AN OVERCOMPLETE WDFT SINUSOIDAL BASIS FOR PERCEPTUALLY MOTIVATED SPEECH ENHANCEMENT[*)]

Adam Borowicz and Alexander Petrovsky

Department of Real-Time Systems, Bialystok Technical University
Wiejska 45A, 15-351, Bialystok, Poland
phone: + (48) 085 746-90-50, fax: + (48) 085 746-90-57, email: borowicz@ii.pb.bialystok.pl
web: www.pb.bialystok.pl

ABSTRACT

This paper considers an application of the warped discrete Fourier transform (WDFT) in the perceptually motivated speech enhancement. Namely, the problem of signal distortions generated by WDFT synthesis block is of interest. Spectral features of the reconstructed signal are analyzed and discussed in context of the perceptual processing. Our proposition is to construct an overcomplete WDFT sinusoidal basis in order to minimize reconstruction error. The new approach is validated in practical speech enhancement system. The results show that the proposed algorithm outperforms both conventional DFT and pure WDFT solutions.

1. INTRODUCTION

The discrete Fourier transform (DFT) is a powerful tool in the uniform spectral analysis. However, a variety of applications requires nonuniform frequency decomposition. An example is perceptually motivated speech enhancement.

Recently proposed the warped discrete Fourier transform (WDFT) [1] enables non-uniform sampling the *z*-transform of finite length sequence by using allpass function. In the context of auditory based speech enhancement, there is a need for perceptual warping that allocates frequency samples in good accordance with psychoacoustic scale (Bark or ERB) [2].

Most of existing noise reduction systems, work in the frequency domain using well known spectral weighting technique. Although these methods are very simple and easy to implement, their weak point is the residual noise also known as musical tones. The verified approach is to modify the weighting rule to keep musical tones slightly below the masking threshold [3].

In our previous work [4] we employed WDFT in perceptual noise reduction system not only as a basis for masking model but also as frequency decomposition tool. In that way, overall processing is performed solely in the warped spectrum domain and transformations between different frequency scales are no longer needed. It results in simplified system architecture. Moreover speech processing performed in critical band domain is more accurate in the context of psychoacoustic modelling. The performance of the WDFT based speech enhancement system seems to be quite satisfactory in the case of speech signals with dominant low frequency components. Unfortunately, for wideband signals the distortions in high frequency range may be noticeable and further improvements are needed.

The possible cause of the degradation of high frequency components is imperfect WDFT synthesis block which only approximates inverse transform. Therefore, our proposition is to construct an overcomplete WDFT sinusoidal basis in order to minimize reconstruction error. In that way we define a new extension of WDFT which is more suitable for perceptual speech processing than ordinary WDFT.

2. WDFT

The warped discrete Fourier transform (WDFT) is a special case of nonuniform DFT. It has frequency samples allocated nonuniformly but regularly over the unit circle. The WDFT can be defined using the matrix representation as follows

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[N-1] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \hat{z}_0^{-1} & \cdots & \hat{z}_0^{-N+1} \\ 1 & \hat{z}_1^{-1} & \cdots & \hat{z}_1^{-N+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \hat{z}_{N-1}^{-1} & \cdots & \hat{z}_{N-1}^{-N+1} \end{bmatrix}}_{\mathbf{D}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}, \qquad (1)$$

with $\hat{z}_k$ being the images of allpass transformed equidistant points of the unit circle

$$z_k^{-1} = e^{-j2\pi k/N} \quad \rightarrow \quad \hat{z}_k^{-1} = A(z_k), \qquad k = 0 \ldots N-1 . \qquad (2)$$

$A(z)$ can be an arbitrary allpass function. Note that $\mathbf{D}$ is the Vandermonde matrix. The determinant of such a matrix has non-zero value for distinct points $\hat{z}_k$. As this condition is satisfied, the invertibility of the WDFT is guaranteed from the theoretical point of view. But for perceptual warping [2] which is obtained using first order allpass function

$$A(z) = \frac{z^{-1} - a}{1 - az^{-1}}, \quad |a| < 1 , \qquad (3)$$

some *z*-transform points are placed close together on the unit circle and numerical problems arise.

