Phase Unwrapping Based Speech Enhancement

Rui Cheng, Changchun Bao
Speech and Audio Signal Processing Laboratory, Faculty of Information Technology,
Beijing University of Technology, Beijing, 100124, China
E-mail: chengrui@emails.bjut.edu.cn, baochch@bjut.edu.cn

***Abstract*— Speech enhancement is a vital technology for reducing the noise in speech communication. Most speech enhancement methods only estimate magnitude spectrum of clean speech from noisy speech and combine noisy phase spectrum to recover the enhanced speech. In this paper, considering the importance of recovering the phase of clean speech in speech enhancement, a phase recovery method of speech is proposed by combining phase unwrapping and deep neural network (DNN). By integrating the recovered phase of clean speech into conventional magnitude enhancement methods, the performance is improved effectively. The verification is conducted by several types of noises at different signal-to-noise ratio (SNR) levels. The experimental results also confirmed that the recovered phase of clean speech resulted in an obvious improvement on the speech quality and intelligibility compared to the noisy phase.**

I. INTRODUCTION

In speech signal processing, speech enhancement is mainly used to remove various noises in speech communication. With the development of digital signal processing technology, many classic speech enhancement methods have emerged for decades. For example, spectral subtraction method [1], Wiener filtering method [2], subspace method [3], statistical model-based method [4] and so on. These methods are classic methods in the field of speech enhancement because of their simple principle and easier implementation. However, these methods are based on a certain degree of ideal hypothesis, such as using the phase of noisy speech as the enhanced speech phase in speech reconstruction. So, they all ignore the importance of phase information [5, 6] in speech signal. In this way, the error of noisy phase to clean phase can be ignored when the SNR level is high. However, if the SNR is at a relatively low level, the phase error will cause a certain amount of negative impact for the enhanced speech.

In recent years, the researchers have paid more attention to the phase estimation of speech in speech enhancement, such as phase-locked loop-based phase estimation [7], harmonic enhancement-based phase reconstruction [8] and so on. Some results have shown that effective phase recovery of speech could considerably improve speech quality [9]. With the development of deep learning, the DNN have been extensively applied into speech enhancement. Some people have tried to combine DNN with traditional methods to reconstruct the phase of speech in speech enhancement system. For example, Magron proposed a phase constraint method based on sinusoidal-model and neural network for speech separation [10], Wang proposed a neural network and unfolded iterative-based phase reconstruction method [11]. These methods were proved that the importance of phase information in speech enhancement. In addition, some scholars have proposed to use DNN to estimate the phase indirectly. For example, Wang proposed a DNN-based complex ideal ratio masking (cIRM) method [12]. In this method, the magnitude and phase of speech were converted into a complex form and the real and imaginary parts of the short-time Fourier transform (STFT) of noisy speech were used as targets of the DNN. Another DNN-based example for indirect estimation of phase is to treat phase estimation problem as a classification problem by discretizing phase values and assigning class indexes [13]. Thus, it can be seen that phase recovery of speech may be a breakthrough of speech enhancement, that is, the phase estimation is indispensable for speech enhancement.

In this paper, a speech enhancement method is proposed through phase unwrapping based on Cellular-Automata principle. In this method, noisy speech phase is first unwrapped so that the phase value is not limited to the interval from $-\pi$ to π. Secondly, the DNN is used to estimate phase from the unwrapped noisy phase for getting the corresponding speech phase, and this speech phase is re-wrapped between $-\pi$ and π. Finally, combined with the enhanced magnitude spectrum, the speech signal is obtained by combing the recovered speech phase.

The rest of this paper is organized as follows. In Section 2, the classic magnitude enhancement methods are described. In Section 3, the details of the proposed phase recovery method are discussed. Experiments and results are provided in Section 4, and the conclusions are given in Section 5.

II. REVIEW OF MAGNITUDE ENHANCEMENT

In the most speech enhancement methods including unsupervised and supervised methods, the short-term magnitude spectrum of speech is considered, whereas the equally important short-term phase spectrum is ignored. Whether it is the unsupervised Wiener filtering-based speech enhancement method [14] or the supervised ideal ratio mask (IRM)-based speech enhancement method [15], the following transfer function is adopted to reduce or mask noise,

$$H(t,k)=\frac{|\mathbf{S}(t,k)|^2}{|\mathbf{S}(t,k)|^2+|\mathbf{N}(t,k)|^2} \quad (1)$$

where t is the frame index and k is the frequency bin. $|\mathbf{S}(t, k)|^2$ and $|\mathbf{N}(t, k)|^2$ are the estimated power spectra of speech and noise at the $(t, k)^{th}$ time-frequency (T-F) bin, respectively.

