Experimental Study of Generalized Subspace Filters for the Cocktail Party Situation

Christensen, Knud Bank; Christensen, Mads Græsbøll; Boldt, Jesper B.; Gran, Fredrik

Published in:
Proceedings IEEE ICASSP 2016

DOI (link to publication from Publisher):
10.1109/ICASSP.2016.7471709

Publication date:
2016

Document Version
Early version, also known as pre-print

EXPERIMENTAL STUDY OF GENERALIZED SUBSPACE FILTERS FOR THE COCKTAIL PARTY SITUATION

Knud B. Christensen[1]*, Mads G. Christensen*[1]*, Jesper B. Boldt*[2] *and Fredrik Gran*[2]

[1]Audio Analysis Lab, AD:MT
Aalborg University, Denmark
{kbc,mgc}@create.aau.dk

[2]GN ReSound A/S
Lautrupbjerg 7, DK-2750 Ballerup
{jboldt,fgran}@gnresound.com

ABSTRACT

This paper investigates the potential performance of generalized subspace filters for speech enhancement in cocktail party situations with very poor signal/noise ratio, e.g. down to -15 dB. Performance metrics output signal/noise ratio, signal/distortion ratio, speech quality rating and speech intelligibility rating are mapped as functions of two algorithm parameters, revealing clear trade-off options between noise, distortion and subjective performances and a recommended choice of trade-off. Given sufficiently good noise statistics, SNR improvements around 20 dB as well as PESQ quality and STOI intelligibility rating improvements exceeding 1.0 and 0.2 points respectively are found. This shows the potential of the method.

Index Terms— Speech enhancement, subspace signal processing, optimal filtering, babble noise, cocktail party

1. INTRODUCTION

The topic of this paper is to show the achievable objective and subjective performance of generalized subspace filters [1, 2] in single-channel speech enhancement applications for the cocktail party situation.

Oral communication systems play a vital role in society in the shape of telephony, intercom systems, public address systems, hearing instruments, etc. The increasing availability of real-time digital signal processing (DSP) for low-cost and/or low-power applications has facilitated the development and use of ever more advanced noise reduction and speech enhancement algorithms in everyday communication systems [3, 4]. And the trend continues.

One application which is pressed hard on DSP ressources, and where much may thus still be gained, is hearing instruments. And the most demanding use-case for a speech enhancement algorithm is that where the noise source is also speech, in babble noise or, in hearing instrument terms, the cocktail party situation.

Thanks to Innovation Fund Denmark, Grant No. 99-2014-1.

Analog intercoms, hearing aids and PA systems have traditionally attempted enhancement by simple bandpass filtering and noise gating. In 1960 Herman Schroeder [5] suggested an analog signal processing apparatus which would suppress both inharmonic and inter-formant harmonic noise- and distortion components in a speech signal.

The advent of DSP boosted the field, and soon spectral subtraction, optimal filtering and speech model based enhancement methods were discovered [6, 7, 8]. Also, the concept of subspace processing came up [9, 10, 7]. In [2, 4] subspace processing is given a filter matrix formulation, and automatic whitening with generalized SVD (aka joint diagonalization), maintaining performance with colored noise, are mentioned. In [2, 11], a generalized subspace filter (GSF) class is presented, providing two control parameters: The reconstruction rank controlling the main trade-off between noise suppression and distortion, and a second parameter providing continuous trade-off between minimum-distortion and Wiener-like solutions for any given rank.

The strength of this subspace processing is, that if the clean speech covariance matrix is known and rank deficient while that of the noisy speech is full-rank, the noise content of the noise-only subspace dimensions can be removed without *any* change to the speech signal by choosing the correct reconstruction rank. However, any reconstruction rank between 1 and full-rank is feasible.

In this paper we study the performance of a state-of-the-art filter algorithm (the GSF) in terms of both noise, distortion, speech quality and speech intelligibility for the cocktail party situation. In order to keep this investigation separated from the problems of noise estimation, "Oracle" noise covariance estimates based on short-time analyses of the *pure noise signal* is used. Thus the results can be considered the upper bound of the method's capability. We have limited our study to single-channel, time-domain filtering.