Fortunately, there are number of methods for approximate inversion of ill-conditioned matrices. They exploit the singular value decomposition (SVD) defined as

$$\mathbf{D} = \sum_{i=1}^{N} \mathbf{u}_i \sigma_i \mathbf{v}_i^H , \qquad (4)$$

where $\mathbf{u}_i$, $\mathbf{v}_i$ are orthogonal columns and H denotes Hermitian transpose, σ_i are singular values. The error amplification can be measured using a matrix condition number

$$\operatorname{cond}(\mathbf{D}) = \|\mathbf{D}\| \|\mathbf{D}^{-1}\| = \sigma_{\max} / \sigma_{\min} . \qquad (5)$$

The singular value distributions for several different WDFT matrices are depicted in Fig. 1. It can be easily verified that many singular values are close to zero.

[*)] This work was supported by Bialystok Technical University under the grant W/WI/2/04.

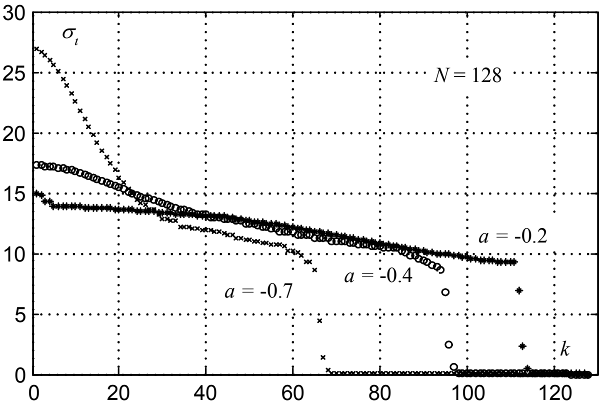

Figure 1. Distribution of singular values as function of warping coefficient for fixed transform size.

The SVD components not only provide many useful insights into matrix ill-conditioning but also, they can be used to form a pseudoinverse matrix

$$\mathbf{D}^{\dagger} = \sum_{i=1}^{N} f_i \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^H . \qquad (6)$$

The symbols f_i denote the so-called filter factors and they all have to equal unity for precise inverse. The regularization theory [5] advises to eliminate the influence of small singular values by weakening their contribution in (6). This is done by setting f_i appropriately. Unfortunately, as we show in the next Section, this operation changes the spectral content of the data.

3. RECONSTRUCTION ERROR ANALYSIS

Using matrix notation, we can define the signal distortion vector as the difference between an original and a reconstructed signal column vector

$$\mathbf{d} = \mathbf{x} - \hat{\mathbf{x}} = \left(\mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D}\right)\mathbf{x} . \qquad (7)$$

Corresponding spectral domain measure can be formulated as the average power spectrum of the signal distortion

$$S_{dd}(\omega) = \frac{1}{N} E\left\{\left|\mathbf{e}(\omega)\mathbf{d}\right|^2\right\} = \frac{1}{N}\mathbf{e}(\omega)\mathbf{R}_{\mathbf{dd}}\mathbf{e}(\omega)^H , \qquad (8)$$

where $E\{\}$ is expectation symbol and

$$\mathbf{e}(\omega) = \begin{bmatrix} 1 & e^{-j\omega\cdot 1} & e^{-j\omega\cdot 2} & \dots & e^{-j\omega\cdot(N-1)} \end{bmatrix} \qquad (9)$$

is DFT-related sinusoidal basis vector. $\mathbf{R}_{\mathbf{dd}}$ denotes the covariance matrix of the distortion signal. Let $\mathbf{Q} = \mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D}$, then

$$\mathbf{R}_{\mathbf{dd}} = \mathbf{Q}\mathbf{R}_{\mathbf{xx}}\mathbf{Q}^H , \qquad (10)$$

where $\mathbf{R}_{\mathbf{xx}}$ is covariance matrix of the processed signal. It is clear that the spectral distortion (8) defined as absolute error, depends on input signal characteristic and fidelity of the inverse WDFT approximation. Theoretically, for non-zero $\mathbf{R}_{\mathbf{xx}}$ an exact inverse is possible when all elements of $\mathbf{Q}$ are equal to zero. Since, the pseudoinverse is computed using SVD approach (6)

$$\mathbf{Q} = \mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D} = \mathbf{I} - \sum_{i=1}^{N} f_i \mathbf{v}_i \mathbf{v}_i^H , \qquad (11)$$

appropriate setting of filter factors may be helpful. Unfortunately, in the case of perceptual warping, choice of the regularization parameters has minimal impact on the spectral distortion level. We found that, even unstable approximation of the inverse WDFT, produces relatively high reconstruction error and it is magnified by further stabilization.

