INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Single-Channel Speech Enhancement Using Double Spectrum

Martin Blass[†], *Pejman Mowlaee*[†] , *W. Bastiaan Kleijn*[‡]

[†]Signal Processing and Speech Communication Lab, Graz University of Technology
[‡]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
mblass@student.tugraz.at pejman.mowlaee@tugraz.at bastiaan.kleijn@ecs.vuw.ac.nz

Abstract

Single-channel speech enhancement is often formulated in the *Short-Time Fourier Transform* (STFT) domain. As an alternative, several previous studies have reported advantages of speech processing using pitch-synchronous analysis and filtering in the modulation transform domain. We propose to use the *Double Spectrum* (DS) obtained by combining pitch-synchronous transform followed by modulation transform. The linearity and sparseness properties of DS domain are beneficial for single-channel speech enhancement. The effectiveness of the proposed DS-based speech enhancement is demonstrated by comparing it with STFT-based and modulation-based benchmarks. In contrast to the benchmark methods, the proposed method does not exploit any statistical information nor does it use temporal smoothing. The proposed method leads to an improvement of 0.3 PESQ on average for babble noise.
Index Terms: speech enhancement, double spectrum, modulation transform, pitch-synchronous analysis

1. Introduction

In various speech processing applications including speech coding, automatic speech recognition and speech synthesis the underlying signal representation determines the accuracy and efficiency of a certain algorithm. Good representations often require relatively few coefficients per unit time for an accurate description of the speech signal, but are complete and hence able to describe any signal. We argue that the *Short-Time Fourier Transform* (STFT), the predominant choice in speech enhancement (see e.g. [1] for an overview), while complete, generally does not lead to a sparse signal representation for speech.

An alternative to the STFT domain is pitch-synchronous analysis, with successful results reported both for speech coding [2, 3] and speech enhancement [4]. It was shown that frame theory can be used to understand this representation [3].

Another alternative is to process speech in the *Short-Time Modulation* (STM) domain. Speech enhancement proposals in modulation domain are spectral subtraction [5], *Minimum Mean Square Error* (MMSE) of *Short-Time Modulation Magnitude* (STMM) Spectrum [6], MMSE speech enhancement using real and imaginary parts of STM [7]. These STM-based methods, compared to their STFT counterparts, showed less musical noise or spectral distortion with improved perceived quality.

Inspired by the advantages of modulation and pitch-synchronous transforms, a key research question is then how to exploit these in a speech enhancement framework. In this paper, therefore, we propose *Double Spectrum* (DS) signal representation consisting of pitch-synchronous and modulation transforms. We propose single-channel speech enhancement in DS domain. To demonstrate the potentials and advantages of the proposed method, we compare its performance versus the previous STFT-based and modulation-based benchmarks.

The remainder of the paper is organized as follows; Section 2 places our work in the context of earlier work. In Section 3 we provide fundamentals of the *Double Spectrum* (DS) approach. Section 4 presents the proposed DS speech enhancement, Section 5 shows the results and Section 6 provides conclusions.

2. Relation to Previous Works

Separating slowly varying and rapidly varying pitch-cycle waveform components formed the basis of *Waveform Interpolation* (WI), which resulted in high quality speech coding [2]. A more general pitch-synchronous modulation representation was introduced in [3]. This two-stage transform representation was further refined by Nilsson et al. [8]. The two-stage transform led to a solid performance in speech coding and prosodic modification. In such speech representation the fundamental frequency is the key feature resulting in a sparse speech-signal representation. The block diagram for the two-stage transform representation, shown in Figure 1, consists of four processing blocks: *Linear Prediction* (LP) analysis, constant pitch warping, pitch-synchronous transform and modulation transform.

The two-stage transform, consisting of pitch-synchronous and modulation transforms exploits the features of the warped residual to achieve a highly energy concentrated representation and will be described in more detail in Section 3.2. The combination of pitch-synchronous and modulation transform results in lapped frequency transforms, which approximates the *Karhunen-Loève Transform* (KLT) for stationary signal segments [9]. The KLT maximizes the coding gain, which can be seen as a particular form of energy concentration [8].

