COMPARISON OF COMPLEX-DFT ESTIMATORS WITH AND WITHOUT THE INDEPENDENCE ASSUMPTION OF REAL AND IMAGINARY PARTS

Richard C. Hendriks, Jan S. Erkelens and Richard Heusdens

Delft University of Technology,
2628 CD Delft, The Netherlands
{R.C.Hendriks, J.S.Erkelens, R.Heusdens}@TUDelft.nl

ABSTRACT

MMSE estimators for DFT-domain based single-microphone speech enhancement can broadly be classified in those that estimate the complex-DFT coefficients and those that estimate the DFT magnitudes. Existing complex-DFT MMSE estimators have generally been derived under assumptions that are in conflict with measured histograms and that are inconsistent with the assumptions made to derive DFT magnitude estimators. Recently it has been shown that these inconsistencies can be eliminated, i.e., no independency has to be assumed between real and imaginary parts of DFT coefficients if the phase of DFT coefficients is assumed uniformly distributed. In this paper we discuss the assumptions that underly the different complex-DFT estimators and show that the uniform phase assumption matches actual speech data. Furthermore, we show experimentally that the estimators without the independence assumption lead to a lower mean-square error.

Index Terms— Speech enhancement, complex-DFT estimators, independence assumption.

1. INTRODUCTION

Single-microphone DFT-based speech enhancement methods have attracted an increased interest for improving the quality of digital speech processing devices. These methods estimate complex-valued clean speech DFT coefficients by processing the noisy DFT coefficients on a frame-by-frame basis. The first methods that were proposed for single-microphone DFT-based speech enhancement were based on spectral subtraction, see, e.g., [1][2]. Later, somewhat more sophisticated methods were proposed, where minimum mean-square error (MMSE) estimators were derived by exploiting (assumed) densities of the speech and noise DFT coefficients, see, e.g., [3][4][5]. These statistical methods estimate either the complex clean speech DFT coefficients [4][5] or their magnitude [3][5].

In [5] it was shown that all known DFT-domain MMSE estimators can be derived as special cases under the generalized-Gamma speech prior density. This holds for both the class of complex-DFT estimators and for the class of magnitude estimators. The complex-DFT estimators were derived by assuming that real and imaginary parts of DFT coefficients are independent and are both distributed as a double-sided generalized-Gamma density. DFT Magnitude estimators were derived by assuming that the magnitude has a single-sided generalized-Gamma density and the phase has a uniform density. However, as already mentioned in [5], there do exist inconsistencies between the densities assumed in the cartesian domain and in the polar domain. More specifically, the independence assumption of real and imaginary parts is generally inconsistent with the assumption of a uniform phase in the polar domain. Furthermore, generalized-Gamma in the cartesian (polar) domain does not always correspond to generalized-Gamma in the polar (cartesian) domain. The uniform phase distribution seems to be the most realistic assumption; measured histograms show that phase is uniform and that real and imaginary parts of speech DFT coefficients are uncorrelated, but not independent [5][6].

In [7] a new theoretical framework is presented that eliminates the aforementioned inconsistencies in distributional assumptions by adopting the uniform phase distribution and dropping the independence assumption of real and imaginary parts. The basic assumptions used in this framework are that speech and noise are independent, that the speech phase is uniformly distributed and that the noise DFT coefficients are Gaussian distributed. In order to derive a complex-DFT estimator using this framework it is sufficient to specify only the density for the speech DFT magnitudes, i.e., it is not necessary to find an exact expression for the distribution of the complex DFT coefficients although this expression can in some cases be derived from the density of speech DFT magnitudes. The advantages are that it is not necessary to make the independence assumption of real and imaginary parts of DFT coefficients and that there is a direct transformational relation between the assumed density of the complex speech DFT coefficients and the density of the magnitude of DFT coefficients. In this paper we discuss the assumptions of independent real and imaginary parts of DFT coefficients on the one hand and uniform phase on the other hand. We show that the uniform phase assumption corresponds better to actual speech data. Further, we evaluate the new estimators with the conventional ones in a speech enhancement setting. The results show that with the new estimators a lower mean-square error (MSE) can be obtained.

