Complex Recurrent Variational Autoencoder for Speech Enhancement

Yuying Xie, Thomas Arildsen, Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark
yuxi@es.aau.dk, tari@its.aau.dk, zt@es.aau.dk

Abstract

Commonly-used methods in speech enhancement are based on short-time fourier transform (STFT) representation, in particular on the magnitude of the STFT. This is because phase is naturally unstructured and intractable, and magnitude has shown more importance in speech enhancement. Nevertheless, phase has shown its significance in some research and cannot be ignored. Complex neural networks, with their inherent advantage, provide a solution for complex spectrogram processing. Complex variational autoencoder (VAE), as an extension of vanilla VAE, has shown positive results in complex spectrogram representation. However, the existing work on complex VAE only uses linear layers and merely applies the model on direct spectra representation. This paper extends the linear complex VAE to a non-linear one. Furthermore, on account of the temporal property of speech signals, a complex recurrent VAE is proposed.

The proposed model has been applied on speech enhancement. As far as we know, it is the first time that a complex generative model is applied to speech enhancement. Experiments are based on the TIMIT dataset, while speech intelligibility and speech quality have been evaluated. The results show that, for speech enhancement, the proposed method has better performance on speech intelligibility and comparable performance on speech quality.

Index Terms: speech enhancement, complex recurrent neural network, variational autoencoder

1. Introduction

Complex neural networks provide a possible way to take full advantage of complex representations. As a pioneer work, Trabelsi et al. proposed the essential components of complex neural networks, and applied in complex convolutional feed-forward neural networks and convolutional LSTMs [1]. Wolter and Yao proposed a basic complex recurrent neural network (RNN) definition and complex gated recurrent unit (GRU) model, which shows excellent stability and convergence property.

As a very common and powerful analysis tool, short-time fourier transform (STFT) has been applied a lot in speech signal processing. According to its definition, STFT representation is naturally complex-valued, and can be expressed in rectangular form or polar form. Since the previous study [2] shows that magnitude is more important and structured than phase, typical deep learning methods in speech enhancement are based on magnitude spectrogram, using features like log-magnitude spectrogram [3]. In these works, only magnitude spectra are processed in the neural network, and phase from the input is kept in time signal re-synthesis. However, phase has shown its importance in speech signal processing [4].

Based on different representations of complex spectrogram in rectangular coordinate system and polar coordinate system, some works utilize magnitude spectrogram and phase [5], but most works use real and imaginary spectrogram [7][11]. Multiple strategies have been applied for complex spectrogram processing. In some existing work concatenated real and imaginary spectra are fed into a real-valued neural network [7][8], while others use a real-valued neural network to process real and imaginary spectra separately [9]. Since there is inherent relationship between real spectra and imaginary spectra, a sauer way for processing complex spectrograms is to use complex neural networks thoroughly, i.e. weights, biases and calculation should all be complex-valued. There also exists some work using complex-valued neural networks in signal speech processing. For instance, paper [10] applied complex feed-forward neural network in speech enhancement. Paper [11] proposed a complex model by combining a complex convolutional neural network with a real-valued recurrent neural network in an autoencoder framework.

Variational autoencoder (VAE) [12], as a generative model, has been applied widely in speech signal processing, including voice conversion [13], speech enhancement [14], etc. The vanilla VAE is composed by an encoder and a decoder, and assumes latent variable space and data space all follow a Gaussian distribution. As an extension of VAE theory, paper [15] mathematically derived complex variational autoencoder (VAE). All parameters and calculation in the complex VAE are complex-valued, and a complex normal distribution is used to model latent variable space and data space. Experiment results of complex spectrogram reconstruction in [15] are positive. However, the neural network used in the complex VAE is a linear neural network, which may degrade the performance of the model. Meanwhile, as an pilot study, paper [15] uses complex VAE only for direct speech representation without any specific application.