The two-stage transform was extended to speech enhancement [4], where its ability to separate periodic and aperiodic signals were exploited to improve speech quality. Noise reduction was achieved by adaptive weighting of the coefficients in different modulation bands, which restored harmonicity of noise corrupted speech. The method was capable of separating the speech signal into voiced and unvoiced components using a best-basis selection that optimized the energy concentration of the transform coefficients.

Throughout this paper, the signal representation obtained by two-stage transform (pitch-synchronous and modulation transform) will be referred to as *Double Spectrum* (DS). Figure 1 shows the DS framework highlighted in a light gray block as the basis of the proposed speech enhancement system. Our

The work was supported by Austrian Science Fund (P28070-N33). The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFW, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and Vienna Business Agency. The programme COMET is conducted by Austrian Research Promotion Agency (FFG)

http://dx.doi.org/10.21437/Interspeech.2016-234

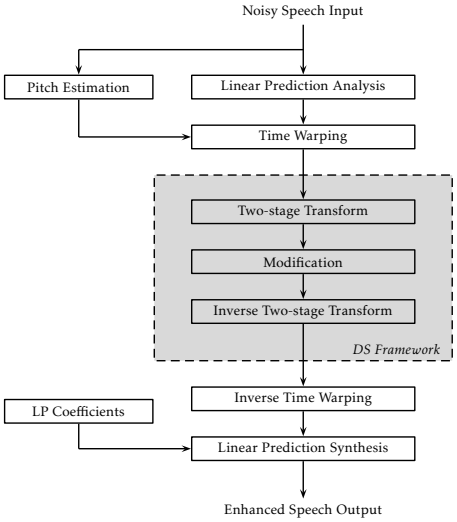

Figure 1: Block diagram for a canonical speech representation system [8]. The highlighted block shows DS framework using a two-stage transform and signal modification in DS domain.

goal is to find a framework where the two-stage transform is directly applied on the noisy signal. In contrast to [4, 8], our method relies on fixed analysis time blocks (no LP analysis, nor time warping), which makes the method simpler and faster.

3. Double Spectrum: Fundamentals

First, the pitch is extracted and stored within the coefficients of the two-stage transform. Since pitch is time-varying and both transforms do not adapt to this property, we introduce block processing under the assumption of quasi-stationarity of speech, explained in the following.

3.1. Time Block Segmentation

Given a fundamental frequency f_0, the first step in calculating DS is pitch-synchronous *Time Block Segmentation* (TBS). The TBS step separates the input speech into L time blocks of variable length. The length of each time block is an integer multiple of $P_0 = f_s/f_0$, where f_s is the sampling frequency and P_0 is the fundamental period in samples. A time block is further subdivided into $\mathcal{L}$ frames, each of length P_0. To avoid discontinuities at the transition of consecutive blocks overlapping is introduced.

3.2. Two-stage Transform

Each time block is analyzed in terms of a two-stage transform. The pitch-synchronous transform is implemented as a *Modulated Lapped Transform* (MLT) [9]. Since pitch varies over time, this means that we ignore its local variation of pitch during TBS. The MLT is implemented using a DCT-IV in combination with square-root Hann window following [8]. This facilitates a critically sampled uniform filter bank with coefficients that are localized in time and frequency. The usage of a square-root window at analysis and synthesis stage as a matched filter satisfies the power complementarity constraint needed for perfect reconstruction.

Let $\nu = 0, 1, \ldots, 2P_0 - 1$ be a time index and let $x_\ell(\nu)$ be the ℓ'th pitch-synchronous time frame, i.e. $x_\ell(\nu) = x(\ell P_0 + \nu)$. The first-stage transform coefficients $f(\ell, k)$ are then obtained as

$$f(\ell,k) = \sum_{\nu=0}^{2P_0-1} \tilde{x}_\ell(\nu)\sqrt{\frac{2}{P_0}}\cos\left(\frac{(2k+1)(2\nu - P_0 + 1)\pi}{4P_0}\right), \quad (1)$$

where $\ell = 0, 1, \ldots, \mathcal{L}-1$ and $k = 0, 1, \ldots, P_0 - 1$ denote time frame index and frequency band index, respectively, and $\tilde{x}_\ell(\nu) = x_\ell(\nu)w(\nu)$ as the windowed signal segment.