2. NOTATION AND BASIC ASSUMPTIONS

We assume an additive noise model, i.e., $Y(k,i) = X(k,i) + N(k,i)$, where Y, X and N are the noisy speech, clean speech and noise DFT coefficient, respectively, and where k is the frequency index and i the time frame index. The DFT coefficients Y, X and N are assumed to be complex zero-mean random variables and X and N are assumed to be independent. Although all expressions in this paper are per time index i and frequency index k, we will leave out these indices for notational convenience.

For the random variables in question we use the following notation in the cartesian and polar domain:

$$Y = Y_R + jY_I,\ |Y| = Re^{j\Theta}, \tag{1}$$

The research was supported by MultimediaN.

$$X = X_R + jX_I,\ |X| = Ae^{j\Phi}, \tag{2}$$

$$N = N_R + jN_I, |N| = De^{j\Delta}, \tag{3}$$

where $j = \sqrt{-1}$ and where the subscripts R and I indicate the real and imaginary part of a DFT coefficient. We will use uppercase letters for random variables and the corresponding lowercase letters for their realizations. It is assumed that the noise DFT coefficients N have a Gaussian distribution, with independent and identically distributed real and imaginary parts with

$$\sigma_N^2 = \sigma_{N_R}^2 + \sigma_{N_I}^2, \text{ and } \sigma_{N_R}^2 = \sigma_{N_I}^2. \tag{4}$$

Further, the *a priori* SNR and the *a posteriori* SNR are defined as $\xi = \sigma_X^2/\sigma_N^2$ and $\zeta = |y|^2/\sigma_N^2$, respectively.

3. MMSE ESTIMATION OF COMPLEX-DFT COEFFICIENTS

3.1. Conventional Complex-DFT Estimators Assuming Independent Real and Imaginary Parts

Conventionally, complex-DFT estimators have been derived by assuming that the real and imaginary parts of speech DFT coefficients are statistically independent, e.g., [4][5]. This implies that the phase distribution of the complex DFT coefficients is in general not uniform. In Section 4 we will discuss the validity of the independence assumption. Under this assumption it follows that the MMSE estimator of clean speech DFT coefficients is given by [4]

$$E\{X|y\} = E\{X_R|y_R\} + jE\{X_I|y_I\}, \tag{5}$$

i.e., the MMSE estimator consists of the sum of the individual MMSE estimators of the real and imaginary part. The conditional expectation $E\{X_R|y_R\}$ is given by

$$E\{X_R|y_R\} = \frac{\int_{x_R} x_R f_{Y_R|x_R}(y_R|x_R) f_{X_R}(x_R) dx_R}{\int_{x_R} f_{Y_R|x_R}(y_R|x_R) f_{X_R}(x_R) dx_R}. \tag{6}$$

Here $f_{X_R}(x_R)$ is the assumed density of the real part of speech DFT coefficients, and the density $f_{Y_R|x_R}(y_R|x_R)$ is determined by the fact that the noise DFT coefficients are assumed Gaussian distributed and independent from the speech. This leads to

$$f_{Y_R|x_R}(y_R|x_R) = (2\pi\sigma_{N_R}^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_{N_R}^2}(y_R - x_R)^2\right).$$

Similar expressions can be given for the imaginary parts.