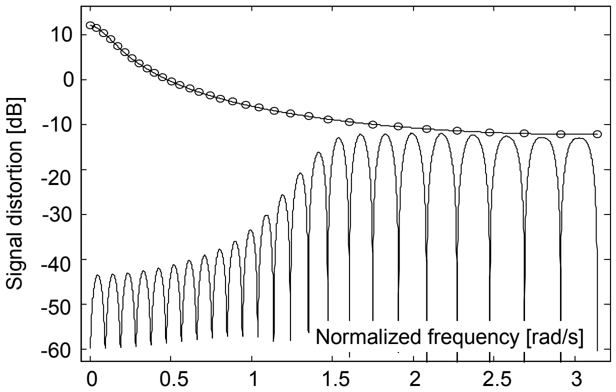

Figure 2. Power spectrum of original signal and corresponding synthesis error (solid line).

The only way to minimize spectral distortion is to modify the transform matrix in such way that a number of almost zero singular values will be decreased. It is identical to reducing eccentricity of SVD ellipsoid that is an image of the unit sphere in N-dimensional space.

Fig. 2 depicts power spectrum of an example input signal (colored Gaussian noise) and corresponding spectral distortion computed analytically using (8). It can be observed that the level of distortions at a given frequency depends on distance between neighbouring WDFT bins. Signal is perfectly reconstructed only at transform points and the spectral distortions are evident especially in stretched frequency regions while in the compressed regions the synthesis error seems to be acceptable.

4. OVERCOMPLETE WDFT VECTOR BASIS

4.1 Geometric signal theory

The k-th row of the WDFT matrix is in fact a complex sinusoidal vector

$$\mathbf{s}_k = \begin{bmatrix} 1 & \hat{z}_k^{-1} & \hat{z}_k^{-2} & \dots & \hat{z}_k^{-N+1} \end{bmatrix} . \qquad (12)$$

Theoretically for distinct transform points the rows of the WDFT matrix are linearly independent (due to non-zero determinant of the Vandermonde matrix). Therefore, WDFT sinusoidal vectors (12) form a non-orthogonal basis $S = \{\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{N\text{-}1}\}$ for complex vector space $\mathbb{C}^N$, so that any input vector can be expressed as linear combination of them.

If we relax the constraint that the vectors have to be independent we can construct an overcomplete non-orthogonal basis. Although, such a basis is redundant for defining vector space, it can prove invaluable for synthesis error correction. Namely, by appropriate selection of vector basis we can modify the singular value distributions of the corresponding transform matrix.

Note, that since we suppose overcomplete basis, the new WDFT matrix is no longer square and number of rows $M > N$ is increased. Matrix representation of the extended WDFT can be formulated as follows

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[M-1] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \hat{z}_0^{-1} & \cdots & \hat{z}_0^{-N+1} \\ 1 & \hat{z}_1^{-1} & \cdots & \hat{z}_1^{-N+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \hat{z}_{M-1}^{-1} & \cdots & \hat{z}_{M-1}^{-N+1} \end{bmatrix}}_{\mathbf{D}_{M\times N}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} , \qquad (13)$$

A generalized inverse [6] of the rectangular matrix can be easily found using the same SVD procedure as for square WDFT matrix.

