# The role of window length and shift in complex-domain DNN-based speech enhancement

*Celia García-Ruiz*[1], *Juan Manuel Martín-Doñas*[2], *Angel M. Gómez*[1]

[1] Dpt. Signal Theory, Telematics and Communications, University of Granada, Spain
[2] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

{cgr14,amgg}@ugr.es,jmmartin@vicomtech.org

## Abstract

Deep learning techniques have widely been applied to speech enhancement as they show outstanding modeling capabilities that are needed for proper speech-noise separation. In contrast to other end-to-end approaches, masking-based methods consider speech spectra as input to the deep neural network, providing spectral masks for noise removal or attenuation. In these approaches, the Short-Time Fourier Transform (STFT) and, particularly, the parameters used for the analysis/synthesis window, plays an important role which is often neglected. In this paper, we analyze the effects of window length and shift on a complex-domain convolutional-recurrent neural network (DCCRN) which is able to provide, separately, magnitude and phase corrections. Different perceptual quality and intelligibility objective metrics are used to assess its performance. As a result, we have observed that phase corrections have an increased impact with shorter window sizes. Similarly, as window overlap increases, phase takes more relevance than magnitude spectrum in speech enhancement.

**Index Terms**: Speech enhancement, Deep neural network, Short Time Fourier Transform, Complex spectral masking

## 1. Introduction

The ubiquitous use of voice-based services, fostered by the spread use of smartphones, smart-TVs or home assistant devices, often results in acquired signals contaminated with acoustical noise. This noise causes a reduction of the speech quality and intelligibility, making speech enhancement a necessity. Classical approaches based on statistical signal processing [1], such as spectral-subtractive, statistical modeling or subspace algorithms, were first introduced to deal with acoustical noise. Weiss et al. [2] were pioneers in proposing the first spectral-subtractive algorithm based on signal correlation, and later [3] based on the Fourier transform. Some statistical techniques as the maximum-likelihood estimation or Wiener filter were proposed to approximate the Fourier transform of clean speech, given the noisy signal [4], as well as statistics of superior order applied to Wiener filtering were suggested in [5] afterwards.

At the present time, Deep Neural Network (DNN) based techniques have dominated the speech enhancement area due to their excellent modeling capabilities and performance in speech-noise separation tasks. In general, approaches based on DNNs can be divided into those that only consider the time domain, providing an end-to-end solution, and those which instead consider a transformed time-frequency domain, applying spectral masking methods [6]. Nonetheless, there exist other approaches which explore alternative domains learned by the DNN itself [7]. In end-to-end solutions, DNNs are applied as black boxes in which process is carried out directly in the time domain without computing any intermediate features [8].

In contrast, spectral masking approaches make use of the Short-Time Fourier Transform (STFT) to obtain a time-frequency representation of the signal which is employed as input to the DNN architecture, while a mask for that representation is obtained as output [9]. Since spectra are complex-valued, magnitude of the spectra was often used instead. However, due to the relevance of the phase-spectrum [10], complex-valued DNNs have been proposed to cope with the full spectrum [11]. This way, the provided mask aims to correct both magnitude and phase.

In this paper we focus on these complex-valued spectral masking approaches. In particular, we are interested in the role that the STFT transform plays in them. As suggested by other authors [12], the STFT analysis window, particularly, the window length and shift considered, can have a noticeable effect on the DNN performance which is often neglected. Motivated by this, the aim of the paper is to carry out an in-depth analysis of the role of the STFT parameters in complex spectral masking. To this end, we have implemented a recently proposed complex-valued DNN, the so-called DCCRN [11] and have trained it considering not only several window lengths but also different shifts. Then, we have assessed the performance of the network when correcting the magnitude spectra only, the phase only and the complete spectra with a variety of objective metrics, involving perceptual and intelligibility ones.

The rest of this paper is organized as follows: in Section 2, we describe the methodology and the signal processing pipeline employed. Section 3 details the experimental framework while Section 4 is devoted to the experimental results. Finally, in Section 5 we summarize the conclusions of this paper.

## 2. Spectral-domain processing

In this paper we have considered the additive noise model, given by $x(n) = s(n)+r(n)$, where $x(n)$, $s(n)$ and $r(n)$ are, respectively, the contaminated signal, clean speech and noise. Signals are processed in the spectral domain by means of three key elements: spectral analysis, masking and overlap and add. Figure 1 depicts a diagram of the speech processing pipeline followed in this paper.