The remainder of this paper is organised as follows. In Section 2 we present the GSF algorithm [2], estimation method and performance metrics. In Section 3 the experimental conditions and setup are given. Results are presented and discussed in Section 4 and Section 5 concludes the paper.

2. ENHANCEMENT METHOD

We consider a real, noisy signal vector $\mathbf{y} = \mathbf{x} + \mathbf{v}$ of length M, where $\mathbf{x}$ is clean speech and $\mathbf{v}$ is noise. Both are assumed quasi-stationary. Vectors $\mathbf{x}$ and $\mathbf{v}$ are assumed uncorrelated, so their covariance matrices add up to that of the noisy signal,

$$\mathbf{\Phi_y} = E(\mathbf{yy^H}) = \mathbf{\Phi_x} + \mathbf{\Phi_v}, \tag{1}$$

where E is the expectation value. $\mathbf{\Phi_v}$ is assumed to be rank M, while $\mathbf{\Phi_x}$ is of rank $P \leq M$.

The speech enhancement objective now is to estimate $\mathbf{x}$ or a known linear transformation of it from $\mathbf{y}$ and $\mathbf{\Phi_v}$, and set up an $M \times M$ filtering matrix $\mathbf{H}$ that will produce the desired speech estimate when applied to $\mathbf{y}$:

$$\mathbf{y_f} = \mathbf{Hy} \approx \mathbf{x}. \tag{2}$$

Provided that $\mathbf{\Phi_v}$ is full-rank and $\mathbf{\Phi_x}$ is positive definite, the joint diagonalization of $\mathbf{\Phi_x}$ and $\mathbf{\Phi_v}$ can be calculated, producing a matrix $\mathbf{B}$ of column eigenvectors $\mathbf{b_i}$ and a diagonal matrix $\mathbf{\Lambda}$ of corresponding real, positive, descending eigenvalues λ_i, $i \in [1, 2, ..., M]$ with the following properties:

$$\mathbf{B^H \Phi_x B} = \mathbf{\Lambda} \text{ and } \mathbf{B^H \Phi_v B} = \mathbf{I}. \tag{3}$$

This is also an eigendecomposition of matrix $\mathbf{\Phi_v^{-1} \Phi_x}$,

$$\mathbf{\Phi_v^{-1} \Phi_x B} = \mathbf{B\Lambda}. \tag{4}$$

From this a number of interesting filter matrices may be readily produced [2] by summing up contributions from decreasingly important eigenpairs up to the desired reconstruction rank:

Maximum SNR (rank 1):
$$\mathbf{H} = \mathbf{\Phi_v b_1 b_1^H}. \tag{5}$$

Classical Wiener:
$$\mathbf{H} = \mathbf{\Phi_v} \sum_{i=1}^{M} \frac{\lambda_i}{1+\lambda_i} \mathbf{b_i b_i}^H. \tag{6}$$

Subspace Wiener:
$$\mathbf{H} = \mathbf{\Phi_v} \sum_{i=1}^{P} \frac{\lambda_i}{1+\lambda_i} \mathbf{b_i b_i}^H. \tag{7}$$

MVDR:
$$\mathbf{H} = \mathbf{\Phi_v} \sum_{i=1}^{P} \mathbf{b_i b_i^H}. \tag{8}$$

Minimum Distortion:
$$\mathbf{H} = \mathbf{\Phi_v} \sum_{i=1}^{Q} \mathbf{b_i b_i^H}, \tag{9}$$

where $1 \leq Q \leq P$.

Generalized (GSF):
$$\mathbf{H} = \mathbf{\Phi_v} \sum_{i=1}^{Q} \frac{\lambda_i}{\mu+\lambda_i} \mathbf{b_i b_i}^H, \tag{10}$$

where $1 \leq Q \leq M$ and $0 \leq \mu \leq 1$.

Note how the GSF collapses to Wiener and Minimum Distortion/MVDR types at μ-values 1 and 0, respectively. We shall focus on this GSF (10), which takes two parameters $Q \in [1, 2, .., M]$ and $\mu \in [0..1]$.