In the unsupervised method, the $H(t, k)$ is used as a gain function to obtain magnitude spectrum of the enhanced speech, while in the supervised method, the $H(t, k)$ is used as the training target of neural network to obtain an IRM for masking the noise. By combining with noisy speech phase of each frequency bin, the enhanced speech can be obtained by doing an inverse STFT. In order to improve the performance of the unsupervised and supervised methods, in this paper, the estimated phase of speech is embedded into enhanced magnitude spectrum. The details are given in the next section.

III. Proposed Phase Recovery Method

A. Cellular-Automata-Based Phase Unwrapping

Generally, the STFT of speech signal makes the values of phase spectrum wrapped between -π and π [6]. Due to the existence of the wrapping phenomenon, the phase spectrum does not have specific structure as the magnitude spectrum again, so it cannot be directly estimated by the neural network. Thus, we have to finish a phase unwrapping to overcome this restriction. The cellular automata gives us an enlightenment. Cellular automata is a simply and discretely mathematical system that can exhibit complex behavior resulting from collective effects of a large number of cells, each of which evolves in discrete time steps according to rather simple local neighborhood rules [16]. So, we can combine this principle with two-dimensional phase spectrum of speech signal, that is, the phase unwrapping can be implemented by following a correction to remove restriction from -π to π. In this paper, a method based on "Strength-of-Vote" [17] with local neighborhood correction is adopted for the phase unwrapping, its details can be obtained by Algorithm 1. Given a frame of phase spectrum containing a finite number of frequency bins, each frequency bin represents a principal value of phase information between -π and π. All frequency bins, derived from T-F transformation (e.g. discrete Fourier transform), in a frame, are updated simultaneously according to the phase difference with adjacent frequency bins to obtain the unwrapped phase by Algorithm 1.

Through the iteration Algorithm 1 of cellular-automata-based phase unwrapping, the constrained phase values are expanded into ones that do not have a limitation from -π to π according to the "Strength-of-Vote" rule. The phase spectrum gets into a specific structure that is suitable for the learning of neural network. The more the iteration number, the larger the dynamic range of unwrapped phase after expansion. For the feasibility of neural network training and the consideration of computational complexity, we set the number of local iteration and global iteration to 20 and 20, respectively. Therefore, obtaining the unwrapped phase information with specific structure and its range of the values is unlimited and deterministic, which are beneficial for neural network.

Algorithm 1: Cellular-Automata-Based Phase Unwrapping

Input: a frame of true phase vector, $\boldsymbol{\theta} = [\theta_{1,0}(t, 0), \ldots, \theta_{1,0}(t, N\text{-}1)]$
Output: a frame of unwrapped phase vector, $\boldsymbol{\theta}_u = [\theta_{m,n}(t, 0), \ldots, \theta_{m,n}(t, N\text{-}1)]$