### 2.1. Spectral analysis

The STFT is widely used to obtain a time-frequency representation of signals [10]. In practice, frequency $\omega$ is preferred to be discrete, being the Discrete Fourier Transform (DFT) then applied. As a result, a spectral representation is obtained per

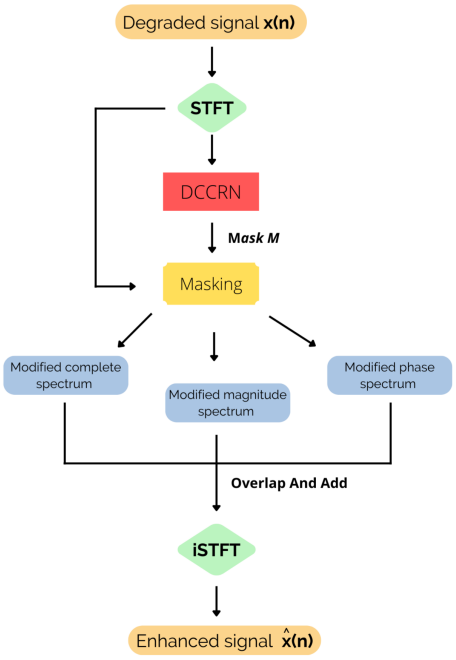



Figure 1: *Spectral domain DNN-based processing pipeline.*

frame as follows, e $e$

$$X(l,k) = \sum_{n=0}^{M-1} x(lH+n)w(n)e^{-j2\pi\frac{kn}{M}}, \quad (1)$$

where $k$ is the frequency index (representing bins of $\frac{2\pi}{M}$ radians per sample), $l$ the frame index, $x(n)$ the input signal and $w(n)$ the analysis window. As can be noted, $x(n)$ is segmented in overlapped frames of length $M$ and shift $H$.

The number of frequency bins is lower-bounded by the window length $M$ ($k = 0, \ldots, M-1$) which is often chosen as a power of two to allow a fast DFT computation via Fast Fourier Transform (FFT). However, in this paper we apply zero padding to keep constant the number of bins, $K$, despite the window length used. This is because the number of the DNN parameters varies with the input size and we want to ensure a fair comparison across them.

### 2.2. Masking and speech synthesis

The masking process is carried out frame-by-frame. It consists of applying the estimated time-frequency mask $\mathbf{M}(k,l)$, which can also be defined in polar coordinates as $\mathbf{M} = |\mathbf{M}|e^{j\phi_\mathbf{M}}$, to the noisy spectrum $\mathbf{X}(k,l)$ (alternatively, $\mathbf{X} = |\mathbf{X}|e^{j\phi_\mathbf{X}}$). Subsequently, we can correct magnitude, phase or both as follows

$$\widehat{\mathbf{S}}^{(complete)} = \mathbf{X} \odot \mathbf{M}, \quad (2)$$

$$\widehat{\mathbf{S}}^{(magnitude)} = |\mathbf{X}| \odot |\mathbf{M}|e^{j\phi_\mathbf{X}}, \quad (3)$$

$$\widehat{\mathbf{S}}^{(phase)} = |\mathbf{X}|e^{j(\phi_\mathbf{X}+\phi_\mathbf{M})}, \quad (4)$$

where $\odot$ stands for element-wise multiplication and $\widehat{\mathbf{S}}$ is the resulting estimated speech spectra. Time-domain signal $\hat{s}(n)$ is approximated from $\widehat{\mathbf{S}}(k,l)$ by means of the widely known Overlap and Add (OLA) method [13]. Here, it must be remarked the possible irreversibly of a modified STFT [14], the effects of which could be enhanced or alleviated by the window length and shift chosen.

## 3. Experimental framework

### 3.1. Speech database

DNN models have been trained and tested on a simulated noisy dataset that uses clean speech from the TIMIT database [15]. This customized dataset is intended for speech enhancement evaluation. It consists of utterances randomly selected from the 630 speakers included in the TIMIT database. Utterances from the same speaker are downsampled to 16 kHz and concatenated so that the resulting signal's length is between 7 and 11 seconds, as recommended for speech evaluation [16]. A total of 200, 50 and 50 length-adjusted utterances are used for the training, validation and testing datasets respectively. Same number of male and female speakers are included in all sets while no speaker is repeated across them. Additive noise is then added to each set at SNRs from -5 to 20 dB in steps of 5db. For the training and validation sets, the noise types considered are bus station, restaurant, car and quiet street. Same ones are included in the test set but extended with subway station, cafe, bus and busy street noises in order to check the performance under unseen noise types. All of them are own recorded and in-house curated noises. Noise recordings are split for training, validation and test sets so that no signal excerpt is repeated across them. In total, 4800 utterances are used for training ($\approx$ 13.3 hours), 1200 for validation and 2400 for testing.