This motivates us to propose complex recurrent VAE. The contribution of this work mainly contains three parts. Firstly, we introduce non-linearity in the complex VAE model. According to basic theories in deep learning and mathematics, non-linearity will increase modelling ability of the model, especially in complicated non-linear situations. Secondly, based on the temporal property of speech signal, this work proposes complex recurrent VAE, using complex recurrent neural network to build complex VAE. Even though application of the proposed method can be broad, speech enhancement has been chosen in this work, as an exploration of using a complex generative model in speech signal processing.

The organization of this paper is as follows: Section2 illustrates the theory of complex neural networks, including complex feed-forward neural networks and complex GRU. Section3 presents the complex recurrent VAE. Experiment results and analysis are presented in Section4. Section5 concludes this paper.
2. Complex neural network

2.1. Complex feed-forward neural network

Compared to real-valued neural networks, all weights and biases in complex neural networks are complex-valued. For a complex-valued feed-forward neural network, if we use $\mathbf{W} = \Re(\mathbf{W}) + i\Im(\mathbf{W})$ to denote complex weight, $x = \Re(x) + i\Im(x)$ to denote input, and $\mathbf{b} = \Re(\mathbf{b}) + i\Im(\mathbf{b})$ to denote bias, then the output of a complex-valued dense layer is:

$$
o = \mathbf{W}x + \mathbf{b} = [\Re(\mathbf{W})\Re(x) - \Im(\mathbf{W})\Im(x) + \Re(\mathbf{b})] + i[\Im(\mathbf{W})\Re(x) + \Re(\mathbf{W})\Im(x) + \Im(\mathbf{b})] = \Re(\mathbf{o}) + i\Im(\mathbf{o}) \tag{1}
$$

2.2. Complex activation functions

Since complex values have different representations in rectangular form and polar form, complex-valued activation functions have a lot of different versions. The first proposed and mostly used complex-valued ReLU is modReLU [1], as shown in (2):

$$f_{\text{modReLU}}(z) = \text{ReLU}(|z| + b)e^{i\theta_z} = \text{ReLU}(|z| + b)\frac{z}{|z|} \tag{2}$$

where $z \in \mathbb{C}$, $|z|$ and $\theta_z$ are magnitude and phase of $z$, $b \in \mathbb{R}$ is a learnable parameter.

Besides, paper [15] proposed a complex-valued sigmoid function, called modSigmoid function, as shown in (3), in which $\sigma(\cdot)$ denotes real-valued sigmoid function:

$$f_{\text{modSigmoid}}(z) = \sigma(\alpha|z| + (1-\alpha)\Im(z)) \text{ with } \alpha \in [0, 1] \tag{3}$$

$\alpha$ used in this work equals 0.5.

2.3. Complex recurrent neural network

The definition of basic complex RNN is:

$$z_t = \mathbf{W}_h z_{t-1} + \mathbf{V} x_t + \mathbf{b} \tag{4}$$

$$h_t = f_a(z_t) \tag{5}$$

where $x_t \in \mathbb{C}^{n_x \times 1}$, $h_t \in \mathbb{C}^{n_h \times 1}$ denote input and hidden unit vector at time $t$ respectively, in which $n_x$, $n_h$ represent the dimension of $x_t$ and $h_t$. $\mathbf{W} \in \mathbb{C}^{n_h \times n_x}$, $\mathbf{V} \in \mathbb{C}^{n_h \times n_y}$ are hidden and input state transition matrices in order, while bias $\mathbf{b} \in \mathbb{C}^{n_h \times 1}$. $f_a(\cdot)$ is the point-wise nonlinear activation function.