m: global iteration number n: local iteration number
N: the number of frequency bins
$\theta_{i,j}(t, k)$: the phase value at the $(t, k)^{th}$ T-F bin of the i^{th} global iteration and the j^{th} local iteration
$\theta_{1,0}(t, k)$: the initial value of true phase at the $(t, k)^{th}$ T-F bin

```
# Subfunction: frequency bin-based iteration
BinUpdate(θ_{1,0}(t, k)):
  # Global iteration
  for i=1 to m
    # Local iteration
    for j=1 to n
      # Phase differences with adjacent frequency bins Δθ_left(t, k) and Δθ_right(t, k),
      Δθ_left(t, k)=θ_{i,j-1}(t, k)-θ_{i,j-1}(t, k-1)   # if k=0, Δθ_left(t, k)=0
      Δθ_right(t, k)=θ_{i,j-1}(t, k)-θ_{i,j-1}(t, k+1)  # if k=N-1, Δθ_right(t, k)=0
      # Strength-of-Vote with adjacent frequency bins N_left(t, k) and N_right(t, k)
      Δθ*_left(t, k)=Δθ_left(t, k)+2πN_left(t, k)
      Δθ*_right(t, k)=Δθ_left(t, k)+2πN_right(t, k)
      let Δθ*_left(t, k)∈[-π, π], get N_left(t, k)
      let Δθ*_right(t, k)∈[-π, π], get N_right(t, k)
      # Update the value at the (t, k)^th T-F bin
      if N_left(t, k)=0 && N_right(t, k)=0
        θ_{i,j}(t, k)=θ_{i,j-1}(t, k)
      if N_left(t, k)+N_right(t, k)≥0 && (N_left(t, k)≠0 || N_right(t, k)≠0)
        θ_{i,j}(t, k)=θ_{i,j-1}(t, k)+2π
      if N_left(t, k)+N_right(t, k)<0
        θ_{i,j}(t, k)=θ_{i,j-1}(t, k)-2π
    end
    # Find average value of two results from the last two local iterations
    θ_{i,n}(t, k)=[θ_{i,n-1}(t, k)+θ_{i,n}(t, k)]/2
    # Preparing for the next global iteration
    θ_{i+1,0}(t, k)=θ_{i,n}(t, k)
  end
  # Return the global iteration result
  return θ_{m,n}(t, k)

# Main function: frame-based iteration
FrameUpdate(θ):
  do BinUpdate() on each frequency bin k consisting of vector θ
     simultaneously to get θ_u
  return θ_u
```

B. DNN-Based Unwrapped Phase Estimation

Since the unwrapped phase spectrum has the specific structure, the DNN can be used for estimating the unwrapped speech phase. In the training stage of the DNN, the unwrapped noisy phase $\theta_{Yu}(t, k)$ is used as input feature, and the ratio mask of the unwrapped phase, $IRM_{phase}(t, k)$, is employed as the training target for more accurate estimation. The training target $IRM_{phase}(t, k)$ is defined as follows:

$$IRM_{phase}(t,k)=\frac{\theta_{Yu}(t,k)}{\theta_{Su}(t,k)} \quad (2)$$

where $\theta_{Su}(t, k)$ is the unwrapped phase of clean speech. The Mean Squared Error (MSE) is chosen as the objective function.

In the enhancement stage, the unwrapped phase of noisy speech is fed into the trained DNN, named as the UPDNN because it is used to estimate the unwrapped phase (UP), to obtain the $IRM_{phase}(t, k)$. By combing with the input feature $\theta_{Yu}(t, k)$, the estimated unwrapped phase $\theta^*_{Su}(t, k)$ of clean speech is given by

$$\theta^*_{Su}(t,k) = \frac{\theta_{Yu}(t,k)}{IRM_{phase}(t,k)} \tag{3}$$

C. Phase Reconstruction

The range of the unwrapped clean phase value estimated by the UPDNN is still unrestricted. But for speech reconstruction, we need to re-wrap the phase to the range of [-π, π]. According to Itoh's analysis of phase wrapping [18], the phase principal values can be regarded as the results of applying a wrapping operator W on the estimated unwrapped phase, i.e., the estimated re-wrapped phase $\theta^*_S(t, k)$ of the clean speech is expressed as follows:

$$\theta^*_S(t,k) = \mathrm{W}[\theta^*_{Su}(t,k)] = \theta^*_{Su}(t,k) + 2\pi l(t,k) \tag{4}$$

where $l(t, k)$ is an integer matrix that makes $\theta^*_S(t, k)$ meet the following condition:

$$-\pi \le \theta^*_S(t,k) \le \pi \tag{5}$$

Based on Eq. (3)-(5), the speech phase used for speech enhancement is recovered by re-wrapping the estimated unwrapped phase of speech.

D. Speech Enhancement Based on Phase Recovery

The block diagram of the proposed method is shown in Figure 1. In the training stage, the unwrapped phase of noisy speech is normalized, and the relationship between this normalized unwrapped phase and $IRM_{phase}(t, k)$ is mapped to train the UPDNN. In enhancement stage, the phase of noisy phase is extracted to get the unwrapped phase. The unwrapped phase is normalized and inputted to the UPDNN. Combining the output of the UPDNN with the unwrapped noisy phase, the phase reconstruction of speech is performed for speech enhancement.