3.2. Complex-DFT Estimators Without Assuming Independent Real and Imaginary Parts

One of the main results in [7] is that complex-DFT estimators can be derived without assuming independence between real and imaginary parts of DFT coefficients. Instead, it is assumed that clean speech magnitude A and phase Φ are independent and that phase is uniformly distributed. Under these assumptions the complex-DFT estimator is given by [7]

$$E\{X|y\} = \frac{1}{r} \frac{\int_0^{+\infty}\int_0^{2\pi} a\cos(\phi-\theta) f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da}{\int_0^{+\infty}\int_0^{2\pi} f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da} y. \tag{7}$$

Here $f_A(a)$ is the assumed density of clean speech DFT magnitudes and $f_{Y|A,\Phi}(y|a,\phi)$ is determined by the fact that the noise DFT coefficients are assumed Gaussian distributed and independent from the speech, that is

$$f_{Y|A,\Phi}(y|a,\phi) = \frac{1}{\pi\sigma_N^2} \exp\left(-\frac{2ar\cos(\phi-\theta) - r^2 - a^2}{\sigma_N^2}\right).$$

An advantage is that only the magnitude distribution needs to be specified. Both magnitude and complex-DFT estimators can then be derived under one and the same statistical model. In other words, in Eq. (7) it is only necessary to specify $f_A(a)$, the transformation to $f_X(x)$ is done implicitly in Eq. (7). Moreover, the uniform phase assumption fits better to measured speech distributions as we will see in Section 4. It is interesting to compare the expression for $E\{X|y\}$ given in Eq. (7) with the general known expression for the MMSE magnitude estimator $E\{A|r\}$, i.e.,

$$E\{A|r\} = \frac{\int_0^{+\infty}\int_0^{2\pi} a f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da}{\int_0^{+\infty}\int_0^{2\pi} f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da}. \tag{8}$$

It then becomes clear that under the same densities the gain function for MMSE complex-DFT estimators is always smaller than the gain function for MMSE magnitude estimators, because of the factor $\cos(\phi - \theta)$ in the numerator of Eq. (7). In other words, MMSE complex-DFT estimators always apply more suppression than MMSE magnitude DFT estimators [7]. In Section 6 we will investigate the influence of the different assumptions on the enhancement performance.

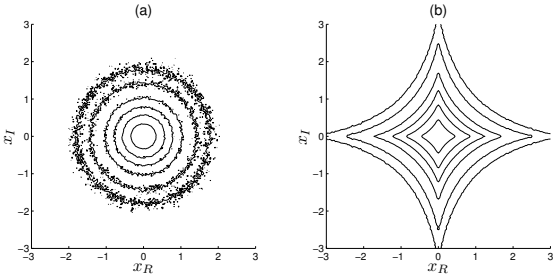

Fig. 1. Contour lines of measured distributions of real and imaginary parts normalized to unit variance [5]: (a) joint distribution; (b) product of marginal distributions.

4. VALIDITY OF DISTRIBUTIONAL ASSUMPTIONS

In order to get insight in the validity of the distributional assumptions that are made by using the conventional framework and by using the new framework, histograms were measured of real and imaginary parts of DFT coefficients. These histograms are shown in Fig. 1 and were measured in a similar way as in [4][6], i.e., only DFT coefficients have been taken into account for which the *a priori* SNR, estimated using the decision-directed [3] approach, was between 19 and 21 dB. The speech signals consisted of the complete TIMIT-TRAIN database, filtered at telephone bandwidth.

Fig. 1(a) shows the measured contours for the joint distribution. We see that the contour lines of the joint pdf of real an imaginary parts of the speech DFT coefficients are circular. A circularly symmetric joint pdf means that the real and imaginary parts are uncorrelated and that the phase distribution is uniform and independent from the magnitude distribution. Fig. 1(b) shows the contours for

the product of the marginal distributions. This plot is different from Fig. 1(a), and therefore, even though real and imaginary parts may be uncorrelated, they are clearly not independent. This validates the assumptions made in Section 3.2 and supports the theoretical framework presented in [7], while it contradicts the assumptions made in Section 3.1. In Section 6 we will show experimentally that the uniform phase assumption also leads to better speech enhancement performance.

