SINGLE CHANNEL SPEECH ENHANCEMENT IN THE MODULATION DOMAIN: NEW INSIGHTS IN THE MODULATION CHANNEL SELECTION FRAMEWORK

Jesper B. Boldt[1]*, Andreas T. Bertelsen*[2]*, Fredrik Gran*

GN ReSound A/S
DK-2750 Ballerup, Denmark

Søren Jørgensen[2]*, Torsten Dau*

Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

Recently, the ideal binary mask has been introduced in the modulation domain by extending the ideal channel selection method to modulation channel selection [1]. This new method shows substantial improvement in speech intelligibility but less than its predecessor despite the higher complexity. Here, we extend the previous finding from [1] and provide a more direct comparison of binary masking in the modulation domain with binary masking in the time-frequency domain. Subjective and objective evaluations are performed and provide additional insight into modulation domain processing.

Index Terms— Speech Enhancement, Speech Intelligibility, Modulation Domain Processing, Binary Masking

1. INTRODUCTION

Improving the intelligibility of speech has been a research topic in signal processing for decades [2], but until recently [3, 4], no substantial gain in intelligibility has been reported by any single channel algorithm. Different approaches to the problem have been taken, e.g. spectral subtraction, Wiener filtering, subspace methods, statistical models, etc., but in general, these methods improve speech quality rather than speech intelligibility [2]. Both quality and intelligibility of a speech signal are important, but they are not necessarily correlated. Some algorithms have increased intelligibility at the expense of reduced quality, e.g. binary masking [5], but in applications like hearing aids and telephony, both intelligibility and quality are of major importance.

From Dudley's famous experiment with the vocoder [6], we know that the temporal envelope of the speech signal is of high importance for speech intelligibility. This finding has been confirmed in many later studies (see e.g. [7, 8]). Two studies, [9, 10], showed the importance of the modulations of the temporal envelope in the frequency range from 1 to 16 Hz, which corresponds approximately to the speed of movements in the vocal tract and the syllabic rate [11]. The energy in this frequency range can be seen in Figure 1.C, but higher frequencies are also present, particularly at onsets and offsets in the speech signal, as seen in Figure 1.D.

The importance of the temporal envelope of the speech signal has also been supported by results in psychoacoustics, neuroscience [12], and by the use of noise and sine vocoders (see e.g. [13]). In psychoacoustics, the modulation filterbank has been introduced [14], and modulations have been used to predict the intelligibility of the speech signal [15].

Knowing that the envelope is important to speech intelligibility makes it intriguing to do speech enhancement in the modulation domain. If the envelope of the speaker of interest can be enhanced so that it is easier to perceive, a gain in intel-

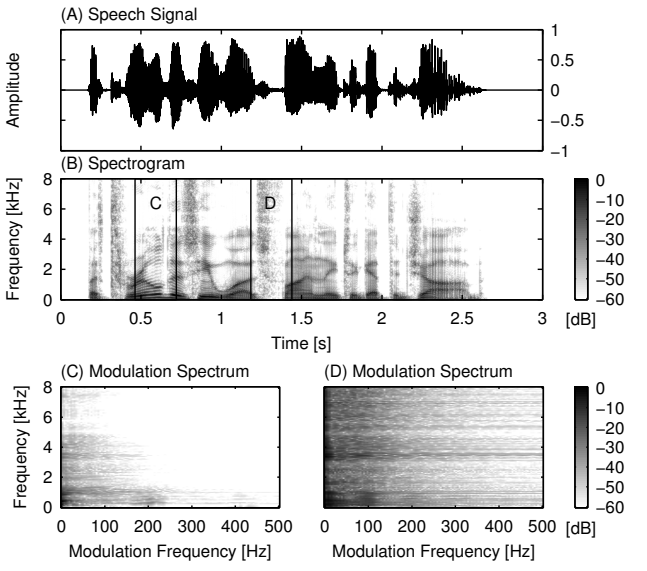

Fig. 1. Speech modulations in the sentence *"but thrilled as he was finally to get the job"*. The modulation spectrum is shown for two segments (C) and (D).

