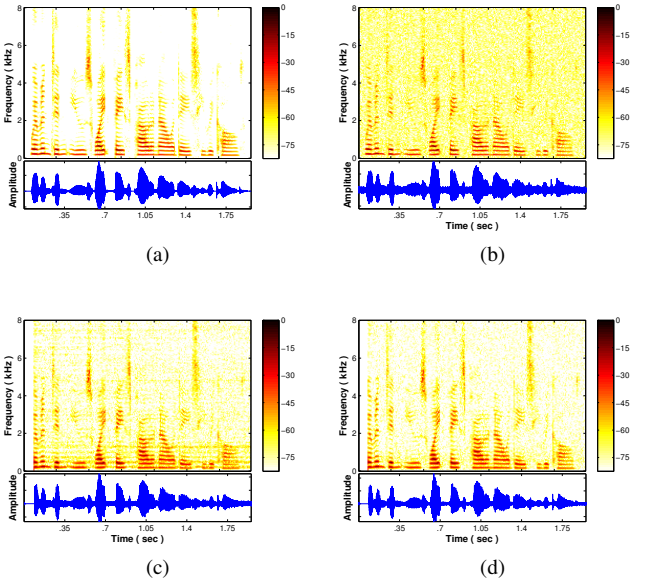

Fig. 6: Spectrograms of (a) clean, (b) noisy data, (c) enhanced signal using the traditional method and (d) enhanced signal using the proposed method for PSD matrix estimation in the case of white Gaussian noise.

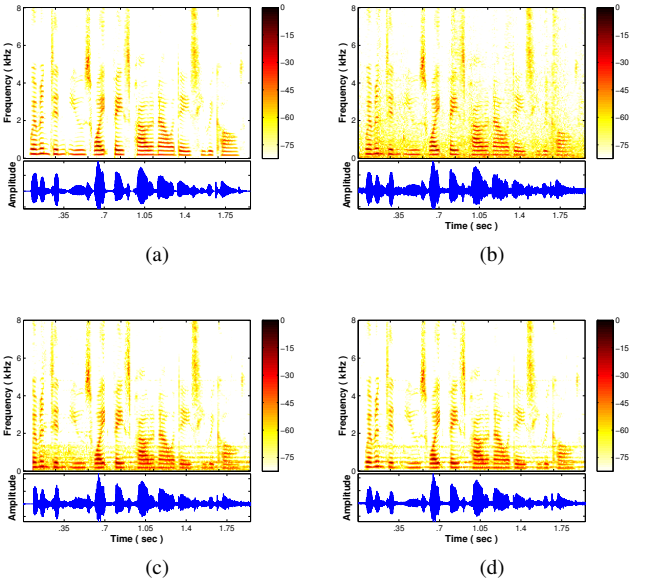

Fig. 7: Spectrograms of (a) clean, (b) noisy data, (c) enhanced signal using the traditional method and (d) enhanced signal using the proposed method for PSD matrix estimation in the presence of factory noise.

REFERENCES

[1] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, July 2011.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[3] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sept 2011.

[4] M. Parchami, W. P. Zhu, and B. Champagne, "A new algorithm for noise psd matrix estimation in multi-microphone speech enhancement based on recursive smoothing," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015, pp. 429–432.

[5] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[6] T. Fingscheidt, C. Beaugeant, and S. Suhadi, "Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, March 2005, pp. 1081–1084.

[7] E. Plourde and B. Champagne, "Multidimensional stsa estimators for speech enhancement with correlated spectral components," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3013–3024, July 2011.

[8] J. Chen and J. Benesty, "Single-channel noise reduction in the stft domain based on the bifrequency spectrum," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 97–100.

[9] H. Momeni, E. A. P. Habets, and H. R. Abutalebi, "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2903–2907.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[11] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1272–1281, Aug 1988.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, 1st ed. Springer-Verlag Berlin Heidelberg, 2008, vol. 1.

[13] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, Aug 1985.

[14] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburgh, MD, Tech. Rep., Dec. 1988.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America Journal*, vol. 65, pp. 943–950, Apr. 1979.

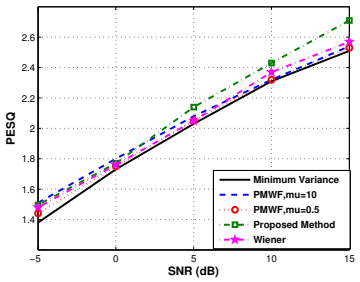

Fig. 8: PESQ at different input SNRs, comparing the effect of parameter μ on the PMWF in the case of white Gaussian noise in anechoic condition.

ISBN 978-0-9928626-7-1 © EURASIP 2017