The output of the first transform is a sequence of MLT coefficients that evolve slowly over time for voiced speech but rapidly for unvoiced speech. Note that due to the pitch-synchronous nature of the time frames, the cardinality of the frequency bands is $K = P_0$.

The modulation transform is a DCT applied to a number of consecutive frames of the frequency coefficients obtained from pitch-synchronous transform [10]. To facilitate the implementation of the modulation transform as a critically sampled filter, we use DCT-II yielding the coefficients $g(q, k)$ given by

$$g(q,k) = \sum_{\ell=0}^{Q-1} f(\ell,k)c(q)\sqrt{\frac{2}{Q}}\cos\left(\frac{(2k+1)q\pi}{2Q}\right), \quad (2)$$

where $q = 0, 1, \ldots, Q-1$ is the modulation band index, $c(0) = 1/\sqrt{2}$ and $c(q) = 1$ for $q \neq 0$. The definition for *Double Spectrum* is now given by $DS(q, k)$, which is equivalent to $g(q, k)$ interpreted as a matrix with K frequency bands as rows and Q modulation bands as columns. Figure 2 schematically visualizes a speech signal in terms of a sequence of Double Spectra, showing $DS^{(l)}(q, k)$ for a set of time blocks $l \in [0, L-1]$.

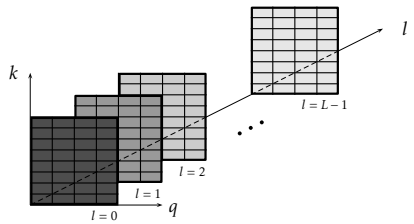

Figure 2: Illustration of a speech signal in Double Spectrum $DS^{(l)}(q, k)$ shown for time blocks $l = 0, 1, \ldots, L-1$.

3.3. Some Useful Properties of Double Spectrum

The useful properties of Double Spectrum are: sparsity, linearity, real-valued coefficients, and facilitates comb filtering.

3.3.1. Property I: Sparsity

For a periodic signal segment $DS(q, k)$ yields a high energy concentration at low modulation bands for frequency channels related to multiples of f_0. In particular, the first modulation band $q = 0$ represents the periodic component of a signal, whereas the other modulation bands describe the aperiodic parts. This property can be explained by assuming a strictly periodic time signal, e.g., a pure sinusoid. Applying the pitch-synchronous transform yields MLT coefficients that are identical for consecutive frames. The subsequent modulation transform is hence applied to a constant data sequence, yielding only one non-zero coefficient for $q = 0$, which can be understood as the DC component of the DCT-II transform. This property may be exploited for voiced-unvoiced decomposition or for restoring the harmonicity of noise corrupted speech by finding an appropriate balance between low and high modulation bands [4].

3.3.2. Property II: Linearity

In the time domain, noisy signal $y(\nu)$ is a superposition of the clean signal $x(\nu)$ and the noise signal $d(\nu)$. In the DS domain this superposition is preserved, since DS is a linear operator:

$$y(\nu) = x(\nu) + d(\nu) \circ\!\!-\!\!\bullet\; DS_y = DS_x + DS_d, \quad (3)$$

where DS_y, DS_x and DS_d denote the DS representation of noisy, clean and noise signal, respectively. Figure 3 shows an example for DS_y, DS_x and DS_d of the same voiced speech segment to illustrate *linearity*.