5. COMPLEX-DFT ESTIMATORS UNDER THE GENERALIZED-GAMMA DENSITY

In this section we give analytical expressions for the estimators of Sections 3.1 and 3.2 for a general class of distributions, namely the generalized-Gamma densities. For details on the derivation of these estimators, see [5] and [7], respectively. For the estimator of Section 3.1 we assume that both X_R and X_I follow a double-sided generalized-Gamma density. That is, for $f_{X_R}(x_R)$ and $f_{X_I}(x_I)$ we assume

$$f_{X_R}(x_R) = \frac{\gamma {\beta'}^{\nu'}}{2\Gamma(\nu')}|x_R|^{\gamma\nu'-1}\exp(-\beta'|x_R|^{\gamma}), \quad \beta' > 0,\ \gamma > 0,\ \nu' > 0, -\infty < x_R < \infty, \tag{9}$$

where $\Gamma(\cdot)$ is the Gamma function. A similar equation holds for $f_{X_I}(x_I)$.

To derive the estimator in Eq. (7), Section 3.2, we assume that A follows a single-sided generalized-Gamma density, that is

$$f_A(a) = \frac{\gamma\beta^{\nu}}{\Gamma(\nu)}a^{\gamma\nu-1}\exp(-\beta a^{\gamma}), \quad \beta > 0,\ \gamma > 0,\ \nu > 0,\ a \geq 0. \tag{10}$$

We will only consider here the case $\gamma = 2$, since for this parameter setting it is possible to show that generalized-Gamma distributed real and imaginary parts lead to generalized-Gamma distributed magnitudes. This means that a direct experimental comparison is possible. This can be derived as follows: if it is assumed that X_R and X_I are independent, we have that $f_X(X) = f_{X_R}(x_R)f_{X_I}(x_I)$, subsequently $f_X(X)$ can be transformed to polar coordinates, i.e., $f_X(X) \Rightarrow f_{A,\Phi}(a,\phi)$. By integrating out the phase ϕ we obtain $\int_\phi f_{A,\Phi}(a,\phi)d\phi = f_A(a)$, i.e., the magnitude density from Eq. (10). The relation between ν' in Eq. (9) and ν in Eq. (10) is then given by $\nu = 2\nu'$. However, notice that by assuming independent X_R and X_I, the phase distribution will be in general not uniform, which contradicts the measurements in Section 4. For other common settings, e.g., $\gamma = 1$, the relation between Eq. (9) and Eq. (10) holds only approximately. However, the results from Section 3.2 remain valid for these settings and the estimator can be derived [7].

Using Eq. (9) with $\gamma = 2$ and [8, Eq. 3.462.1], the estimator in Eq. (6) can be written as

$$E\{X_R|y_R\} = \frac{2\nu'\sigma_{N_R}}{\sqrt{1+2\nu'\xi^{-1}}}\frac{D_{-(2\nu'+1)}(y_-) - D_{-(2\nu'+1)}(-y_-)}{D_{-2\nu'}(y_-) + D_{-2\nu'}(-y_-)}, \tag{11}$$

where $D_\nu(\cdot)$ is the parabolic cylinder function of order ν [9, Ch. 19] and

$$y_- = -\frac{y_R}{\sigma_{N_R}}\left(1 + 2\nu'\xi^{-1}\right)^{-1/2}. \tag{12}$$

Use has been made of the relation $\beta' = 2\nu'/\sigma_X^2$ [5]. Again, a similar expression can be derived for the imaginary part of Eq. (5).

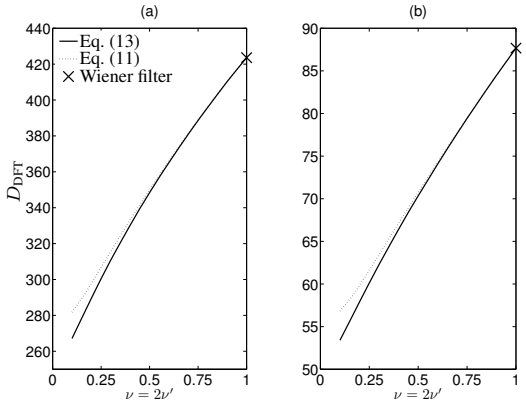

Fig. 2. Comparison between conventional (dotted line) and proposed framework (solid line) in terms of D_{DFT}. The $\times$ indicates the Wiener filter for both cases. The signals are degraded by white noise at an SNR of (a) 5 dB (b) 15 dB.