1) Corresponding author: jboldt@gnresound.com, 2) is now with Oticon A/S, Denmark. Thanks to Tobias May, CAHR, DTU, for valuable input to this work.

ligibility or quality might be obtained. This idea has led to different modulation methods with encouraging results, see e.g. [16, 17, 18]. One of these methods is inspired by the intelligibility improvements obtained using an ideal binary mask method [19], also called ideal channel selection [20]. This binary approach was recently introduced in the modulation domain through the Modulation Channel Selection framework [1], as explained in Section 1.1. Modulation Channel Selection (MCS) improved speech intelligibility substantially, but, despite its additional processing and complexity, the improvement was smaller than that of its predecessor, Ideal Channel Selection (ICS). In this paper, we provide an explanation of the difference in intelligibility results, an in-depth comparison of the MCS and ICS frameworks, and a perspective on modulation domain processing in general.

1.1. Modulation Channel Selection

Modulation domain processing using MCS can be seen as ICS performed in the modulation domain. In ICS, a binary gain is applied to the noisy speech in the time-frequency domain, where the binary gain is computed using ideal knowledge of the clean speech and noise signal. If the time-frequency unit is dominated by speech energy, a gain of one is used; otherwise, a gain of zero is used. With MCS the binary gain is calculated in the modulation domain as a function of time, acoustic frequency, and modulation frequency. Like ICS, the binary gain in MCS is based on ideal knowledge of the clean speech signal, $s(n)$, and noise signal, $d(n)$, before being mixed:

$$G_{k,l}(n_m) = \begin{cases} 1 & \frac{|S_{k,l}(n_m)|^2}{|D_{k,l}(n_m)|^2} \geq \theta \\ 0 & \text{otherwise} \end{cases}, \qquad (1)$$

where $G_{k,l}(n_m)$ is the binary gain at acoustic frequency k, modulation frequency l, and time index n_m. $|S_{k,l}(n_m)|^2$ and $|D_{k,l}(n_m)|^2$ are the modulation energies of the speech and noise signal respectively. θ is a threshold determining the sparsity of the binary mask.

In MCS, the dual-AMS (analysis-modification-synthesis) framework [17], seen in Figure 2, is used to calculate and apply the binary mask. In this framework, the input signal, $x(n) = s(n) + d(n)$, is filtered into subbands using a short time Fourier Transform (STFT), and a second STFT is applied to the magnitude of each subband signal to calculate the modulation frequencies. After modification by $G_{k,l}(n_m)$, the signal is synthesized using two inverse STFTs and the noisy phase from the input signal ($\angle Z_{k,l}(n_m)$ and $\angle X_k(n_a)$). A summary of the parameters used in the dual-AMS framework can be seen in Table 1.

In [1], the intelligibility of MCS processed speech mixed with babble noise was evaluated using $\theta = -5$ dB and $\theta = -10$ dB in Equation 1. The highest improvements were found using $\theta = -10$ dB, where the MCS provided a 14 dB

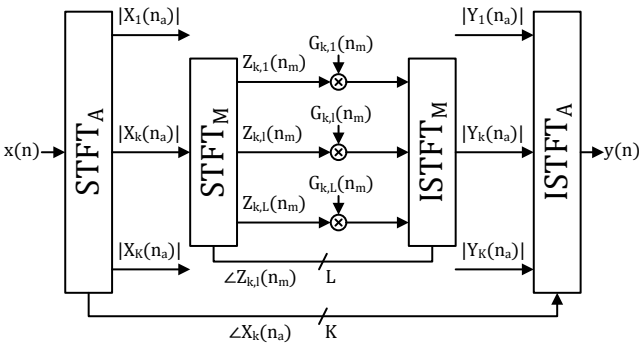

Fig. 2. Dual-AMS framework for modulation domain processing. The input signal, $x(n)$, is split into acoustic subbands, $X_k(n_a)$, and subsequently into modulation subbands, $Z_{k,l}(n_m)$. The three-dimensional gain matrix, $G_{k,l}(n_m)$, is applied in the modulation domain. n, n_a, and n_m are the time indices in the time domain, time-frequency domain, and modulation domain, respectively. K is the number of acoustic frequency bands, L is the number of modulation frequency bands.

Parameter	MCS	ICS	wICS	MCS	ICS	wICS
unit	samples			ms		
w_a	512	512	512	32	32	32
s_a	128	128	128	8	8	8
FFT_a	1024	1024	1024	-	-	-
w_m	32	-	32	256	-	256
s_m	4	-	4	32	-	32
FFT_m	64	-	-	-	-	-

Table 1. Parameters used in the experiment. w_a, s_a, and FFT_a are the window size, window shift, and FFT size in the acoustic domain. w_m, s_m, and FFT_m are the window size, window shift, and FFT size in the modulation domain.

improvement in intelligibility as measured by the speech-reception-threshold (SRT). For comparison, the ICS provided a substantial larger improvement of speech intelligibility with a 28.25 dB improvement in SRT. An example of MCS and ICS processing can be seen in Figure 3.

