INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Maximum a Posteriori Speech Enhancement Based on Double Spectrum

Pejman Mowlaee[†*], *Daniel Scheran*[*], *Johannes Stahl*[*], *Sean U. N. Wood*[*], *W. Bastiaan Kleijn*[‡]

[†]Widex A/S, Nymøllevej 6, 3540 Lynge, Denmark
[*]Signal Processing and Speech Communication Lab, Graz University of Technology, Austria
[‡]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
{pejman.mowlaee,d.scheran,johannes.stahl,sean.wood}@tugraz.at,bastiaan.kleijn@ecs.vuw.ac.nz

Abstract

While the acoustic frequency domain has been widely used for speech enhancement, usage of the modulation domain is less common. In this paper, we investigate single-channel speech enhancement in the recently proposed Double Spectrum (DS) framework and provide insights on the statistical properties of speech and noise in the DS domain. Relying on our statistical analysis in the DS, we derive a maximum a posteriori estimator of speech in the DS domain. By means of experiments, we evaluate the speech enhancement performance of the proposed method and relevant benchmarks in the acoustic frequency and modulation domains and show that the proposed method achieves a good balance between noise attenuation and speech distortion for various SNRs and noise types.

Index Terms: Speech Enhancement, Modulation Domain Processing, Double Spectrum, MAP Estimator

1. Introduction

Single-channel speech enhancement is often formulated in the acoustic frequency domain generally based on a Short-Time Fourier Transform (STFT). In such an analysis-modification-synthesis (AMS) framework, various noise reduction strategies have been studied to enhance noisy speech signals (for a review see e.g. [1]). As an alternative to the acoustic frequency domain, the modulation transform is of particular interest for speech signal processing, but has been studied less [2–9].

The importance of the modulation domain for speech processing and speech perception has been shown in the literature. Dudley's early research on modulations in human speech showed that a speech signal is well-modeled as a radio broadcast transmitter producing amplitude modulation [10] as a result of modification of the message wave amplitude occurring at syllabic rates. The temporal modulations of the speech spectral envelope contain information about articulation and intonation and have been used successfully to develop the modulation transfer function [11] and speech intelligibility predictors [12]. On the receiver side, the auditory system has been reported to be most sensitive to modulation frequencies below 30 Hz for low acoustic frequencies [13]. According to [14], modulation frequencies between 4 and 16 Hz are crucial for speech intelligibility and modulation frequencies outside this range can be attenuated without severely impairing speech intelligibility. Also, physiological research has provided evidence that neural responses relate to both temporal and spectral signal attributes. This has led to the conclusion that a combination of spectral and temporal modulations is a good candidate to model the auditory cortex spectrogram decomposition [15] which is a natural basis for enhancement.

The work was supported by Austrian Science Fund: P28070-N33.

Recently, several speech enhancement methods that operate in the modulation domain have been proposed [2–7]. For example, the conventional AMS-based acoustic domain was extended to the modulation domain by applying a secondary (modulation) AMS framework computed by again using STFT analysis. The procedure called dual AMS framework has been used for processing speech in the Short-Time Spectral Modulation (STSM) domain. Examples of STSM-based speech enhancement are modulation spectral subtraction (ModSpecSub) [2] and minimum-mean square error spectral modulation magnitude estimator (MMSESPU) [3], both proposed by Paliwal et al.. The role of the modulation magnitude and phase spectra towards speech intelligibility was investigated in [4]. Wojcicki and Loizou [5] proposed a modulation channel selection strategy and argued that by retaining the target-dominated modulations within a narrow band of modulation frequencies (0–8 Hz), it is possible to significantly improve speech intelligibility. Later, Boldt et al. [6] studied the limits and potential of channel selection strategies in both the acoustic frequency and modulation domains. They concluded that though the modulation domain offers potential for improvement in speech intelligibility, due to the inherent compromise between the window length and modulation frequency resolution in the dual AMS framework, the ultimate gain in intelligibility was limited.