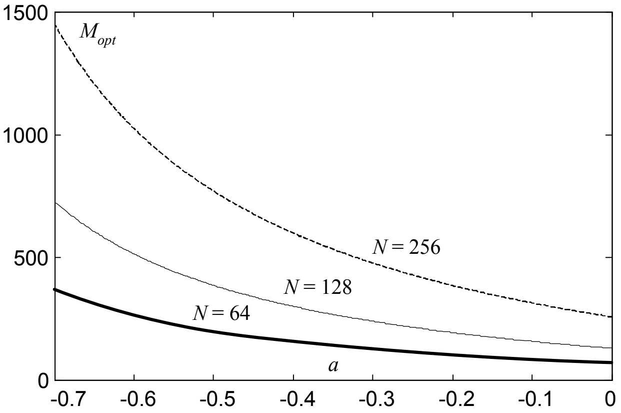

Figure 3. Relation between value of the warping parameter and size of the overcomplete basis.

4.2 Constructing overcomplete basis

From a practical point of view, there are least two conditions for building an overcomplete WDFT vector basis. Firstly, we need to preserve the regularity of transform points (uniform frequency resolution in a psychoacoustic scale). Secondly, size of the new sinusoidal basis should be small as possible to minimize complexity load.

Because, every direction in $\mathbb{C}^N$, indicated by basis vector, corresponds to particular frequency region, it is straightforward to consider the WDFT operation as a critically decimated filter bank. The k-th row of WDFT matrix can be viewed as a finite response (FIR) filter with transfer function given by

$$H_k(z) = \sum_{n=-\infty}^{\infty} A(z_k)z^{-n}, \quad k = 0 \ldots N-1. \tag{14}$$

Since, $A(z)$ is first-order allpass function (3), $H_k(z)$ is a bandpass filter with a center frequency at

$$\hat{\omega}_k = 2\tan^{-1}\left(\frac{1+a}{1-a}\tan\left(\frac{\omega_k}{2}\right)\right), \quad \omega_k = \text{angle}(z_k) \tag{15}$$

and bandwidth about $2\pi/N$.

The new overcomplete basis should be constructed from $M > N$ vectors corresponding to FIR filters whose center frequencies are allocated regularly over unit circle. It can be done by redefining transform points

$$z_k = e^{j\pi k/M}, \quad k = 0 \ldots M-1. \tag{16}$$

If we assume negative value of the allpass parameter, the maximal angular distance between the new z-transform points is

$$\Delta\omega_{\max} = \pi - 2\tan^{-1}\left(\frac{1+a}{1-a}\tan\left(\frac{\pi - 2\pi/M}{2}\right)\right). \tag{17}$$

To obtain the same spectral resolution in the high frequency regions as for conventional DFT we assumed that separation of the z-transform points is not greater than $2\pi/N$. Substituting $\Delta\omega_{\max} = 2\pi/N$ into (17) and solving for M we obtain

$$M = M_{opt} = 2\pi\left[\pi - 2\tan^{-1}\left(\frac{1-a}{1+a}\tan\left(\frac{\pi - 2\pi/N}{2}\right)\right)\right]^{-1}. \tag{18}$$

The relationship between the parameter a and M for different size of analysis window can be observed in Fig. 3. It is easy to see, that for strong warping the size of the overcomplete basis set rapidly increases.

Now, we can back on moment to the reconstruction error analysis presented in Section 3. It is clear that the spectral distortion (8) can be computed for rectangular WDFT matrices as well as for square matrices.

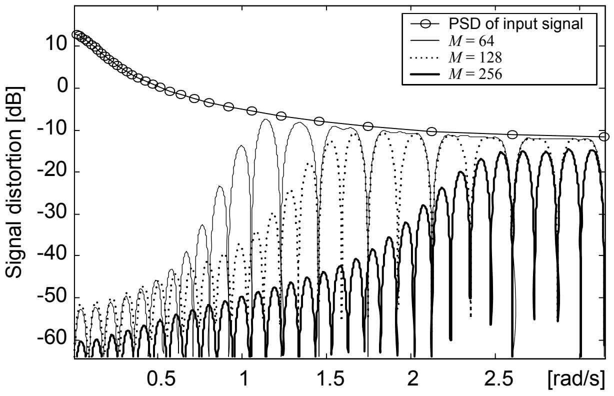

Figure 4. The power spectrum of an example input signal and corresponding spectral distortions calculated for different-size WDFT transforms with constant $N = 64$.