### 3.2. DCCRN architecture

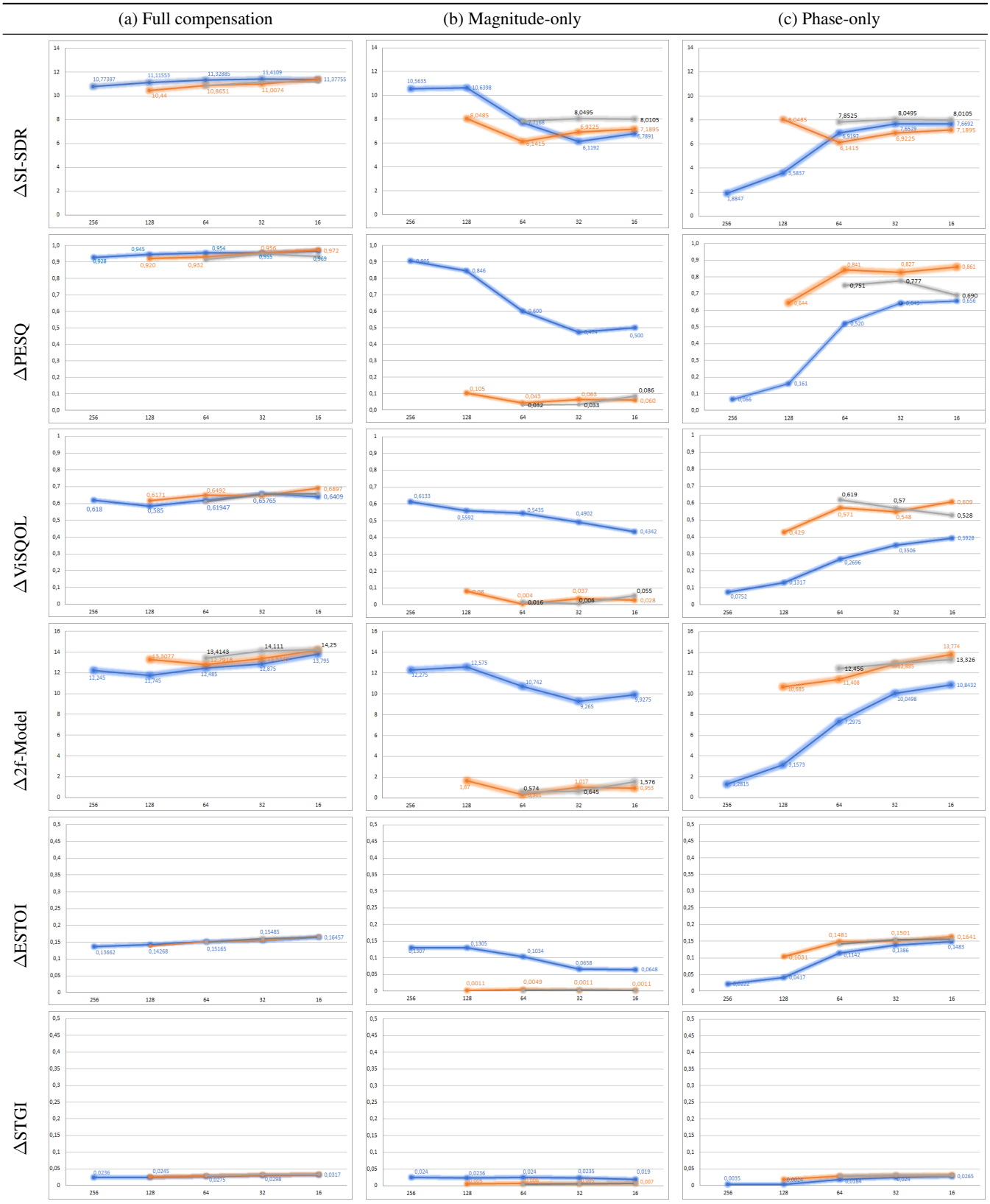
The neural network implemented in this paper is the Deep Complex Convolutional Recurrent Network (DCCRN) from [11]. This network consists of a causal Convolutional Encoder-Decoder (CED) architecture with two Long Short-Term Memory (LSTM) layers between the encoder and the decoder, so that the temporal dependencies can be modeled. The DCCRN is essentially an extension of a Convolutional Recurrent Network (CRN) that includes a common complex part computation instead of considering two isolated real and imaginary parts. In particular, the elements CNN, batch normalization layer in encoder/decoder and LSTM of CRN are complex-valued.