Recently, the authors in [8,9,16,17], inspired by the canonical speech representation presented in [18], proposed a two-stage transform called Double Spectrum (DS) consisting of pitch-synchronous modulation transforms providing a highly energy-concentrated representation of speech that allows separating its periodic and aperiodic components [9]. The desired properties of the DS domain were used for harmonicity enhancement and separation of slowly evolving signal components from quickly evolving ones assuming that signal parts caused by noise change more rapidly than those caused by speech. Based on these assumptions, a Wiener filter in the DS domain (DS-Wiener) approach was proposed [9] assuming a normal distribution for both speech and noise DS coefficients.

The novelty of this paper is threefold: i) we present a statistical analysis for speech and noise DS coefficients and provide insights regarding the separability of speech and noise in the DS domain in comparison to the STSM domain, ii) based on this statistical analysis, we derive a maximum a posteriori (MAP) speech estimator in the DS domain (DS-MAP), and iii) finally, we evaluate the performance of the speech enhancement methods investigated here and discuss their limits and potential in terms of noise attenuation and resulting speech distortion.

The remainder of this paper is organized as follows. Section 2 studies the statistics of speech and noise in the DS domain which are then used to derive a MAP speech estimator in the DS domain in Section 3. Experimental results are presented in Section 4, and Section 5 concludes on the work.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1197

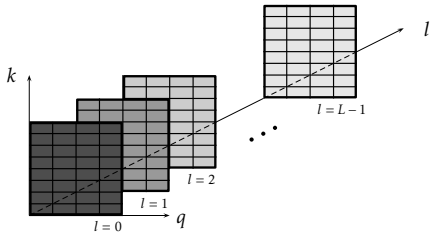

Figure 1: *Double Spectrum tensor after [9]. Using the two-stage transform, a three-dimensional representation is obtained. Each time block l is represented by a $Q \times K$ matrix where Q is the number of modulation bands and K is the number of frequency bands.*

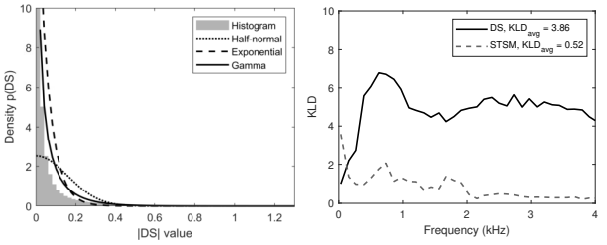

Figure 2: *(Left) histogram of speech DS coefficients with fitted half-normal, exponential, and Gamma distributions, (right) Kullback-Leibler divergence between speech and noise averaged over modulation bands in DS and STSM domains.*

2. Modulation Domains Statistical Analysis

In this Section, the two modulation domains STSM and DS are compared in terms of their statistics and we provide evidence on how well speech and noise are separated in each domain.

2.1. Signal Model and Background on DS

We assume an additive model for noisy speech $x(n)$, i.e. the clean speech $s(n)$ and noise $d(n)$ are additive and independent:

$$x(n) = s(n) + d(n), \tag{1}$$

where n is the discrete-time index.

To obtain the DS coefficients, the noisy signal is processed as follows. Let f_s be the sampling frequency and f_0 be the fundamental frequency of speech in Hz. First, the signal is segmented into L time blocks of variable length by time block segmentation (TBS) with the length of an integer multiple of the fundamental period $P_0 = \lfloor f_s/f_0 \rceil$ in samples. Each block is further subdivided into frames of length P_0. To avoid discontinuities at the transition of consecutive time blocks, 50% overlap is introduced [18]. For unvoiced segments, the fundamental frequency is set to a minimum of 70 Hz. Each block then undergoes a two-stage transform [9]. The first stage, namely the pitch-synchronous transform, is implemented as a Modulated Lapped Transform (MLT) [19], in the form of a Discrete Cosine Transform (DCT-IV), motivated by its effective energy concentration [18]. The MLT is defined as follows [9]:

$$f(k,l) = \sum_{n=0}^{2P_0-1} x_l(n)\, w(n) \sqrt{\frac{2}{P_0}} \cos\left[\frac{(2k+1)(2n-P_0+1)\pi}{4P_0}\right], \tag{2}$$

where $x_l(n)$ is the input signal, l denotes the lth pitch-synchronous time block, k is the acoustic frequency band index, and $w(n)$ is a square-root Hann window satisfying the power complementarity constraint [20].