Noise- (SNR) and distortion (SDR) metrics are based on filtering the clean speech and noise signals with the GSF, yielding $\mathbf{x_f}$ and $\mathbf{v_f}$. Thus these block metrics in dB become:

$$\mathrm{oSNR_{dB}} = 20 \log \frac{\|\mathbf{Hx}\|}{\|\mathbf{Hv}\|} = 20 \log \frac{\|\mathbf{x_f}\|}{\|\mathbf{v_f}\|}, \tag{11}$$

$$\mathrm{SDR_{dB}} = 20 \log \frac{\|\mathbf{Hx}\|}{\|\mathbf{Hx} - \mathbf{x}\|} = 20 \log \frac{\|\mathbf{x_f}\|}{\|\mathbf{x_f} - \mathbf{x}\|}. \tag{12}$$

However, the algorithm is evaluated on a whole-file basis:

$$\mathrm{oSNR_{dB}} = 10 \log \frac{\sum_{i=1}^{L}(x_{f,i}^2)}{\sum_{i=1}^{L}(v_{f,i}^2)}, \tag{13}$$

$$\mathrm{SDR_{dB}} = 10 \log \frac{\sum_{i=1}^{L}(x_{f,i}^2)}{\sum_{i=1}^{L}((x_{f,i} - x_i)^2)}, \tag{14}$$

where L is the file length in samples.

Speech Quality is measured by the PESQ method [12, 13] implemented in the MATLAB program `pesq.m` [4]. Based on a clean and a degraded version of the same speech file, this produces an estimate $\mathrm{mos_{PESQ}}$ of the Mean-Opinion-Score of a perceptual evaluation of speech quality rating on a 0-5 scale.

The Short-Time Objective Intelligibility (STOI) measure is done by the MATLAB program `taal2011.m` from the Auditory Modeling Toolbox [14] based on Taal et.al. [15]. This also takes a clean and a degraded version of the speech signal and returns an intelligibility metric on a 0-1 scale.

3. EXPERIMENTAL DETAILS

The purpose of the experiment was to map the above mentioned performance metrics of the GSF as functions of (Q, μ) revealing how that parameter space provides useful trade-offs between performance metrics. This was repeated with different values of input SNR (iSNR).

The clean speech signal was taken from the NOIZEUS corpus [16]. The babble noise (10 interfering speakers) was created by mixing five english male and female speakers from the EU-ROM corpus [17].

Experiment 1 used a concatenation of all 30 NOIZEUS tracks as test signal, performing filtering and performance measurements, and saving audio input/output files for all combinations of the parameters in Table 1.

Experiment 2 chose two fixed points in (Q, μ)-space and measured performance metrics for each of these for *each* individual NOIZEUS track as functions of iSNR. With 30 track results at each chosen parameter point, the mean value and 95% confidence interval of each performance metric was assessed.

Table 1. Parameters and signal conditions for Experiment 1

Parameter	Value Set
iSNR(dB)	-15,-12,-9,-6,-3,0,3,6
Q	1,2,4,8,16,32,64,128,160
μ	0,0.05,0.1,0.25,0.5,0.75,0.9,0.95,1

Test signals were scaled to the desired iSNR on a whole-file basis with clean speech levels set to 1.

The length-M signal vectors were read with 50% overlap from length-L ($L \gg M$) audio files, and the filtered results were Hanning-windowed and overlap-added to form a length-L audio output.

Sample covariance matrices were taken time-symmetrically around the current input vector over $2M+1$ vector samples, ensuring full-rank noisy matrices, i.e. for signal vector $\mathbf{y}$ starting at sample n_0:

$$\mathbf{\Phi_y} = \frac{1}{2M+1}\sum_{i=0}^{2M}\tilde{\mathbf{y}}_\mathbf{i}\tilde{\mathbf{y}}_\mathbf{i}^\mathbf{H}, \quad (15)$$

where $\tilde{\mathbf{y}}_\mathbf{i} = [y_{n_0+i-M}, y_{n_0+i-M+1}, .., y_{n_0+i-1}]^H$.