Fig. 4 depicts the power spectrum of an example input signal and corresponding spectral distortions calculated for different-size WDFT transforms. As it is shown, the synthesis error decreases for high M and can be completely neglected for $M > M_{opt}$.

5. SPEECH ENHANCEMENT EXPERIMENTS

5.1 System overview

The complete implementation details of the WDFT based noise reduction system were described in our previous work [4]. Here, only brief outline is presented.

Common psychoacoustically motivated spectral weighting technique [7] is used for noise reduction. If we denote predefined residual noise level by ζ_n, the weighting function can be expressed as follows

$$H^{IND}(\omega) = \min\left\{1, \sqrt{\frac{R_{TT}(\omega)}{R_{nn}(\omega)}} + \zeta_n\right\}, \quad 0 \le H^{IND}(\omega) \le 1. \tag{19}$$

To calculate weighting coefficients the estimates of the masking threshold $R_{TT}(\omega)$ and noise power spectral density $R_{nn}(\omega)$ are only needed. The simple modification of Johnston's perceptual entropy model [8] is used as a basis for clean speech masking threshold estimation. As a pre-processor for masking model, log spectral amplitude (LSA) estimator was implemented. The noise power spectrum estimation is performed via minima controlled recursive averaging (MCRA) approach [9].

Three variants of the noise reduction system were tested: the first with conventional DFT based analysis/synthesis block, the second with ordinary WDFT block and the last with extended WDFT block. In all cases the sampling frequency was 8 kHz. Input samples were partitioned into frames of length $N = 256$ (32 ms) with 50% overlap and multiplied by Hanning window. In order to reduce computational complexity, the size of the analysis window of extended WDFT system was reduced to 128 samples. It can be seen in Fig. 3 that, optimal size of overcomplete WDFT basis for $N = 128$ and the allpass parameter $a = -0.4$ is close to 300. From practical reasons, this value can be slightly decreased. In this way, the final size of the extended WDFT matrix was 256×128.

5.2 Performance evaluation

For our experiments the set of eight speech sentences with strong high frequency components was selected. The sentences were about a 5-8 s long. The coloured noise was added to the clean signals such that the segmental signal to noise ratio (SEGSNR) was between -5 dB and 20 dB.

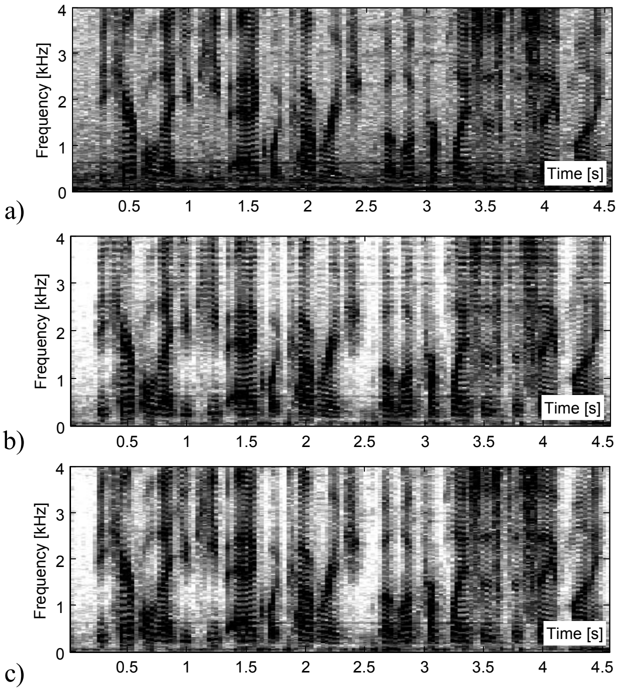

Figure 6. Spectrograms of noisy speech (a), speech enhanced with pure WDFT based method (b) and proposed method (c).

Noise attenuation factor (NA) defined as the mean ratio of the input to output noise power, was used to evaluate the suppression capabilities of tested systems. Speech distortions were measured using SEGSNR, where the noise was interpreted as a difference between original and enhanced speech [10]. The higher value of this factor indicates the weaker speech distortions.