The MLT coefficients $f(k,l)$ of the output evolve slowly over time for voiced speech and rapidly for unvoiced speech. In order to analyze the temporal fluctuations, the modulation transform, performed with a DCT-II which takes into account the rapid onsets of speech [18], is applied to Q consecutive frames of MLT coefficients, using a boxcar window:

$$g(q,k,l) = \sum_{l=0}^{Q-1} f(k,l) c(q) \sqrt{\frac{2}{Q}} \cos\left[\frac{(2k+1)q\pi}{2Q}\right], \tag{3}$$

with $q = 0, 1, ..., Q-1$ is the modulation band index and $c(0) = 1/\sqrt{2}, c(q) = 1$ for $q \neq 0$. The number of modulation bands Q depends on the block length and is determined by energy concentration measurement as described in [9]. In further notation, we will refer to the set of L subsequent DS frames as $DS(q,k,l)$ with K as the number of frequency bands as rows and Q modulation bands as columns. An example of a DS tensor is shown in Figure 1. For further details on DS transform and its properties we refer to [9].

After processing the noisy DS coefficients, e.g., with a gain function, the enhanced coefficients undergo the inverse two-stage transform, followed by overlap-add (OLA) [9].

2.2. Separation of Speech and Noise in DS and STSM

In this Section, we point out the benefits of performing speech enhancement in the DS domain as opposed to STSM domain. We do so by evaluating empirical speech and noise distributions and comparing their separability in the respective domain.

As noise types we consider white, modulated pink, factory, and babble noise. As speech files we used 72 sentences from the TIMIT train database [21], spoken by male and female speakers and sampled at 16 kHz. As it is important that the samples used to create the histogram obey the same distribution model, i.e. have the same mean and variance, they were normalized with their standard deviation $\sigma_s(q,k,l)$. The variance $\hat{\sigma}_s^2(q,k,l)$ was estimated using recursive averaging [22],

$$\hat{\sigma}_s^2(q,k,l) = \alpha\, \hat{\sigma}_s^2(q,k,l-1) + (1-\alpha)\, \widehat{DS}_s^2(q,k,l), \tag{4}$$

with the smoothing factor chosen as $\alpha = 0.98$. We use $Q = 4$ modulation bands, which was found to be a good compromise in the resulting spectral resolution of the pitch-synchronous analysis provided for male and female speech [9]. We assume the sign of DS coefficients to be uniformly distributed, i.e. $p(DS_s/|DS_s| = -1) = p(DS_s/|DS_s| = 1) = 0.5$, therefore it is sufficient to use absolute values for creating the DS histograms which makes them comparable to the STSM histograms. The smoothing of the variance estimate was performed similarly for the STSM coefficients.

To compare speech and noise separability in the DS versus the STSM domain, we compute the empirical symmetric Kullback-Leibler divergence (KLD) [23]:

$$\text{KLD} = \sum_{i=1}^{N_{bin}} (p_s(i) - p_d(i)) \log_2\left(\frac{p_s(i)}{p_d(i)}\right), \tag{5}$$

where $p_s(i)$ and $p_d(i)$ are the ith histogram bins of the speech and noise histograms, respectively, and N_{bin} is the number of histogram bins. The higher the KLD score, the better the separation between speech and noise distributions.