Based on this, the complex GRU model is presented in the form of (6)-(7):

$$\tilde{z}_t = \mathbf{W}(g_r \odot h_{t-1}) + \mathbf{V} x_t + \mathbf{b} \tag{6}$$

$$h_t = g_z \odot f_s(\tilde{z}_t) + (1 - g_z) \odot h_{t-1} \tag{7}$$

$\odot$ represents Hadamard product. According to the experiment results and analyses in papers [16] and [17], for stability, $f_a(\cdot)$ in (7) is modReLU, and state transition matrices are unitary in this work. $g_r$ and $g_z$ represent reset gate and update gate, respectively, as shown in (8)-(9):

$$g_r = f_g(\tilde{z}_t), \quad z_t = \mathbf{W}_r h + \mathbf{V}_r x_t + \mathbf{b}_r \tag{8}$$

$$g_z = f_g(z_t), \quad z_t = \mathbf{W}_z h + \mathbf{V}_z x_t + \mathbf{b}_z \tag{9}$$

$f_g(\cdot)$ represents the modSigmoid function [3]. $\mathbf{W}_r, \mathbf{V}_r, \mathbf{W}_z, \mathbf{V}_z \in \mathbb{C}^{n_h \times n_h}$ are state-to-state transition matrices. $\mathbf{V}_r, \mathbf{V}_z \in \mathbb{C}^{n_h \times n_y}$ are input-to-state transition matrices. $\mathbf{b}_r, \mathbf{b}_z \in \mathbb{C}^{n_h}$ are biases.

2.4. Initialization

In this paper, the initialization of weights in both complex dense layers and complex GRU layers uses a uniform initializer like papers [16] and [18], i.e. weights are sampled from uniform distribution $U[-\delta, \delta]$, where $\delta = \sqrt{6}/(\tilde{n}_\text{in} + \tilde{n}_\text{out})$. $\tilde{n}_\text{in}$ and $\tilde{n}_\text{out}$ are the dimensions of input and output, respectively. All biases are initialized as 0, except $\mathbf{b}_r$ and $\mathbf{b}_z$ in GRU layers, for which 4 is used as initialization value like paper [19].

2.5. Gradient calculation

Assume function $f(z)$ is real-valued with complex variable $z$, i.e. $f : \mathbb{C} \rightarrow \mathbb{R}$, then $f(z)$ is non-holomorph long as $f(z) \neq 0$. Presume $z = x + iy$, $x, y \in \mathbb{R}$. Wirtinger calculus is used for partial derivative of non-holomorphic function:

$$\frac{\partial f}{\partial z} = \frac{1}{2}(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}) \tag{10}$$

$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2}(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}) \tag{11}$$

3. Complex recurrent variational autoencoder

3.1. Complex variational autoencoder

In complex-valued VAE framework, $x \in \mathbb{C}^{n_x}$, $z \in \mathbb{C}^{n_z}$ are used to represent data and latent variable respectively, in which $n_x$ and $n_z$ denote dimensions of data and latent variable. As with conventional VAE, the loss function of the complex-valued VAE is:

$$\log p(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\theta(x|z)}[\log p(x|z)] - KL(q_\theta(z|x)||p(z)) \tag{12}$$

$\phi$ and $\theta$ denote parameters in inference model and generative model individually, and $p$ and $q$ represent respectively the prior and posterior distribution.

Compared to real-valued VAE, complex VAE not only uses complex-valued parameters in the whole framework, but it also changes the assumption of data space and latent variable space distribution. A multivariate complex normal distribution is used to model latent variable and data in complex-valued VAE. If a complex random variable $\mathbf{h} \in \mathbb{C}^{D}$ follows a multivariate complex normal distribution, i.e., $\mathbf{h} \sim \mathcal{N}(\mathbf{a}, \Gamma, \mathbf{C})$, in which $\mathbf{a} \in \mathbb{C}^{D}$, $\Gamma \in \mathbb{C}^{D \times D}$, $\mathbf{C} \in \mathbb{C}^{D \times D}$ denote mean vector, covariance matrix and pseudo-covariance matrix in order, then the probability density of $\mathbf{h}$ can be written as:

$$p(\mathbf{h}) \triangleq \frac{1}{\pi^D \sqrt{\det(\Gamma) \det(\mathbf{C} - \mathbf{C}^H \Gamma^{-1} \mathbf{C})}} \exp\left\{-\frac{1}{2} \left[ \begin{array}{c} \mathbf{h} - \mathbf{a} \\ \mathbf{R} - \bar{\mathbf{a}} \end{array} \right]^H \left[ \begin{array}{c} \Gamma \\ \mathbf{C}^H \end{array} \right]^{-1} \left[ \begin{array}{c} \mathbf{h} - \mathbf{a} \\ \mathbf{R} - \bar{\mathbf{a}} \end{array} \right] \right\} \tag{13}$$