Similarly for $\mathbf{\Phi_v}$; and finally $\mathbf{\Phi_x}$ was estimated as

$$\mathbf{\Phi_x} = \mathbf{\Phi_y} - \mathbf{\Phi_v}. \quad (16)$$

The estimated $\mathbf{\Phi_x}$ (16) is rarely positive definit, producing some *negative* $\lambda_i s$. Therefore the summation in (10) should be stopped at the *last positive* λ_i, see (17,18), lest the algorithm become unstable when $\mu + \lambda_i \approx 0$. Thus, the practical filter is formed as

$$\mathbf{H}(Q,\mu) = \mathbf{\Phi_v}\sum_{i=1}^{R}\frac{\lambda_i}{\mu+\lambda_i}\mathbf{b_i}\mathbf{b_i}^H, \text{where} \quad (17)$$

$$R = \min[\mathrm{Q}, \max(\{\mathrm{i} \in \mathrm{N} | \lambda_\mathrm{i} > 0\})]. \quad (18)$$

This *dynamic rank limiting* (DRL), born out of the practical need for stability, provides an estimate of signal rank P and modifies e.g. a Classical Wiener solution, $(Q,\mu) = (M,1)$, see (6), to a subspace Wiener implementation (7).

The sampling rate was 8 kHz, and the block size M was 160, corresponding to a block duration of 20 ms. All signals were highpass filtered at 80 Hz (4^{th}-order Butterworth) before scaling in order to ensure overall zero-mean signals and avoid irrelevant low-frequency content with period times longer than the block size M.

Everything was done in MATLAB.

4. RESULTS

Experiment 1 results selected uniformly with regard to iSNR are shown in Figs. 1-4. Note the non-uniform parameter axes. Surface plots have been spline-interpolated for increased resolution, and the peak locations marked with red dots. In perfect accordance with theory, performance metrics iSNR and SDR peak in opposite corners of the (Q,μ) plane. And *both* the perceptual metrics peak in the "Wiener-corner" near (160,1). Also note the break-even thresholds drawn in red, beyond which filtering is *deteriorating* signals. Contrary to the case for oSNR, a GSF with the wrong choice of parameters may do more harm than good for *perceptual* quality. All increase distortion, because this was by definition absent from the start.

It has been reported that it is hard to improve both speech quality and intelligibility with the same algorithm [18]. That is not the case here. This could be due to the use of "Oracle" noise estimation, but is nonetheless of interest.

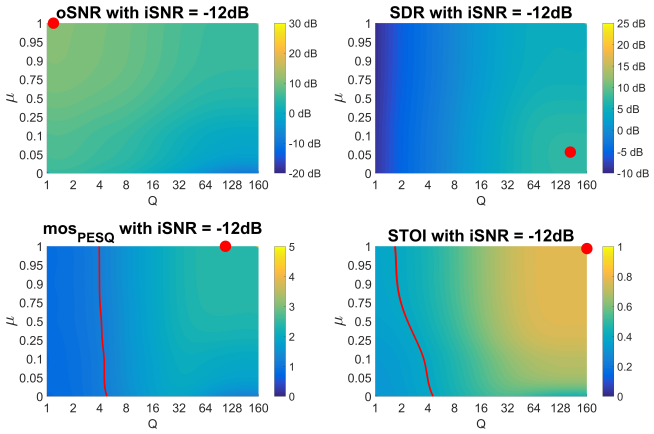

Fig. 1. Performance surfaces for iSNR = -12 dB

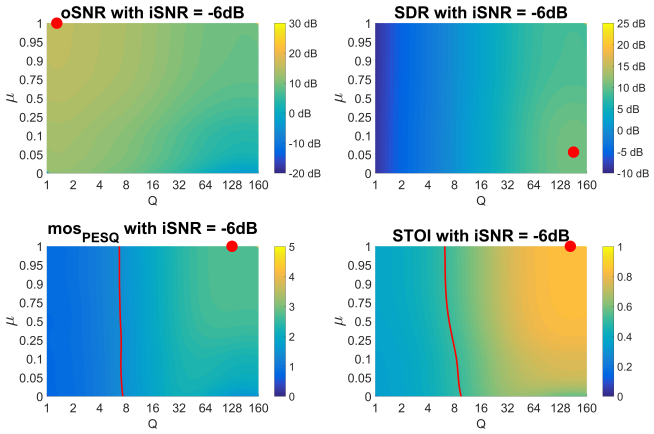

Fig. 2. Performance surfaces for iSNR = -6 dB

Experiment 2 maps performance as function of iSNR for two fixed (Q,μ) points, Point 1 at the perceptually optimal (160,1) and Point 2 halfway towards the SNR optimum corner, at (13,1). Results are shown in Fig. 5.