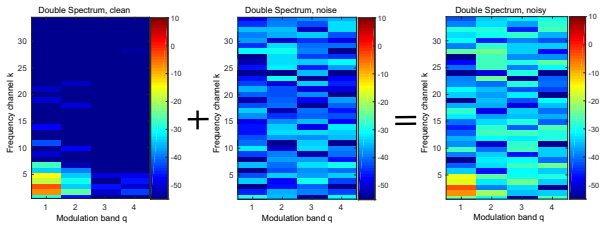

Figure 3: Linearity of DS operator given in (3): (Left) clean, (Middle) noise and (Right) noisy DS.

3.3.3. Property III: Real-Valued Coefficients

The coefficients of $DS(q,k)$ are real-valued and symmetrically distributed around zero as mean value.

3.3.4. Property IV: Facilitates Comb Filtering

Another property is the pitch-synchronous filter bank which allows comb filtering. Since an analysis frame of length of $2P_0$ yields $K = P_0$ frequency bands, $k_{f_0} = 2$ denotes the frequency band corresponding to f_0 and we have:

$$k_{f_0} = \frac{2K}{f_s} f_0. \quad (4)$$

4. Speech Enhancement in DS Domain

In this Section we present the essential tools for speech enhancement in DS domain comprised of pitch estimation, speech presence probability estimation, and the DS weighting function.

4.1. Pitch Estimation

The segmentation used in DS requires a fundamental frequency estimate. If the time blocks are segmented erroneously due to errors in pitch estimation, then the energy of periodic speech segments is no longer concentrated in the low modulation bands, but leaks into higher bands. We propose an f_0-estimator that relies on a periodicity measure calculated in the DS domain, called the *Modulation Band Ratio* (MBR). The MBR compares the summed energy of the first modulation band E_1 to the total energy $E_{1:Q}$

$$\text{MBR}(K) = \frac{E_1}{E_{1:Q}} = \frac{E_1}{E_1 + E_{2:Q}}, \quad (5)$$

where $E_1 = \sum_{k=0}^{K-1} |DS(0,k)|^2$ and $E_{1:Q} = \sum_{q=0}^{Q-1} \sum_{k=0}^{K-1} |DS(q,k)|^2$. For periodic frames the MBR reaches values close to 1, while for non-periodic frames the mean MBR is $1/Q$ (close to 0). This allows us to derive an f_0-estimator by searching for an optimal frequency index K^* that maximizes the MBR:

$$K^* = \arg\max_K \text{MBR(K)}. \quad (6)$$

Using (4), the fundamental frequency estimate is $f_0^* = \frac{f_s}{K^*}$. Since using this f_0-estimator should serve as a proof of concept only, we skipped further evaluation steps.

4.2. Speech Presence Probability Estimation

Many common speech enhancement systems use information about the speech presence probability (SPP). In the design of our filter method we also take into account SPP to selectively modify regions of speech presence or absence. The SPP is computed in the DS domain using the MBR measure, which discriminates voiced and unvoiced speech even in heavy noise scenarios. MBR yields values close 1 for voiced and close to 0 for unvoiced, hence is a good measure for SPP.

4.3. Adaptive Weighting based on Energy Smoothing

Our proposed speech enhancement, referred to as *Double Spectrum Weighting* (DSW), is an adaptive weighting scheme corresponding to filtering in time domain. The weighting coefficients $G(q,k)$ are applied to the noisy coefficients $DS_y(q,k)$ and yield the clean speech estimate $\widehat{DS_x}(q,k)$:

$$\widehat{DS_x}(q,k) = G(q,k) DS_y(q,k), \quad (7)$$

where $G(q,k)$ is a cascade of two weighting schemes: $W_e(q,k)$ to dampen noise-dominant coefficients, and $W_q(q,k)$ to enhance harmonicity, each described in the following.

4.3.1. $W_e(q,k)$: Energy-based coefficient weighting

The first weighting, $W_e(q,k)$ is an energy based coefficient weighting $W_e(q,k)$ which compares the energy of each DS-coefficient with respect to the mean energy of $DS_y(q,k)$, resulting in the relative energy $E_{rel}(q,k)$ defined as

$$E_{rel}(q,k) = KQ \, \frac{|DS(q,k)|^2}{E_{1:Q}}. \quad (8)$$

Since E_{rel} shows a broad dynamic range, we apply the decadic logarithm as a non-linear mapping function. Additionally, we constrain the weights to non-negative numbers by adding 1 to E_{rel}:

$$W_e(q,k) = \log_{10}(E_{rel}(q,k) + 1). \quad (9)$$

Note that this coefficient compression is empirically chosen and motivated by works like [11, 12].