By combining the recovered phase with the enhanced magnitude spectrum obtained by Wiener filtering-based method [14] and IRM-based method [15], the enhanced speech is obtained by performing an inverse STFT

IV. Experimental Setup and Evaluation Metrics

A. Experimental Setup

In the experiments, the TIMIT [19] corpus is used to evaluate performance of the proposed method. 4620 sentences from different speakers in the TIMIT corpus are used as the clean speech. 102 noise types including 100 environmental

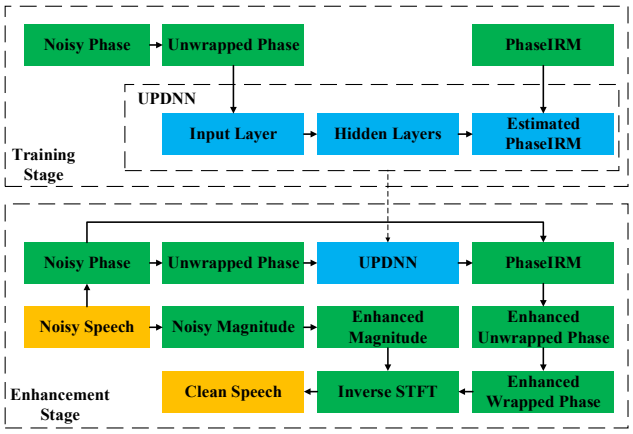


Figure 1: *Block diagram of the proposed speech enhancement method.*

noises [20], *Babble* and *F16* noise [21] are used as the training set of noise. Speech signal is down-sampled to 8 kHz. Moreover, the speech and noise are artificially mixed to obtain noisy speech at four different SNR levels from -5 to 10dB spaced by 5dB. 9216 noisy speech sentences are selected randomly to build an 8-hour training set.

In enhancement stage, another 201 sentences from the TIMIT test set are chosen randomly as clean speech. Babble and F16 noises in the training set and other two noises Factory and White [21] outside the training set are combined with clean speech to get noisy speech for the test. The noisy speech is formed at three different SNR levels ranging from -5 to 5dB in terms of 5dB step. Additionally, the length of each sentence in the test set is about 10 minutes.

For the training of neural network, 129-dimensional normalized speech magnitude spectrum is used as input features for the IRM-based supervised method to enhance magnitude spectrum, and the normalized unwrapped phase signal from the 129-dimensional phase spectrum is used as input features for the UPDNN to enhance the unwrapped phase, which are extracted using a window length of 32ms (256 samples) and a frame shift of 16ms (128 samples). The structure of the proposed UPDNN is composed of three hidden layers and each layer contains 512 neurons with rectified linear unit (ReLU) [22] as activation function. The Adaptive Moment Estimation (Adam) algorithm [23] is chosen to update the parameters of the neural network. For the training target $IRM_{phase}(t, k)$, as indicated in Eq. (2), since the numerator and denominator components of $IRM_{phase}(t, k)$ are obtained by the same unwrapping rule and their ranges are similar, the ratio is in a very small range of values and can be used as training target of neural networks without normalization.

B. Evaluation Metrics

The performance of the enhanced speech is evaluated by perceptual evaluation of speech quality (PESQ) [24], short-time objective intelligibility (STOI) [25] and phase error (PE). The PESQ is used to measure subjective quality of speech, while STOI is utilized to test intelligibility of speech. The PE

reflects the error between the recovered phase and clean phase. The PE is defined by

$$PE = \frac{1}{MN}\sum_{t=1}^{M}\sum_{k=1}^{N}\left|\theta_S(t,k) - \theta_S^*(t,k)\right| \quad (6)$$

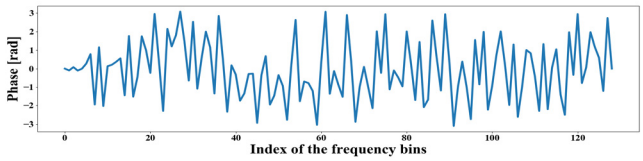
where M and N are the numbers of frames and frequency bins, respectively. $\theta_S(t, k)$ and $\theta_S^*(t, k)$ denote phase spectra of clean speech and the recovered speech at the $(t, k)^{th}$ T-F bin.

V. Experimental Results and Discussion

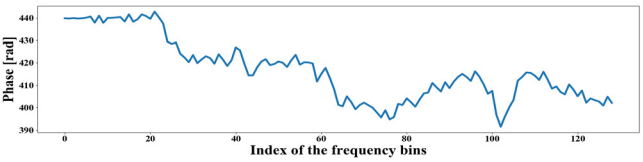
A. Feasibility of Phase Unwrapping

For phase recovery based on phase unwrapping, the loss of phase information must be within a sufficiently small error. So, we need to verify whether the proposed phase unwrapping and phase reconstruction can be a transform pair.