For data $x$, if we assume its prior distribution $p_0(x|z) = \mathcal{N}_c(x; \mathbf{a}, \mathbf{I}, 0)$, according to (13), we get:

$$\mathbb{E}_{q_\theta(x|z)}[\log p_0(x|z)] \approx -||x - \mathbf{a}||^2_2 + K \tag{14}$$

in which $K$ is a constant.

For latent variable $z$, we assume its posterior follows a multivariate complex normal distribution with diagonal covariance and pseudo-covariance matrices, i.e., $\mathcal{N}(z; \mu, \Delta(\sigma), \Delta(\delta))$, in which $\mu \in \mathbb{C}^{n_z}$, $\sigma \in \mathbb{R}_+^{n_z \times n_z}$, $\delta \in \mathbb{R}_+^{n_z \times n_z}$.

$$\mu = \mathbb{E}_q[z] = \mathbb{E}_q[z|\theta, \phi; x] = \mathbb{E}_q(z|x) = \mathbb{E}_{q_\theta(z|x)}[x] \tag{15}$$
C^\infty$, and $\Delta(\cdot)$ represents diagonal matrix. Assume the prior of $z$ follows a standard complex normal distribution, i.e., $p(z) \triangleq \mathcal{N}_c(0, I, 0)$. Then, the KL divergence in (12) can be written as:

$$
KL(q_\theta(x)||p(x)) = KL(\mathcal{N}_c(\mu, \Delta(\sigma), \Delta(\delta)||\mathcal{N}_c(0, I, 0))
$$

(15)

In real-valued VAE, the ‘re-parameterization trick’ introduced in [12] has been used to make back-propagation achievable from decoder to encoder. A similar trick is also needed in complex VAE. Under the assumption of the estimated latent variable $\tilde{z} \sim \mathcal{N}_c(z; \mu, \Delta(\sigma), \Delta(\delta))$, and the elements of latent variable $z$ are independent of each other, then

$$
\tilde{z} = \mu + k_\gamma \odot \epsilon_r + k_\sigma \odot \epsilon_i
$$

(16)

$$
k_\gamma = \frac{\sigma + \delta}{\sqrt{2\sigma + 2\Re(\delta)}}
$$

(17)

$$
k_\sigma = \frac{-\sqrt{\sigma^2 - |\delta|^2}}{\sqrt{2\sigma + 2\Re(\delta)}}
$$

(18)

$\epsilon_r$ and $\epsilon_i$ are random variables following a standard Gaussian distribution, i.e., $\epsilon_r, \epsilon_i \sim \mathcal{N}(0, 1)$. And $\sqrt{\cdot}$ in (17)-(18) means element-wise square root.

3.2. Proposed method

As a rising topic, using complex neural networks, especially complex generative model, for speech signal processing is at the starting stage. Paper [15] proposed complex VAE with linear layers for direct representation learning of complex spectrogram. However, basic theories in the deep learning and math tell us in order to learn something interesting, nonlinear calculation should be used in the framework. Since the recurrent neural network has natural advantage in time-series signal processing, this paper proposes complex recurrent VAE by using complex-valued GRU in a complex VAE framework. Figure 1 shows the structure of the complex recurrent VAE. Both encoder and decoder are composed of two layers, one is a complex-valued GRU layer as mentioned in 2.3, another is a complex-valued fully-connected layer. The encoder outputs parameters of the posterior of the latent variable $z$, i.e. mean $\mu$, covariance $\sigma$ and pseudo-covariance $\delta$. The latent variable $z$ resulting from the re-parameterization trick is fed into the decoder. As an exploration of using a complex generative model in speech signal processing, speech enhancement has been chosen to test the performance of the proposed framework.