Figure 2 (left panel) shows the histograms of clean speech

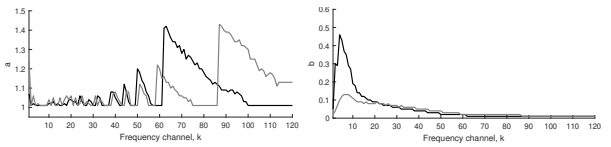

Figure 3: *(Left) shape and (right) scale parameters of the generalized gamma distribution, obtained by histogram fitting over frequency channels k, shown for low ($q = 0$) and high ($q = 3$) modulation bands marked with black and gray color, respectively.*

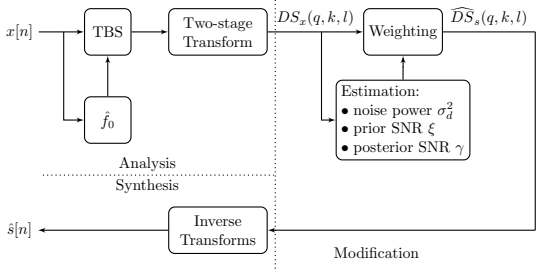

Figure 4: *AMS framework for speech enhancement in Double Spectrum domain (after [9]).*

DS coefficients, averaged over all frequency and modulation bands. Figure 2 (right panel) shows the average KLD over frequency. Averaging the KLD also over frequency yields 3.86 (DS) versus 0.52 (STSM) indicating a better separation between speech and noise in DS domain.

2.3. Distribution of Speech and Noise DS Coefficients

Following our statistical analysis for speech and noise in the DS and STSM domains, in this Section, we fit a parametric distribution to the empirical probability density functions (pdf) of speech and noise in DS domain. We then formulate a speech estimator in DS that takes into account the so-obtained prior knowledge. The statistical model of the DS speech prior is obtained by fitting a pdf to the histograms of the clean speech DS coefficients $DS_s(q,k)$. As the DS coefficients can be also negative, we choose a two-sided generalized gamma distribution which represents a flexible distribution and yields an analytically tractable cost function. The two-sided gamma distribution is defined as (with $\gamma = 1$ in Eq. (1) in [24])

$$p(z|a,b) = \frac{1}{2b^a\Gamma(a)}|z|^{a-1}\exp\left(-\frac{|z|}{b}\right), \tag{6}$$

with $a > 0$ and $b > 0$ as the shape and scale parameters, respectively, and $\Gamma(\cdot)$ denoting the Gamma function. The values for a and b were determined for each acoustic frequency-modulation frequency combination separately by fitting a PDF to the respective histogram. We used the same data as in Section 2.2. According to Figure 3, both a and b are dependent on both k and q.

A similar statistical analysis was conducted on noise signals concluding that a normal distribution is a good fit.

3. DS-MAP Estimator

In this Section, we focus on speech enhancement in the DS domain. The MAP estimator resulting from the statistics of speech and noise will be derived. The AMS framework for speech enhancement in DS is shown in Figure 4.

As statistical model for speech, we use (6) with $z = \mathrm{DS}_s/\sigma_s$, resulting in $p(\mathrm{DS}_s/\sigma_s)$ as prior distribution of the speech DS coefficients.[1] The parameters of the prior are chosen according to the distribution fit performed in Section 2.3. For noise DS coefficients we consider the prior distribution following a zero-mean normal distribution, as

$$p(DS_d) = \frac{1}{\sqrt{2\pi\sigma_d^2}}\exp\left(-\frac{DS_d^2}{2\sigma_d^2}\right), \tag{7}$$

where σ_d^2 is the variance of noise. Based on these models, a MAP solution

$$\widehat{DS_s} = \arg\max_{DS_s}\frac{p(DS_x|DS_s)p(DS_s/\sigma_s)}{p(DS_x)}, \tag{8}$$

can be found. Since $p(DS_x)$ is not a function of DS_s, we only need to maximize the numerator of (8) [25]. Plugging the speech and noise prior distributions in (8), we obtain

$$p(DS_x|DS_s)p\left(DS_s/\sigma_s\right) = \frac{1}{2b^a\Gamma(a)\sqrt{2\pi\sigma_d^2}}\left(\left|\frac{DS_s}{\sigma_s}\right|\right)^{a-1} \exp\left(-\frac{(DS_x-DS_s)^2}{2\sigma_d^2} - \frac{1}{b}\left|\frac{DS_s}{\sigma_s}\right|\right). \tag{9}$$