The estimator in Eq. (7) can be derived by inserting Eq. (10) with $\gamma = 2$ into Eq. (7), followed by using [8, Eqs. 8.431.5, 6.643.2, 9.210.1, and 9.220.2], leading to

$$E\{X|y\} = \frac{\nu\xi}{\nu+\xi}\frac{\mathcal{M}(\nu+1;2;\frac{\zeta\xi}{\nu+\xi})}{\mathcal{M}(\nu;1;\frac{\zeta\xi}{\nu+\xi})}y, \tag{13}$$

where $\mathcal{M}(\cdot)$ is the confluent hypergeometric function [9, Ch. 13] and where use is made of the relation $\beta = \nu/\sigma_X^2$ [5]. For $\nu = 1$, Eq. (13) leads to the Wiener filter. An interesting observation that can be made from Eq. (13) is that for $\nu \ll \xi$ the estimator becomes almost independent of ξ.

6. EXPERIMENTAL RESULTS

In Section 4 it was shown that the uniform phase assumption matches measured speech DFT histograms. In this section experimental results are presented to demonstrate the influence of these better assumptions on the speech enhancement performance.

The speech and noise signals that we used to generate the experimental results originate from the Noizeus [10] database. We extended this database with white noise. All signals are filtered at telephone bandwidth and sampled at 8 kHz. The noisy time domain signals are divided in frames of 256 samples with 50 % overlap. For both analysis and synthesis a square root Hann window is used. For estimation of the *a priori* SNR we use the decision-directed approach [3] and for estimation of the noise variance we use the DFT-subspace based method presented in [11].

As a first comparison we measure the performance of both complex-DFT estimators using the square-error distortion measure

$$D_{\mathrm{DFT}} = \sum_{(k,i)\in\mathcal{Q}} |x(k,i) - \hat{x}(k,i)|^2, \tag{14}$$

where $\mathcal{Q}$ is an index set that denotes the DFT bins with energy no less than 50 dB below the maximum bin energy in the particular speech signal. This is done to reduce the influence of noise-only regions on the experimental results.

Fig. 2 compares Eq. (11) and Eq. (13) in terms of D_{DFT} as a function of the ν-parameter. In these experiments ν ranges from $\nu = 0.1$ up to $\nu = 1$. Both estimators comprise the Wiener filter as a special

case. This is at $\nu = 1$ (indicated by the symbol $\times$). The results are shown in Fig. 2(a) and Fig. 2(b) for speech signals degraded by white noise at an SNR of 5 and 15 dB, respectively. We see that at both SNRs, the distortion D_{DFT} is reduced using the complex-DFT estimators under the new framework. Both estimators perform best at small values of ν, where we also get the largest improvements with the new estimator of Eq. (13).

For further comparison of enhancement performance we follow a similar approach as in [6] and measure speech segmental SNR as

$$\text{SSNR}_{\text{seg}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} 10 \log_{10} \left(\frac{\|\mathbf{x}_t(i)\|_2^2}{\|\mathbf{x}_t(i) - \tilde{\mathbf{x}}_t(i)\|_2^2} \right), \qquad (15)$$

where $\tilde{\mathbf{x}}_t(i)$ is a frame of the time domain signal that is the result of applying the gain functions to the clean speech frame. To discard non-speech frames, an index set $\mathcal{P}$ is used of all clean speech frames with energy within 50 dB of the maximum frame energy in a particular speech signal. $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$. Similarly, noise segmental SNR is measured as

$$\text{NSNR}_{\text{seg}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} 10 \log_{10} \left(\frac{\|\mathbf{n}_t(i)\|_2^2}{\|\tilde{\mathbf{n}}_t(i)\|_2^2} \right), \qquad (16)$$

where $\mathbf{n}_t(i)$ is a noise frame, and $\tilde{\mathbf{n}}_t(i)$ is the residual noise frame resulting from applying the noise suppression filter to the noise only.