4. Experiment

4.1. Dataset

The TIMIT dataset [19] is used for the following experiments. All 4620 utterances, 500 utterances and 192 utterances in the TIMIT training, development and core test sets are used for training, validation and test separately. The noise dataset, the same one used in paper [20], contains 6 different noise types: babble (BBL), cafeteria (CAF), street (STRI), speech shaped noise (SSN), bus (BUS) and pedestrian (PED). The first four noise types are used for training, development and test set generation as seen noise type, while the last two types are used as unseen noise type only in test set generation. To generate the noisy speech dataset, for each utterance in the training and development sets, SNR is selected uniformly at random from -10 dB to 10 dB with 1 dB as step size. For model evaluation, noisy speech data with SNR equal to {−6, −3, 0, 3, 6} dB has been produced separately. Clean speech signals are all normalized to have unit RMS power each before adding noise, and only the speech-active region of clean speech signals has been used to calculate the scale of noise to attain the target SNR. The sampling rate equals 16kHz. 200-dimensional complex spectrogram is used as input for the complex neural network, while the real-valued neural network utilizes the same dimension log-magnitude spectrogram.

4.2. Baseline

To compare a real-valued neural network to a complex-valued neural network, one baseline in this work is a real-valued recurrent VAE, denoted R-RVAE in the following. Both encoder and decoder of R-RVAE contain a GRU layer and a dense layer, like in Fig. 1 but real-valued. The prior of the latent variable $z$ is the standard Gaussian distribution, and the posterior of $z$ is $\mathcal{N}(\mu, \sigma^2)$, in which $\mu$ and $\sigma$ are from neural network. The reconstruction loss uses L2 loss.

Inspired by work in [10], complex-valued feed-forward neural network, denoted C-FFNN in the following, has been chosen as another baseline. For a fair comparison, the structure of C-FFNN contains four complex dense layers. ModReLU has been used as activation function after every layer except the output one. And a dropout of 0.2 is applied after each hidden layer. The objective function of C-FFNN is L2 loss.

4.3. Implementation details

The proposed complex recurrent VAE is denoted C-RVAE in the following. For all models except C-FFNN, the batch size equals 200, while the learning rate is $10^{-5}$. For C-RVAE, the GRU layers contain 512 units in both encoder and decoder, while the latent variable dimension equals 512. To have the same parameter number, R-RVAE uses 769-unit GRU layers in encoder and decoder, and the latent variable dimension also equals 769. For model C-FFNN, inspired by the settings in [10], batch size equals 4096, while the learning rate is 0.0002. To make a fair comparison, the first three layers of C-FFNN contain 512 units. For all models, training stops when the loss has not decreased for 50 epochs. We work in Tensorflow, and the gradient optimizer used here is Adam optimizer. The weights initialization of complex neural networks all follows section 2.4.

4.4. Results and discussion

Results of speech enhancement have been evaluated by extended short-time objective intelligibility (ESTOI) measure [21] (range from 0 to 1), scale-invariant signal-to-distortion ratio (SI-SDR) measure in dB [22] (range from $-\infty$ to $+\infty$), perceptual evaluation of speech quality (PESQ) measure [23] (range from 1 to 4.5) to evaluate speech intelligibility, signal quality and speech quality respectively. Higher score means better performance for all measures.

All models are broadly trained under 4 seen noise conditions and tested under all noise conditions. Tables 1 and 2 show the results of the different models under conditions of seen noise types and unseen noise types, respectively. The number in every cell is the evaluation results averaged over scores from test sets with 5 different SNRs. The best result in each row is highlighted by black bold font, and gray bold font indicates a result very close to the best one. The unprocessed noisy speech
In summary, in comparison to R-RVAE and C-FFNN, C-RVAE shows positive results on speech intelligibility (ESTOI) and signal quality (SI-SDR), and comparable performance on speech quality (PESQ), under both seen-noise conditions and unseen-noise conditions.