As the logarithm is a monotonically increasing function, we can alternatively maximize the logarithm of (9) [25] given by:

$$\frac{\partial}{\partial DS_s}\left\{(a-1)\log(|DS_s|) - \frac{(DS_x-DS_s)^2}{2\sigma_d^2} - \frac{|DS_s|}{b\sigma_s}\right\} = 0 \tag{10}$$

Finding the stationary point yields:

$$\frac{(a-1)}{DS_s} - \frac{DS_s - DS_x}{\sigma_d^2} - \frac{1}{b\sigma_s}\frac{DS_s}{|DS_s|} = 0. \tag{11}$$

Maximization of the log-posterior, i.e., terms in the brackets in (10) with respect to the sign leads to the minimization of the term $(DS_x - DS_s)^2$ as this is the only term dependent on the DS sign yielding $\frac{DS_s}{|DS_s|} = \frac{DS_x}{|DS_x|}$. In fact, this is also inferred from the fact that both prior and likelihood are symmetric around zero. Now, plugging this DS sign into (11) we get a quadratic function for DS_s:

$$DS_s^2 - DS_s\left(1 - \frac{\sigma_d^2}{b\sigma_s|DS_x|}\right)DS_x - \sigma_d^2(a-1) = 0 \tag{12}$$

Solving (12) for DS_s by using the definitions of the a priori SNR $\xi = \sigma_s^2/\sigma_d^2$ and the a posteriori SNR $\zeta = DS_x^2/\sigma_d^2$, we finally obtain the MAP speech estimate:

$$\widehat{DS_s} \quad = \quad G_{\mathrm{MAP}}(\xi,\zeta,a,b)\cdot DS_x, \tag{13}$$

$$G_{\mathrm{MAP}}(\xi,\zeta,a,b) = \frac{u}{2} + \sqrt{\left(\frac{u}{2}\right)^2 + \frac{a-1}{\zeta}},\ u = 1 - \frac{1}{b\sqrt{\xi\zeta}}. \tag{14}$$

The gain function is real-valued defined for $a \geq 1$. In order to prevent artifact we lower-bound the gain to a minimum value G_{min}. Our informal listenings revealed that a low fixed-value -25 dB for G_{min} results in audible artifacts at low modulation bands as they mainly contain periodic signal contents. On the other hand, a larger G_{min} as -8 dB provides not enough noise attenuation. Therefore, as a trade-off we consider an adaptive choice of G_{min} implemented by assigning its value between -8 and -25 dB for $q \in \{0,1\}$ depending on the prior SNR estimate following a logistic function with steepness set to 0.27. For $q \in \{2,3\}$ a fixed value of -25 dB was chosen.

[1] For the sake of readability, the time frame index l and the modulation band index q will be omitted in the following.

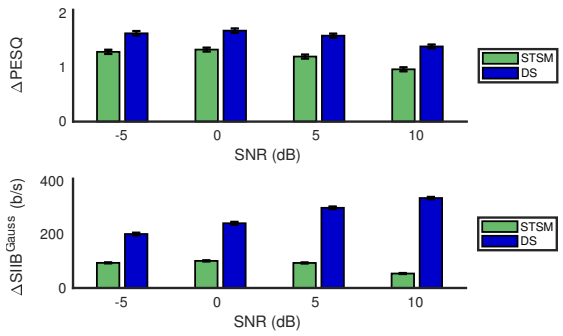

Figure 5: *Improvements in (top) PESQ and (bottom) SIIB^{Gauss} versus noisy speech (AWGN) in STSM and DS domains using binary masking.*

4. Experiments and Results

In this Section, we evaluate modulation-based speech enhancement methods in three ways: i) comparing the maximum achievable improvement in terms of objective perceived quality and speech intelligibility in the STSM and DS domains using ideal channel selection (assuming speech and noise known), ii) visual comparison of spectrograms, and iii) reporting noise attenuation versus speech distortion in a blind scenario.