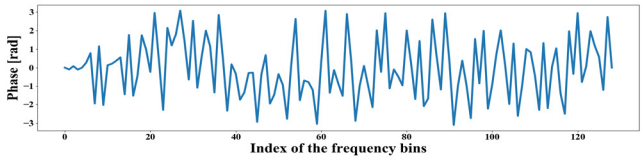
Figure 2 depicts an example of a random frame of clean speech phase in the TIMIT corpus, it includes the true clean phase, the unwrapped clean phase and the reconstructed clean phase directly obtained from the unwrapped clean phase without DNN estimation process. It can be clearly seen from Figure 2 (a) and Figure 2 (c) that the reconstructed phase is almost the same as true one. Table 1 shows the results of speech quality and intelligibility of clean speech with true clean phase and reconstructed clean phase directly obtained from the unwrapping without DNN estimation. This implies that the speech obtained by the reconstructed clean phase is nearly same as clean speech within a very small error range. Therefore, the transform pair is valid within a tolerant error and can be used for phase transformation and estimation.



(a) The true speech phase



(b) The unwrapped speech phase



(c) The reconstructed speech phase

Figure 2: *A frame of clean speech phase. (a) The true speech phase; (b) The unwrapped speech phase; (c) The reconstructed speech phase directly obtained from the unwrapped clean phase without DNN estimation process.*

Table 1: *The PESQ and STOI results of clean speech with true speech phase and the reconstucted speech phase directly obtained from unwrapping without DNN estimation process.*

Speech	PESQ	STOI
CleanMagnitude+CleanPhase	4.5	1
CleanMagnitude+ReconstuctedPhase	4.45843	0.99998

B. Speech Quality and Intelligibility

Table 2 presents the average PESQ and STOI results of the Wiener filtering-based method (referred to WF) [14] and the IRM-based supervised method (referred to IRMDNN) [15] with noisy phase (NP) and the proposed phase (UP) at three different SNR levels (-5dB, 0dB, 5dB). In four noises (*Babble*, *F16*, *Factory* and *White*) tests, the *Babble* and *F16* are the trained noises, while the *Factory* and *White* are the unseen noises.

From the PESQ results, it is clear that although the Wiener filtering-based method and IRMDNN-based method achieved great improvement comparing with noisy speech, the scores of the proposed method is still superior to the reference method. The increase in PESQ scores of the proposed method is due to effective estimation of phase compared with Wiener filtering-based method and the IRM-based supervised method. They help the enhanced speech to have more accurate phase information for the reconstruction of speech, especially in the case of low SNR level. Because, in this case, employing the phase of noisy speech will lead to greater errors.

From the STOI results, we can find that the Wiener filtering-based speech enhancement method does not have an obvious improvement, even the intelligibility is reduced compared with noisy speech. This is due to Wiener filtering method can introduce speech distortion and music noise. By combing the recovered phase, even so, there still have a certain degree of improvement. For the IRM-based supervised method, the proposed method has considerable improvement in intelligibility compared to noisy phase, especially in the case of low SNR. This indicates that the recovered phase contributes higher intelligibility to speech than the noisy phase at lower SNR levels.

C. Phase Error

Based on the definition of phase error Eq. (6) in this work, the results of error analysis on the test set are given in Figure 3. In Figure 3, the blue dashed line shows the average phase error results with noisy phase and the black solid line shows the average phase error results with the recovered phase. We can see that combing noisy phase into magnitude spectrum results in the large errors due to the overwhelming of speech by noise. However, the proposed phase recovery method reduces phase error so that the more accurate phase is obtained for speech reconstruction, especially in the case of lower SNR level, where the noisy phase has large error. Therefore, compared with the noisy phase, the proposed phase recovery method can effectively reduce the error of clean speech phase.