2. METHOD

Although the MCS method needs more, and more complex processing, it is not able to outperform the ICS method in terms of SRT improvement. Comparing the methods in more detail reveals a difference that may explain why: to be able to analyze and modify modulation frequencies with a high resolution at low frequencies, each frame in the modulation domain has a length of 256 ms. The large frame length introduces time smearing in the MCS processed signal, as seen in Figure 3.C. This could explain the limited improvement in

intelligibility compared to ICS (3.D). The ICS-processed signal has less time smearing, and by inspection appears closer to the original signal.

To test the hypothesis that the long modulation frames reduce the benefit from MCS processing, a new type of processing is introduced in the present work, making it possible to directly compare the benefit of binary decisions in the time frequency domain vs. the modulation domain. This new method is ICS with a binary decision at the same time resolution as MCS. This is done by time weighting the magnitudes, $|X(m,n)|$, with the window, w_a, used for modulation analysis. For this reason, we refer to this method as weighted ICS (wICS). This new method makes it possible to compare the benefit of keeping specific modulations instead of the more simple decision based on the total energy of each time-frame. Additionally, three different noise signals with varying modulation content are used to compare the benefit of ICS, wICS, and MCS.

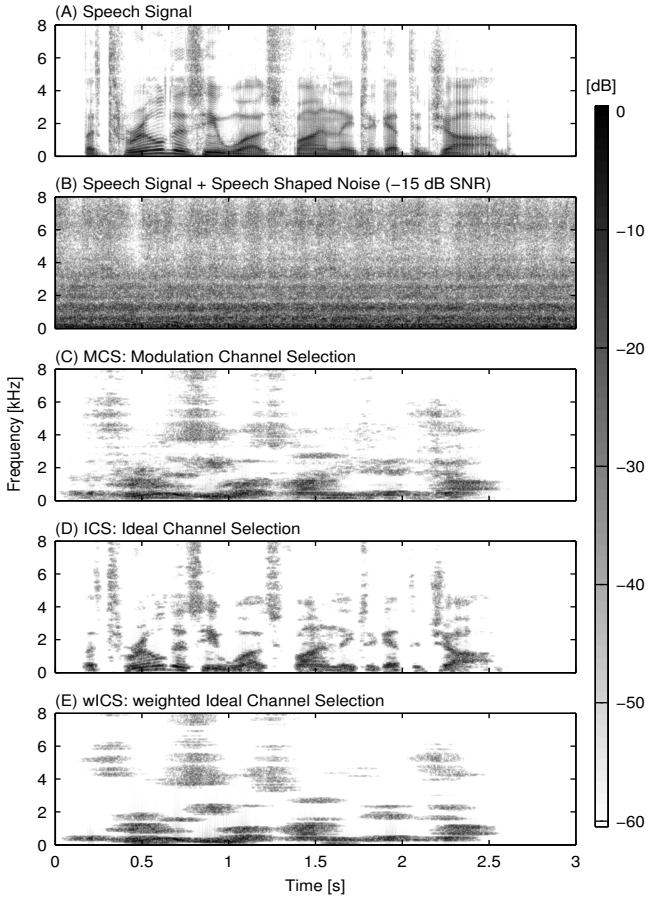

Fig. 3. The speech signal *"but thrilled as he was finally to get the job"* (A) is mixed with speech shaped noise (B) and processed using MCS (C), ICS (D), and wICS (E). Clearly, MCS and wICS introduce time-smearing when compared to the ICS. ICS has the highest similarity with the clean speech signal, and onsets and offsets can more easily be identified.