### 3.3. Training

The network is trained to estimate the mask $\mathbf{M}(k,l)$ which is then applied to enhance the noisy spectra. The Scale-Invariant Signal-Distortion Ratio (SI-SDR), which evaluates the distortion in the time domain ignoring general attenuation effects [17] [18], is used as loss function [17]. To this end, the output mask is applied and time-domain signal is synthesized by means of the OLA method (see Section 2.2). The resulting enhanced speech signal is then compared with the clean signal by using the aforementioned loss.

To address the aim of this paper, we have considered different window lengths of 512, 256 and 128 samples along with shifts of 256, 128, 64 and 16 samples. Since the sampling frequency is $F_s = 16kHz$, these window sizes correspond to 32 ms, 16 ms and 8 ms, respectively. The number of DFT bins is fixed to $K = 257$ (spectral symmetry of real time-signals is accounted here) independently of the window size (see Section 2.1). The window type chosen for the STFT analysis and synthesis is the square-root Hanning window, while the minimum frame overlap considered is 50% (e.g. a 256 shift cannot be

Table 1: *Performance results of the speech enhancement assessed by SDR, PESQ, ViSQOL, 2f-model, ESTOI and STGI metrics (gain with respect to noisy speech). First column (a) groups the results from a complete spectrum correction, the second one (b) those from to magnitude-only correction and the last one (c) from phase-only. Horizontal axis refers to the window shifts considered. Blue, orange and grey plots correspond to 512-, 256- and 128- sample window lengths, respectively. Confidence intervals at* 95% *are included as a color band for each line.*



used with a 256 window length).

DCCRNs are trained with the ADAM optimization algorithm [19] with a batch size is of 10 utterances, a learning rate of $10^{-3}$ and a dropout factor of $0.5$. We have also applied early stopping with a patience of 20 epochs (i.e. training is stopped after 20 epochs with no improvement in the validation set).

### 3.4. Evaluation

We have evaluated the resulting enhanced signals in terms of perceptual quality and intelligibility by using several objective metrics. All of these metrics consider the clean speech signal (i.e. intrusive metrics) to provide a score, while higher scores mean better quality and/or intelligibility.

#### *3.4.1. Perceptual quality*

Three different metrics has been used to quantify the perceptual quality of the speech signal in this paper: Perceptual Evaluation of Speech Quality (PESQ), Virtual Speech Quality Objective Listener (ViSQOL) and the 2f-model. The PESQ algorithm was proposed with the objective of providing an estimation of narrowband speech quality in [16]. Later on, it was extended to deal with wideband speech signals too. Alternatively, the ViSQOL metric was introduced in [20] and it is an objective measure centered in modeling the human speech quality perception of the signals. It provides an alternative and more recent model to assess the speech quality.

Finally, the 2f-model [21] evaluates the noticed quality of audio signals by means of an auditory model. As such, it evaluates the quality of audio signals, not only speech. It is based on two Model Output Variables (MOVs) of an audio quality assessment method called PEAQ [22]. The results from [23] suggest that this method is the best correlated with speech quality scores provided by real listeners in multiple databases.

#### *3.4.2. Intelligibility*

Intelligibility metrics correlate with the percentage of words that a native speaker would be able to correctly identify in the audio signal. In this paper, we have considered two intelligibility metrics: the Extended Short-Time Objective Intelligibility (ESTOI) [24] and the Spectro-Temporal Glimpsing Index (STGI).

The ESTOI metric is a well known monaural intelligibility prediction algorithm of speech contaminated by noise signals with a time-domain weighting. The STGI is a measure of intelligibility based on the detection of glimpses in short-time segments. While ESTOI is a fairly established method, STGI is a recent metric which shows higher correlation in recent subjective experiments [25]. Moreover, STGI can be employed not only with additive uncorrelated noise but also in a broad range of degradation conditions such as modulated noise, noise reduction processing or reverberation.

## 4. Experimental results

Performance results achieved by the DCCRN networks in terms of the aforementioned metrics are depicted in Table 1. These are expressed as a gain with respect to their noisy speech counterpart (in absolute terms) of the mean value across all tested SNRs. Each column shows the results obtained by a complete spectrum enhancement (2), and by magnitude (3) and by phase estimation only (4) across different window shifts and sizes (blue, orange and gray colors). Confidence intervals at $95\%$ are also shown as colored bands.