4.5. Ablation study

As an ablation study, to study the impact of complex recurrent neural network, we use a complex nonlinear fully-connected VAE for comparison. This model denotes C-VAE in the following. The structure of C-VAE is similar to Fig. 1 but using complex dense layers rather than complex GRU. The activation function used after the first dense layers in encoder and decoder is ModReLU. C-VAE has 512-unit complex dense layers in encoder and the first layer of decoder. The other settings are the same with C-RVAE.

Table 3: Comparison of C-VAE and C-RVAE

| Framework | Seen noise | Unseen noise |
|-----------|------------|--------------|
| ESTOI     | 0.454      | 0.509        |
| SI-SDR(dB)| -19.97     | -19.31       |
| PESQ      | 2.05       | 2.22         |

Averaged results for four seen noise type conditions and two unseen noise type conditions are shown in Table 3. C-RVAE shows its advantage on speech intelligibility in terms of ESTOI scores under both seen noise type and unseen noise conditions. For SI-SDR scores, C-VAE slightly outperforms C-RVAE. And both models show similar performance on speech quality evaluation. This indicates the complex recurrent neural network in the complex VAE framework does improve speech intelligibility in enhancement.

5. Conclusions

Motivated by the temporal property of speech signals, this work extended complex linear VAE to nonlinear, and proposed complex recurrent VAE. The proposed method has been applied in speech enhancement. ESTOI, SI-SDR and PESQ are utilized as evaluation methods for speech intelligibility, signal quality, and speech quality, respectively. Experiments show that, compared to real-valued recurrent VAE and complex feed-forward neural network, complex recurrent VAE has advantages on speech intelligibility and signal quality, and shows comparable performance on speech quality. An ablation study shows that recurrent layers in complex VAE improves intelligibility.

6. Acknowledgements

The work of Yuying Xie is supported by China Scholarship Council.
7. References

[1] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, “Deep complex networks,” in ICLR, 2018.

[2] 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.

[3] M. Kolbæk, Z.-H. Tan, and J. Jensen, “On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 283–295, 2019.

[4] K. Paiwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Communication, vol. 53, no. 4, pp. 465–494, 2011.

[5] J. Le Roux, “Phase-controlled sound transfer based on maximally-inconsistent spectrograms,” Signal, vol. 5, p. 10.

[6] N. Zheng and X.-L. Zhang, “Phase-aware speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 63–76, 2018.

[7] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single- and multi-channel speech enhancement and robust asr,” IEEE/ACM transactions on audio, speech, and language processing, vol. 28, pp. 1778–1787, 2020.

[8] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in MLSP, 2017.

[9] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, “A fully convolutional neural network for complex spectrogram processing in speech enhancement,” in ICASSP, 2019.

[10] A. Pandey and D. Wang, “Exploring deep complex networks for complex spectrogram enhancement,” in ICASSP, 2019.

[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,” in INTERSPEECH, 2020.

[12] D. Kingma and M. Welling, “Auto-encoding variational bayes,” in ICLR, 2014.

[13] Y. Xie, T. Arildsen, and Z.-H. Tan, “Disentangled speech representation learning based on factorized hierarchical variational autoencoder with self-supervised objective,” in MLSP, 2021.

[14] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “A recurrent variational autoencoder for speech enhancement,” in ICASSP, 2020.

[15] T. Nakashika, “Complex-valued variational autoencoder: A novel deep generative model for direct representation of complex spectra.” in INTERSPEECH, 2020.

[16] M. Wolter and A. Yao, “Complex gated recurrent neural networks,” in NeurIPS, 2018.

[17] M. Arjovsky, A. Shah, and Y. Bengio, “Unitary evolution recurrent neural networks,” in ICML, 2016.

[18] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in AISTATS, 2010.

[19] J. S. Garofolo, L. F. Lamel, W. Fisher, J. Fiscus, and D. Pallett, “Darpa tumit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1,” NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

[20] M. Kolbæk, Z.-H. Tan, and J. Jensen, “Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification,” in IEEE SLT Workshop, 2016.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.