4.3.2. $W_q(q,k)$: Harmonicity Enhancement

As the second weighting, we propose $W_q(q,k)$ to enhance the harmonicity of noisy speech. To this end, we need a harmonicity indicator. Similar to (5), we consider the Modulation Band Ratio of the respective frequency band, MBR_k given by

$$\text{MBR}_k = \frac{|DS(0,k)|^2}{\sum_{q=0}^{Q-1} |DS(q,k)|^2}. \quad (10)$$

In contrast to the fixed-weighting in [4], we propose an exponentially decaying modulation weighting, motivated by statistical observations of voiced DS data. Therefore, we use

$$W_q(q,k) = e^{-\text{MBR}_k q}, \quad (11)$$

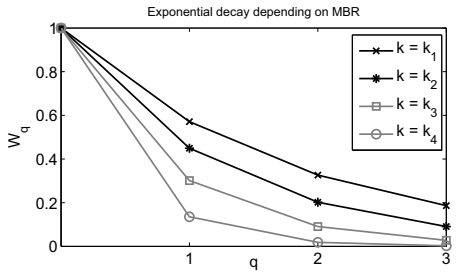

Figure 4: $W_q(q,k)$ as a function of q shown for different values of $k_1 = 2000$ Hz, $k_2 = 700$ Hz, $k_3 = 500$ Hz, $k_4 = 200$ Hz.

where MBR_k serves as the decay factor of the exponential weighting. Figure 4 exemplifies the exponential decaying characteristic in $W_q(q,k)$ for different frequency channels k and across all modulation bands q.

To have a selective noise suppression, similar to conventional DFT-based speech enhancement [1], we utilize DS-based SPP as described in 4.2 and apply it as a scaling factor on the cascade weighting outcome

$$G(q,k) = \text{SPP} \cdot W_e(q,k) W_q(q,k). \tag{12}$$

Finally, we restrict $G(q,k)$ to a lower limit $G_{\min} = 0.178 \triangleq -15$ dB [13] which yields

$$G(q,k) = G_{\min} \quad \text{if} \quad G(q,k) < G_{\min}. \tag{13}$$

Following (7) we apply these weighting coefficients on the noisy DS to obtain $\widehat{DS_x}$. To obtain the enhanced time signal inverse transforms are applied followed by an overlap-and-add routine.

5. Results

In this Section, we demonstrate the effectiveness of the proposed DS-based speech enhancement in a blind scenario and compare its performance versus the STFT-based and modulation-based benchmarks. To check the robustness of the method we provide results for f_0-known versus blind scenario.

5.1. Experimental Setup

Clean speech utterances were taken from Noizeus speech corpus [14] consisting of 30 phonetically-balanced sentences uttered by three males and three female speakers (average length of 2.6 seconds). The speech files were downsampled from the original sampling frequency of 25 kHz to 8 kHz to simulate telephony speech. To obtain noisy files, the clean speech was corrupted in babble noise mixed at SNRs of 0, 5 and 10 dB. As evaluation criteria, we chose *Perceptual Evaluation of Speech Quality* (PESQ) measure [15] and the *Short-Time Objective Intelligibility* (STOI) measure [16]. We report results in terms of improvement in ΔPESQ and ΔSTOI as comparison to the outcome from the noisy (unprocessed) input speech.

To demonstrate the effectiveness of the proposed method, we include three benchmarks: 1) *MMSE-STSA* [17], 2) *ModSpecSub* [5] referring to spectral subtraction in STM, as speech enhancement benchmark, and 3) we report results of fixed-weighting following specification in [4] without LP and time-warping stages. For *MMSE-STSA* a decision-directed scheme was used with a Minimum Statistics noise estimator [18] with a 16 ms frame shift, a 32 ms window length and a Hamming window. For *ModSpecSub* we used the implementation provided by Paliwal et al. [5].