Clearly, the "Wiener-corner" Point 1 filter produces sig-

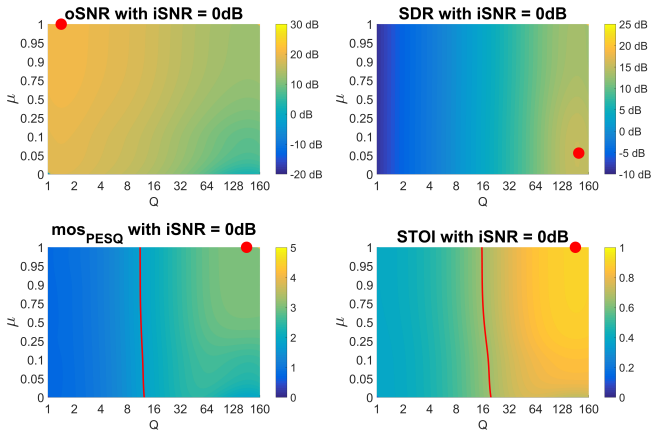

Fig. 3. Performance surfaces for iSNR = 0 dB

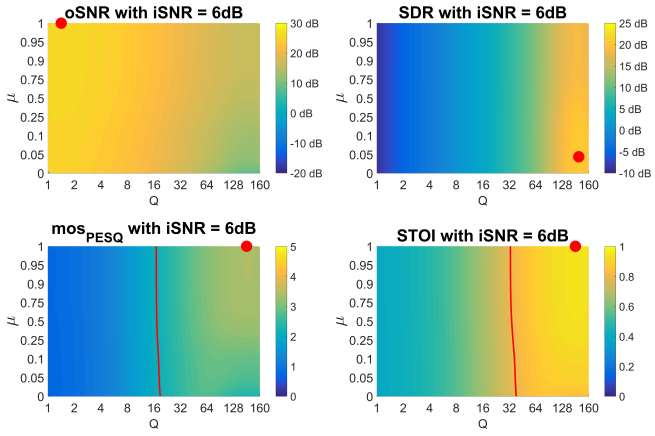

Fig. 4. Performance surfaces for iSNR = 6 dB

nificant improvements in both quality and intelligibility ratings. It also improves the SNR range from -15 to 6 dB to 5 to 17 dB. Moving the parameter choice to Point 2, in the direction of better oSNR, deteriorates both quality, intelligibility and distortion performance with only a moderate improvement in oSNR. Informal listening tests confirm that Point 1 is preferable to Point 2, which - due to the low reconstruction rank - sounds synthetic, even at high SNR.

As mentioned earlier, the dynamic rank limiting (17,18) effectively turns a classical Wiener solution (6) into a subspace Wiener solution (7) which is Wiener-optimal to the extent that the simple signal rank estimation (18) is correct. So how much performance is gained from all this subspace complexity, compared to a classical, much cheaper Wiener solution? As shown in Fig. 6, quite a lot is gained on oSNR and SDR, while the perceptual performance metrics are essentially unchanged.

The results presented in this section show that subspace filters are capable of very significant improvements on single-channel speech with very severe amounts of babble noise, *provided* that the short-term noise statistics is well known.