In general, similar tendencies are observed along all the metrics considered, although these seem more evident over perceptual ones (PESQ, ViSQOL and 2f-Model). As expected, performing a complete spectrum enhancement leads to better and more consistent results. Despite the complete spectrum enhancement does not seem severely affected by the window size and shift employed, slightly better results across all the metrics are obtained when shorter window shifts are considered. On the other hand, a completely different behavior is observed when magnitude-only or phase-only corrections are considered. Thus, as can be noted, phase corrections benefit from shorter window lengths, whereas worse results are obtained with magnitude-only compensation. This is particularly noticeable when window length reduces from 512 to 256 (32 to 16 ms). These results are also in accordance with the findings shown [12].

Regarding to window shift, it can be observed that, when only the spectra magnitude is compensated, a longer frame overlap (i.e. shorter shift) negatively affects speech enhancement. This could be explained by the limitations predicted in [14] with respect to the modified STFT inversion. Thus, inconsistencies across overlapped segments could make the time-domain signal estimation harder. On the contrary, decreasing window shift leads to improvements when phase-only is corrected. This effect is very noticeable with a long window size (32 ms) while seems erratic with short ones (8 ms). According to these results, we can argue that some kind of balance is achieved when both magnitude and phase are corrected (complete spectrum enhancement) under longer frame overlapping, leading to marginally better results.

## 5. Conclusions and future work

In this paper we have evaluated the effect of window size and shift in spectral domain DNN-based speech enhancement. A masking approach with a complex-valued DNN has been used for signal denoising. The time-frequency complex mask provided by the network has been applied to enhance magnitude-only, phase-only and the complete spectra independently, in order to assess the role of the STFT window size and shift under such approach. To this end, multiple perceptual quality and intelligibility objective metrics have been employed. From the results, it can be concluded that window size and shift parameters of the STFT transform play a notable role when training a neural network, particularly when only magnitude is corrected. Surprisingly, phase-only enhancement can provide competitive results with respect to magnitude-only correction when shorter frame shifts are chosen.

As future work, other complex-valued DNN networks will be tested to further confirm these conclusions. In addition, it would be interesting to replicate our analysis when a spectral-domain function loss is used for the DNN training instead of SI-SDR.

## 6. Acknowledgments

This work has been supported by the project PID2019-104206GB-I00 funded by MCIN/ AEI /10.13039/ 501100011033.

## 7. References

[1] P. C. Loizou, *Speech Enhancement – Theory and Practice*. Florida: Board, 2007.

[2] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and the development of the intel technique for improving speech intelligibility," in *Nicolet Scientific Corp.*, 1975.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

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

[5] J. M. Salavedra Molí, *Técnicas de Speech Enhancement considerando estadísticas de orden superior*. Polytechnic University of Catalonia, 1995.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *Interspeech*, 2018, pp. 1136–1140.

[8] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020.

[9] P. Mowlaee and J. K. Stahl, "Single-channel speech enhancement with correlated spectral components: Limits-potential," *Speech Communication*, vol. 121, pp. 58–69, 2020.

[10] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[11] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv:2008.00264*, 2020.

[12] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *arXiv:2203.16222*, 2022.

[13] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, p. 99–102, 1980.

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] J. Garofolo, L. Lamel, W. Fisher, J. F. adn D. Pellett, N. Dehlegren, and V. Zue, "The structure and format of the DARPA TIMIT CD-ROM Prototype," *in Proceedings of NIST*, pp. 1–9, 1988.

[16] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recomendation P.862*, 2000.

[17] S. Li, H. Liu, Y. Zhou, and Z. Luo, "A SI-SDR loss function based monaural source separation," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, vol. 1. IEEE, 2020, pp. 356–360.

[18] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of 3rd International Conference on Learning Representations*, 2015, pp. 1–13.

[20] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "VISQOL: The virtual speech quality objective listener," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[21] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 95–99.

[22] C. Colomes, C. Schmidmer, T. Thiede, and W. C. Treurniet, "Perceptual quality assessment for digital audio: PEAQ-the new ITU standard for objective measurement of the perceived audio quality," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999.

[23] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1530–1541, 2021.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] A. Edraki, W. Y. Chan, J. Jensen, and D. Fogerty, "A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction," in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association, 2021, pp. 2738–2742.