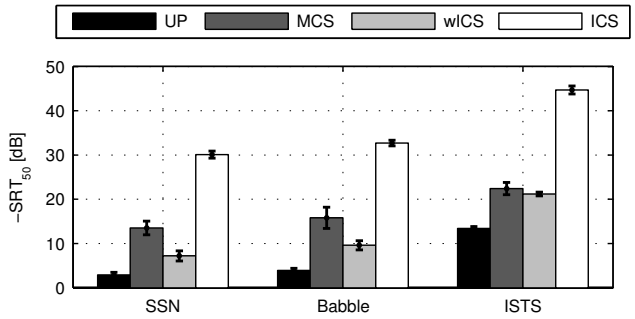

Fig. 4. Mean $\text{SRT}_{50\%}$ for unprocessed (UP), MCS, wICS, and ICS. The three noise types are speech shaped noise (SSN), multi-talker babble, and a female speech sound (ISTS). ICS provides the largest gain in speech intelligibility in all conditions. The benefit of MCS over wICS is mainly seen in the SSN and Babble noise conditions.

2.1. Subjective Evaluation

To measure the intelligibility of wICS relative to ICS and MCS, a subjective listening experiment was carried out using three different noise types: Speech shaped noise (SSN), multitalker babble [21], and a single female speaker sound [22]. Intelligibility was measured by 50% SRT using the Danish speech intelligibility test CLUE [23]. Five persons with normal hearing participated in the experiment, which took place in a sound treated room. All sounds were presented through headphones at 65dB SPL. A training session was conducted prior to the listening experiment. The parameters used for MCS, ICS, and wICS are presented in Table 1, and the threshold, θ, in Equation 1 was kept at -10 dB to enable comparison with the results from [1].

3. RESULTS

The results from the listening test are presented in Figure 4. An Analysis of Variance (ANOVA) test [24] suggested that performance in our babble noise condition did not differ from performance in [1]. As expected, ICS performs significantly better than MCS in babble noise, as well as in the SSN and ISTS noise conditions. Figure 4 shows that when the noise is modulated, ICS provides the largest gain in intelligibility, as measured by change in SRT. This result is in line with previous studies [25]. In MCS, we find the opposite result; the smallest SRT benefit is found in highly modulated speech (ISTS). Comparing MCS with wICS yields a similar conclusion: the benefit of doing modulation domain processing is largest for the more stationary noises, SSN and Babble. With ISTS, the difference between MCS and wICS is non-significant [24], indicating that modulation domain processing is most efficient when the modulations in the target and noise signals are different.

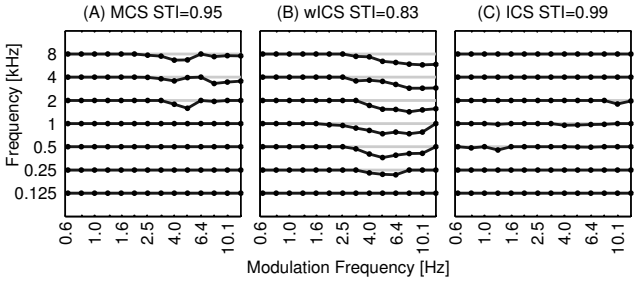

Fig. 5. Modulation Transfer Functions for each of the processing types when speech is mixed with SSN at -15 dB. Each Figure shows the 98 modulation reduction factors in the processed signal relative to the clean speech signal. If all modulations are preserved, the black lines will be on top of the gray lines. The scale between the gray lines is one.

3.1. Modulation Transfer Function

Additional insight into the three types of processing can be obtained using the modulation transfer functions (MTF) from the speech transmission index (STI) [26]. With STI, intelligibility is predicted by the weighted average of the reduction of modulations at 7 acoustic frequencies (octave-spaced between 125 and 8000 Hz) and 14 modulation frequencies (1/3-octave-spaced between 0.63 and 12.7 Hz). The 98 modulation reduction factors provide insight into the effect of the three processing methods, as seen in Figure 5. This figure shows that ICS is better at preserving speech modulations than wICS or MCS. wICS shows the highest reduction in modulations, mainly at high modulation frequencies. This may be explained by the severe time smearing obtained using wICS.

4. DISCUSSION

The results obtained in the current study provide additional insight into modulation domain processing and MCS in particular. We do see a benefit of doing modulation domain processing rather than time-frequency processing when comparing them on the same time-scale. However, the benefit is smallest for the ISTS signal (1.2 dB) and largest for the SSN signal (6.3 dB). In contrast, the ICS obtains best performance with the highly modulated ISTS signal compared to the more stationary SSN signal, which corresponds to previous results [25]. This finding can be explained by differences in the modulation content of the different noise signals used in this study: when speech is mixed with ISTS, the modulations from the target signal and noise signal might be similar, whereas in speech mixed with SSN, the modulations will mainly be generated by the target speech. In other words, the target modulations might be easier to pick out for the MCS algorithm when speech is mixed with the stationary noise (SSN).