4.1. Experiment Setup

We used 100 sentences from the *TIMIT* test corpus [21]. The clean speech was corrupted with babble and factory noise at SNRs of -5, 0, 5, and 10 dB. As benchmark methods, we included the MMSE-STSA method in the acoustic frequency domain [26], ModSpecSub in the STSM domain [2], and DS-Wiener [9]. The minimum statistics noise estimator [27] was used for MMSE-STSA, and an adaptation of the MMSE noise estimator based on speech presence probability [28] was implemented for DS-MAP. For f_0 estimation we used PEFAC [29]. The methods were evaluated using the black box approach [30]. Speech quality was evaluated using Perceptual evaluation of speech quality (PESQ) [31] and speech intelligibility was predicted by speech intelligibility in bits (SIIB^{Gauss}) [32]. As measures for speech distortion and noise attenuation we used speech-to-speech distortion rate (SSDR) versus noise attenuation (NA) as defined in [33]

4.2. Upper bounds: Ideal Channel Selection

We first performed a proof-of-concept experiment to determine the upper bounds of the modulation-based speech enhancement methods. We performed ideal channel selection in both STSM and DS domains by computing the ratio of speech and noise energy and computing binary mask by comparing this ratio against a threshold $\theta = -10$ dB (see [6]). The results depicted in Figure 5 reveal that under oracle conditions, i.e., binary masking applied to noisy speech with known speech and noise power, the DS-enhanced speech achieves higher objective improvements than the STSM-enhanced speech. This is consistent with the fact that the signals are better separated in the DS domain.

4.3. Spectrograms

Figure 6 shows spectrograms for speech enhancement results in blind scenario, provided by Modulation Spectral Subtraction [2] and the proposed DS-MAP method for babble noise at SNR = 5 dB. The STSM enhanced speech introduces more residual noise (blue box). In speech dominated regions, STSM

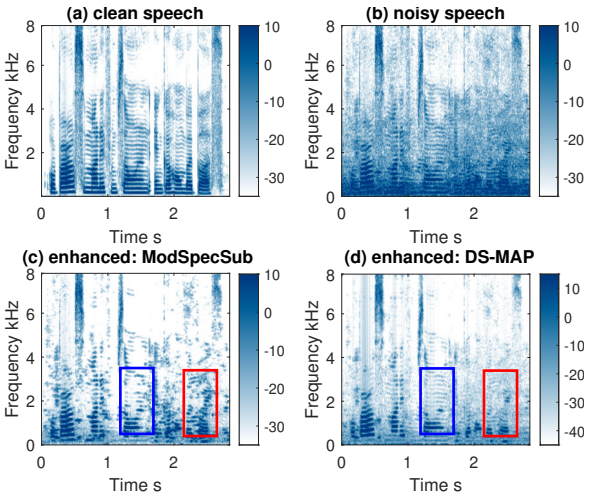

Figure 6: *Spectrograms of (a) clean speech, (b) noisy speech (babble noise, SNR = 5 dB), (c) enhanced speech obtained by Modulation Spectral Subtraction (STSM domain), (d) enhanced speech by DS-MAP.*

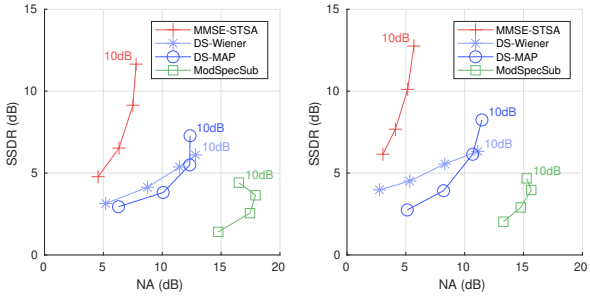

Figure 7: *Noise attenuation vs. speech-to-speech distortion ratio for (left) factory noise and (right) babble noise for STFT and modulation-based speech enhancement methods at SNR* $\in \{-5, 0, 5, 10\}$ *dB.*

introduces more speech distortion while the DS-MAP estimator better preserves these regions (red box).