Fig. 3 shows a comparison between the estimators in Eq. (11) and Eq. (13) in terms of NSNR_{seg} vs. SSNR_{seg} as a function of the ν-parameter. In addition, the NSNR_{seg} vs. SSNR_{seg} trade-off is shown for the MMSE magnitude estimators that can be derived from Eq. (8) with Eq. (10), see [5] for details on the derivation of these estimators. The comparison is shown for speech signals degraded by white noise and street noise at 5 dB and 15 dB SNR. The ν-values in Fig. 3 range from $\nu = 0.1$ (indicated by the symbol $*$) to $\nu = 2$.

We see that estimators under the new framework, with assumptions in line with measurements as discussed in Section 4, lead to improved speech quality in terms of SSNR, but a slightly lower NSNR.

Further, it is shown that the magnitude estimators in general lead to a better speech quality compared to the complex-DFT estimators, but that they lead to less noise reduction as was mentioned in Section 3.2 below Eq. (8).

7. CONCLUDING REMARKS

Existing complex-DFT estimators have generally been derived under the assumption that real and imaginary parts of DFT coefficients are independent. This assumption is in conflict with measured histograms of actual speech data and is also inconsistent with the assumption that the phase of DFT coefficients in the polar domain is uniform. Recently, a framework has been proposed that eliminates these inconsistencies. The assumptions made under this new framework match better with actual speech data. We have shown that estimators derived using this new framework lead to improved speech enhancement performance in terms of mean-square error.

From the website http://ict.ewi.tudelft.nl/%7Erichard a MATLAB toolbox can be downloaded to compute and tabulate gain functions for $E\{X|y\}$ and $E\{A|r\}$ under the assumption that $f_A(a)$ has a density as in Eq. (10) with $\gamma = 1$ or $\gamma = 2$.

8. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

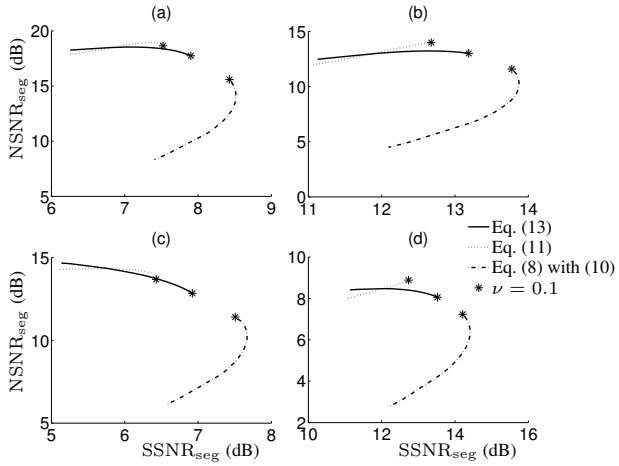

Fig. 3. Performance in terms of NSNR_{seg} vs. SSNR_{seg}. The ν-values range from $\nu = 0.1$ up to $\nu = 2$. The lowest ν-value is indicated by the $*$. Signals are degraded by (a) white noise at 5 dB SNR (b) white noise at 15 dB SNR, (c) street noise at 5 dB SNR (d) street noise at 15 dB SNR.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[5] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1741 – 1752, August 2007.

[6] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, May 2005.

[7] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *Accepted for publication in IEEE Signal Processing Letters*, 2008.

[8] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*, New York: Academic, 6th ed. edition, 2000.

[9] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New-York, ninth dover printing, tenth gpo printing edition, 1964.

[10] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2006, vol. 1, pp. 153–156.

[11] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio Speech and Language Processing*, March 2008.