The result of experiments is depicted in Fig. 7. Extended WDFT system has significantly better noise attenuation (NA) performance than conventional DFT system and comparable to pure WDFT method. In the case of speech distortion measure (SEGSNR), the similar results are obtained for extended WDFT and DFT systems. However, they are much better than for pure WDFT method. It is not surprise since synthesis error is efficiently reduced using overcomplete basis.

Spectral structure of the residual noise and speech distortions can be verified in Fig. 6. In the low frequency region there is no noticeable difference between the systems. Contrary to pure WDFT solution the high frequency distortions are not produced by extended WDFT synthesis block and the performance of speech enhancement system in high frequency range is not degenerated.

6. CONCLUSIONS

The method for cancellation of the synthesis error was developed. The proposed algorithm of constructing the overcomplete WDFT basis provides solution that not only solves the reconstruction problem but also allows obtaining the same spectral resolution in the high frequency regions as for conventional DFT. Experiments were done for extended WDFT speech enhancement system. The results clearly show that the new algorithm outperforms not only conventional DFT method but also pure WDFT system. Now, high quality perceptual speech processing is possible even for wideband signals.

In our case computational complexity was reduced by shortening analysis window but in general case it is not always possible. Commonly, cancellation of the synthesis error is done at a cost of increased computational load. Therefore, further work is aimed at reduction of computational complexity of the new WDFT scheme.

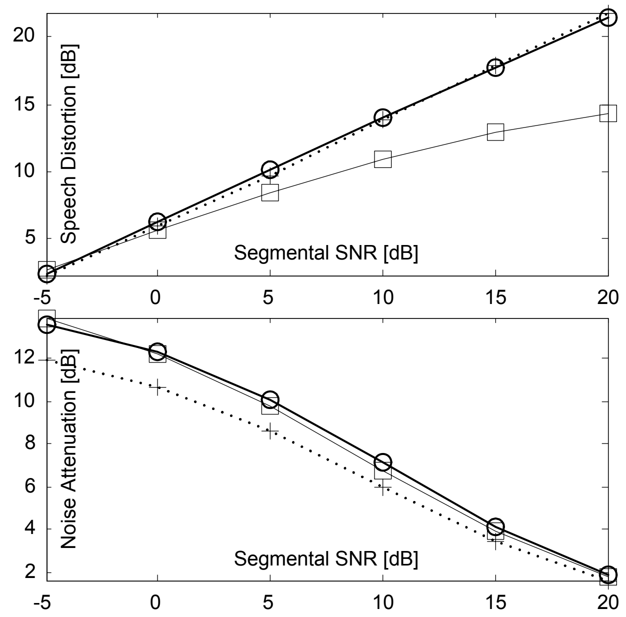

Figure 7. Performance evaluation: DFT based system (crosses), pure WDFT (squares), extended WDFT (circles).

REFERENCES

[1] A. Makur, S.K. Mitra, "Warped Discrete-Fourier Transform: Theory and Applications," *IEEE Trans. Circuits Systems I*, vol. 48, no. 9, pp. 1086-1093, 2001.

[2] J.O. Smith III, J.S. Abel, "Bark and ERB Bilinear Transforms," *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 697-708, 1999.

[3] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, 1999.

[4] A. Petrovsky, M. Parfieniuk, A. Borowicz, "Warped DFT based perceptual noise reduction system," *Proc. AES 116th*, Berlin, Germany, Conv. Paper #6035, 2004.

[5] P.C. Hansen, "The truncated SVD as a method for regularization," *BIT*, vol. 27, pp. 534-553, 1987.

[6] R. Penrose, "A Generalized Inverse for Matrices," *Proc. Cambridge Phil. Soc. 51*, pp. 406-413, 1955.

[7] S. Gustafsson, P. Jax, P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristic," *IEEE Int. Conf. on Acoustic, Speech and Signal Processing ICASSP'98*, Seattle, USA, vol. 1, pp. 397-400, 1998.

[8] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. on Selected Areas in Comm.*, vol. 6, pp. 314-323, 1988.

[9] I. Cohen, B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-15, 2002.

[10] S. Wang, A. Sekey, A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. on Selected Areas in Comm.*, vol. 10, no. 5, pp. 819-829, 1992.