SNR-level (dB)	0	5	10
MMSE-STSA [17]	0.18	0.20	0.22
ModSpecSub [5]	0.12	0.12	0.08
Fixed weighting [4]	0.17	0.19	0.17
DSW (blind)	0.27	0.34	0.30
DSW (f_0-known)	0.37	0.38	0.35

Table 1: ΔPESQ results averaged over SNRs and utterances shown for babble noise and different methods.

SNR-level (dB)	0	5	10
MMSE-STSA [17]	-0.01	0.00	0.00
ModSpecSub [5]	-0.04	-0.04	-0.05
Fixed weighting [4]	0.00	-0.01	-0.02
DSW (blind)	-0.03	-0.04	-0.07
DSW (f_0-known)	0.03	0.00	-0.04

Table 2: ΔSTOI results averaged over SNRs and utterances shown for babble noise and different methods.

The parameter setup used for the proposed DS-based speech enhancement is as follows. The length of the analysis window is $2P_0$ with 50% overlap, i.e., P_0 of the respective time block. Assuming stationarity for short time intervals [19] and taking a typical range for f_0 into account, we set the number of modulation bands to $Q = 4$.

5.2. Speech Enhancement Results

Tables 1 and 2 report the averaged results of ΔPESQ and ΔSTOI for 30 speakers. The following observations are made:

- The proposed method (*DSW*) leads to a 0.3 improvement in PESQ, outperforming both the *MMSE-STSA* [17] and *ModSpecSub* [5] benchmarks.
- Our pitch estimator performs well. Using an oracle f_0 leads to only a minor improvement in performance in PESQ and STOI. For some audio examples we refer to https://www2.spsc.tugraz.at/people/pmowlaee/DS.html.
- In terms of intelligibility, a fixed weighting similar to [4] results in a better STOI compared to the proposed method at the expense of a lower improvement in the perceived quality predicted by PESQ.

6. Conclusions

In this paper, we proposed *Double Spectrum* (DS) speech enhancement that relies on pitch-synchronous and modulation transforms. The linearity of the DS operator results in a sparse representation of speech that provides a means for the identification and separation of rapidly-varying (noise and unvoiced speech) versus slowly varying (voiced speech) component. These properties facilitate selective noise reduction. Our experiments confirm that DS-based speech enhancement outperforms its STFT and modulation-only counterparts.

The linear property of DS suggests the study of DS subtraction as a direction for future work on the DS noise estimator.

7. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013.

[2] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 1, no. 4, pp. 386–399, Oct 1993.

[3] ——, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2000, pp. 1475–1478.

[4] F. Huang, T. Lee, W. B. Kleijn, and Y.-Y. Kong, "A method of speech periodicity enhancement using transform-domain signal decomposition," *Elsevier speech communication*, vol. 67, pp. 102–112, 2015.

[5] K. K. Paliwal, K. Wjcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Elsevier speech communication*, vol. 52, no. 5, pp. 450 – 475, 2010.

[6] K. K. Paliwal, S. Belinda, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Elsevier speech communication*, vol. 54, no. 2, pp. 282–305, 2012.

[7] S. Belinda and K. K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Elsevier speech communication*, vol. 58, pp. 49–68, 2014.

[8] M. Nilsson, B. Resch, M. Y. Kim, and W. B. Kleijn, "A canonical representation of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 849–852, 2007.

[9] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 38, no. 6, pp. 969–978, Jun 1990.

[10] M. Nilsson, "Entropy and speech," Ph.D. dissertation, Royal Institute of Technology (KTH), 2006.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[12] J. G. Lyons and K. K. Paliwal, "Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement." in *INTERSPEECH*. Citeseer, 2008, pp. 387–390.

[13] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[14] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Elsevier speech communication*, vol. 49, no. 7–8, pp. 588–601, 2007.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 749–752.

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

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[18] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[19] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.