The larger improvement by ICS over MCS is also supported by visual inspection of Figure 3, where speech is mixed with SSN at -15 dB SNR. The larger similarity between ICS (D) and the clean speech signal (A) is evident when compared to MCS (C) and wICS (E). The ICS preserves more information about onsets, offsets, pitch, and formants, whereas the onsets and offsets in the processed speech are severely smeared using MCS and wICS. Using ICS, the onset at 1.4 s (the word 'finally') can be seen in the ICS processed signal, whereas with MCS and wICS, the onset is is smeared and difficult to distinguish.

The time smearing introduced by MCS and wICS is a drawback of the long window used for modulation domain analysis in combination with the binary modifications in the modulation domain. If no modifications were applied in the modulation domain, perfect reconstruction could be obtained with the dual-AMS method, but the binary decision and the synthesis using the noisy phase will smear out the signal content. An interesting extension of MCS could be to include the phase information in the modulation processing, similar to what has been seen in recent time-frequency domain speech enhancement algorithms (see e.g. [27, 28]).

The long modulation analysis window is necessary in the MCS framework to obtain a useful resolution at the low modulation frequencies. With the parameters in Table 1, the modulation frequency resolution is 1.95 Hz. This long window makes it a challenge to process more transient speech sounds while maintaining a high time precision. Reducing the size of the analysis window will reduce the modulation frequency resolution and make the MCS processing converge towards that of ICS. Different solutions to this inherent problem in MCS could be studied, e.g. by the use of different analysis and synthesis window lengths, asymmetric windows [29], or multi-resolution window lengths similar to what is being used in the intelligibility predictor SNR_{env} [30].

To be able to compare with previous results, the threshold, θ, was kept at -10 dB in this study. Obviously, this is a limiting factor in the current study, and -10 dB has not be proven to be the optimal value for increasing intelligibility using MCS. However, we expect the ranking of ICS, MCS, and wICS with respect to SRT to be the same at other θ values, which has been supported by preliminary objective evaluations using the STOI measure [31]. These expectations must be verified by the use of subjective listening experiments.

5. CONCLUSION

Modulation domain processing in the form of modulation channel selection (MCS) is able to increase intelligibility of noisy speech by a substantial amount. However, the compromise between modulation frequency resolution and window length limits the gain in intelligibility by the MCS algorithm. This limitation must be carefully considered, and means to avoid or reduce this limitation should be investigated.

6. REFERENCES

[1] K. Wójcicki and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[2] P. C. Loizou, *Speech enhancement: theory and practice*, CRC Press, 1st edition, 2007.

[3] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[5] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[6] H. Dudley, "Remaking speech," *Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[7] R. V. Shannon, F-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[8] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and T. Lunner, "Speech perception of noise with binary gains," *Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[9] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[10] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[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," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[12] S. Shamma, "Encoding sound timbre in the auditory system," *IETE Journal of research*, vol. 49, no. 2/3, pp. 145–156, 2003.

[13] P. Souza and S. Rosen, "Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 792–805, 2009.

[14] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers," *Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[15] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *Journal of the Acoustical Society of America*, vol. 130, pp. 1475, 2011.

[16] S. M. Schimmel and L. E. Atlas, "Target talker enhancement in hearing devices," in *Proc. of ICASSP*. IEEE, 2008, pp. 4201–4204.

[17] K. Paliwal, K. Wójcicki, 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, 2010.

[18] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *Signal Processing, IEEE Transactions on*, vol. 57, no. 11, pp. 4323–4332, 2009.

[19] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[20] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[21] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[22] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ists)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

[23] J. B. Nielsen and T. Dau, "Development of a Danish speech intelligibility test.," *International Journal of Audiology*, vol. 48, no. 10, pp. 729–41, Jan. 2009.

[24] A. T. Bertelsen, "Modulation domain processing for enhanced speech intelligibility," Tech. Rep., Center for Applied Hearing Research, Technical University of Denmark, 2014.

[25] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[26] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[27] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the stft-phase," *Signal Processing Letters, IEEE*, vol. 20, no. 2, pp. 129–132, 2013.

[28] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[29] D. Mauler and R. Martin, "Improved reproduction of stops in noise reduction systems with adaptive windows and nonstationarity detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 2, 2009.

[30] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 2013.

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