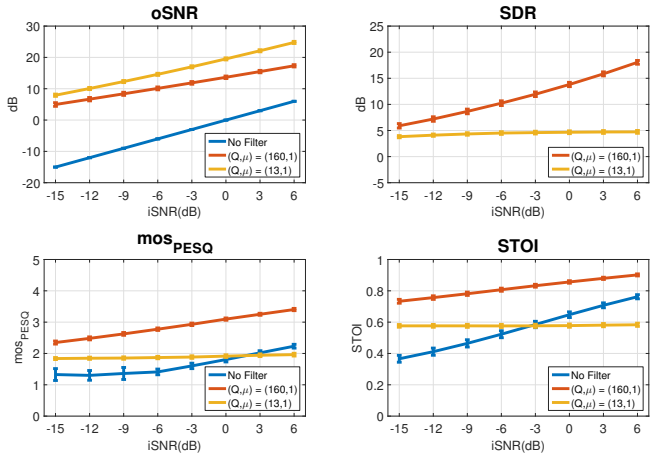

Fig. 5. Mean performance metrics as functions of iSNR with 95% confidence intervals.

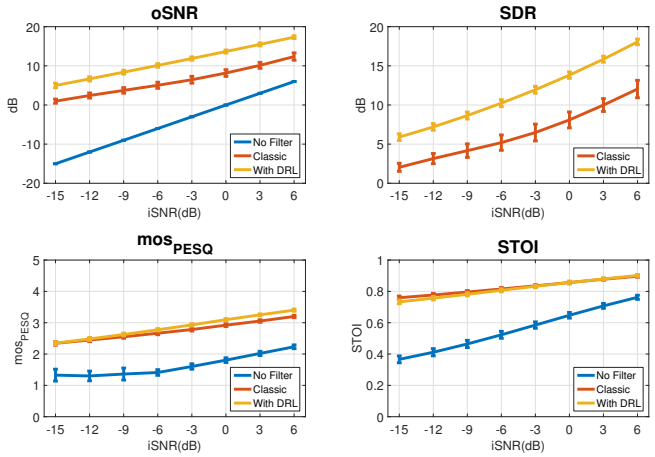

Fig. 6. Mean performance gains from subspace Wiener filter with DRL versus classical Wiener solution

How well this can be estimated in real life and *how* sensitive the filter performance is to estimation errors could be good topics of further investigation.

5. CONCLUSION

We have presented an experimental study of generalized subspace filters applied to single-channel speech enhancement in cocktail party situations with very poor SNR.

It has been shown that the full-rank Wiener subspace filter with Dynamic Rank Limiting can produce large improvements of both noise, speech quality and intelligibility in such cases, significantly out-performing the classical Wiener filter with respect to noise and distortion.

Investigating the filter's sensitivity to noise estimation errors and discovering good noise estimators for severe babble noise conditions are two natural extensions of this work.

6. REFERENCES

[1] T. Long, J. Chen, J. Benesty, and Z. Zhang, "Single-channel noise reduction using optimal rectangular filtering matrices," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1090–1090–1101, February 2013.

[2] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech Enhancement - A Signal Subspace Perspective*, Elsevier Academic Press, 2014.

[3] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, 2008.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2 edition, 17 April 2013.

[5] M. R. Schroeder, "US patent 3,180,936: Apparatus for suppressing noise and distortion in communication signals," 1965.

[6] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, Springer, 2009.

[7] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication, 10*, 1991.

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

[9] G. Carayannis and C. Gueguen, "The factorial linear modelling: A karhunen-loeve approach to speech analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing 1976, vol. 1, pp. 489-492*.

[10] S. Bakamidis, M. Dendrinos, and G. Carayannis, "SVN Analysis by Synthesis of Harmonic Signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 472–477, 1991.

[11] J. R. Jensen, J. Benesty, and M.G. Christensen, "Variable span filters for speech enhancement," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[12] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862.," .

[13] "Mapping function for transforming P.862 raw result scores to MOS-LQO, ITU-T Recommendation P. 862.1," .

[14] P. L. Soendergaard and P. Majdak, "Matlab program taal2011.m," http://amtoolbox.sourceforge.net/, 2014.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 7, pp. 2125 - 2136*, 2011.

[16] "http://ecs.utdallas.edu/loizou/speech/noizeus," .

[17] D. Chan et al., "EUROM - A Spoken Language Resource for the EU," *EUROSPEECH Proceedings, vol. 1, pp. 867-870*, 1995.

[18] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 47–56, Januar 2011.