4.4. Noise Attenuation vs. Speech Distortion

Figure 7 shows the SSDR and NA in blind scenario. MMSE-STSA has the lowest speech distortion at the expense of the lowest noise attenuation. STSM provides the highest noise attenuation while distorting speech the most. DS-based methods provide a good trade-off between noise attenuation and speech distortion. In particular, the proposed DS-MAP achieves higher noise attenuation than STSA and DS-Wiener. It also results in less speech distortion compared to both DS-Wiener and STSM, in particular for mid-high SNRs (listening examples in [34]).

5. Conclusion

We studied the statistical properties of speech and noise in the recently proposed DS signal representation and derived a maximum a posteriori speech estimator in the DS domain. Our experiments showed that the proposed method provides a better trade-off between the noise attenuation and speech distortion for various noise types and SNRs versus its benchmarks in the modulation or acoustic frequency domains. Our preliminary experiments based on oracle sign information in DS suggests significant improvement in speech quality and intelligibility. Therefore, a natural direction of future work is the estimation of the sign from a noisy observation.

6. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: a survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, 2013.

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

[3] K. Paliwal, B. Schwerin, and K. Wojcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *speech communication*, vol. 54, no. 2, pp. 282 – 305, 2012.

[4] ——, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *speech communication*, vol. 53, no. 3, pp. 327–339, 2011.

[5] K. K. Wojcicki and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2904–2913, 2012.

[6] J. B. Boldt, A. T. Bertelsen, F. Gran, S. Jørgensen, and T. Dau, "Single channel speech enhancement in the modulation domain: New insights in the modulation channel selection framework," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 2015, pp. 5748–5752.

[7] Y. Wang and M. Brookes, "Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5225–5229.

[8] M. Blass, P. Mowlaee, and W. B. Kleijn, "Single-Channel Speech Enhancement Using Double Spectrum," *Proc. Interspeech*, pp. 1740–1744, 2016.

[9] P. Mowlaee, M. Blass, and W. B. Kleijn, "New results in modulation-domain single-channel speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 25, no. 11, pp. 2125–2137, Nov. 2017.

[10] H. Dudley, "The carrier nature of speech," *The Bell System Technical Journal*, vol. 19, no. 4, pp. 495–515, Oct. 1940.

[11] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, March 1985.

[12] 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, Sep. 2011.

[13] E. Zwicker, "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones," *Acta Acustica*, vol. 2, no. 3, pp. 125–133, 1952.

[14] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[15] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[16] F. Huang, T. Lee, and W. B. Kleijn, "Transform-domain Speech Periodicity Enhancement with Adaptive Coefficient Weighting," *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–5, Dec. 2011.

[17] ——, "Transform-domain wiener filter for speech periodicity enhancement," pp. 4577–4580, March 2012.

[18] 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, Apr. 2007.

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

[20] M. Nilsson, "Entropy and Speech," Ph.D. dissertation, Sound and Image Processing Laboratory, KTH Electrical Engineering, 2006.

[21] J. Garfolo, L. Lamel, and W. Fisher, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Philadelphia: Linguistic Data Consortium, 1993.

[22] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. International Workshop on Acoustic Signal Enhancement.*, Aug. 2010.

[23] I. Andrianakis and P. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *speech communication*, vol. 51, no. 1, 2008.

[24] J. W. Shin, J. H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 258–261, March 2005.

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

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

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

[28] T. Gerkmann and R. Hendriks, "Noise power estimation based on the probability of speech presence," *Proc. International Workshop on Acoustic Signal Enhancement.*, 2011.

[29] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[30] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards objective quality assessment of speech enhancement systems in a black box approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2008, pp. 273–276.

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

[32] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, Nov 2018.

[33] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

[34] P. Mowlaee, D. Scheran, J. Stahl, S. U. N. Wood, and W. B. Kleijn, "Maximum a posteriori speech enhancement based on double spectrum." [Online]. Available: https://www2.spsc.tugraz.at/people/pmowlaee/DSMAP