Table 2: *Comparison of the average PESQ and STOI results.*

Metrics		PESQ			STOI		
SNR (dB)		-5	0	5	-5	0	5
Methods	Noisy Speech	1.5751	1.8441	2.1376	0.5112	0.6329	0.7484
	WF+NP	1.6449	2.0335	2.4534	0.4776	0.6194	0.7450
	WF+UP	1.7132	2.0994	2.4958	0.4866	0.6259	0.7471
	IRMDNN+NP	1.8069	2.2313	2.6207	0.5761	0.7104	0.8075
	IRMDNN+UP	**1.8986**	**2.2925**	**2.6442**	**0.5823**	**0.7130**	**0.8098**

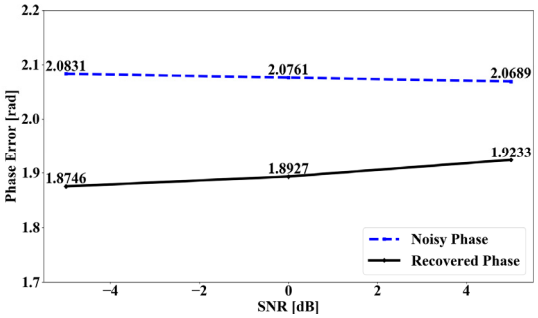


Figure 3: *The average PE results of noisy and the recovered phases.*

From the above quality and intelligibility score and error analysis, we can see that the proposed phase recovery method can be combined with the existing unsupervised and supervised magnitude-based speech enhancement methods to further reduce the error caused by the use of noisy speech phase and improve the enhanced speech quality and intelligibility.

VI. Conclusions

In this paper, a new method for speech phase recovery was proposed based on the cellular-automata and deep neural network. The cellular-automata-based method was used to obtain the unwrapped phase without limitation from -π to π, in which the DNN was used to estimate the unwrapped phase. The enhanced phase was reconstructed by re-wrapped the estimated unwrapped phase, and combined with magnitude enhancement methods to reconstruct speech. In comparison with conventional speech enhancement methods that used noisy phase as the enhanced phase, the proposed method can significantly improve the enhanced speech quality and intelligibility. In the future, we will further improve the method of phase estimation and combine it with other neural network with stronger fitting ability to further improve the accuracy of phase estimation, so as to obtain better quality and better intelligibility speech.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant No. 61831019, No. 61471014 and No. 61231015).

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.

[3] F. Asano, S. Hayamizu, T. Yamada and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE/ACM Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 497-507, Sep 2000.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral magnitude estimator," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec 1984.

[5] P. Mowlaee, R. Saeidi, and Y. Stylianou, "INTERSPEECH 2014 Special Session: Phase Importance in Speech Processing Applications", *Proc. Interspeech*, pp. 1623-1627, 2014.

[6] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55-66, March 2015.

[7] P. Pallavi and R. Rao, "Phase-Locked Loop (PLL) Based Phase Estimation in Single Channel Speech Enhancement", *Proc. Interspeech*, pp. 1161-1164, 2018.

[8] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, "Single-Channel Speech Enhancement With Phase Reconstruction Based on Phase Distortion Averaging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1559-1569, Sept. 2018.

[9] PALIWAL K, WÓJCICKI K, SHANNON B. "The Importance of Phase in Speech Enhancement". *Speech Communication*, vol.53, no. 4, pp.465-494, 2011.

[10] P. Magron, K. Drossos, S. I. Mimilakis, and T. Virtanen, "Reducing Interference with Phase Recovery in DNN-based Monaural Singing Voice Separation", *Proc. Interspeech*, pp. 332-336, 2018.

[11] Z. Wang, J. L. Roux, D. Wang and J. R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction", *Proc. Interspeech*, pp. 2708-2712, 2018.

[12] D. S. Williamson, Y. Wang and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483-492, March 2016.

[13] N. Takahashi, P. Agrawal, N. Goswami and Y. Mitsufuji, "PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation", *Proc. Interspeech*, pp. 2713-2717, 2018.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude

estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, December 1984.
[15] 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, Oct. 2018.
[16] D. C. Ghiglia, G. A. Mastin, and L. A. Romero, "Cellular-automata method for phase unwrapping," *J. Opt. Soc. Am. A*, vol. 4, no. 1, pp. 267-280, 1987.
[17] S. Wolfram, "Cellular Automata", *Los Alamos Science*, vol. 9, pp. 2-21, 1983.
[18] K. Itoh, "Analysis of the phase unwrapping algorithm," *Appl. Opt*, vol. 21, no. 4, pp. 2470-2470, 1982.
[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.
[20] G.H, "100 nonspeech environmental sounds," 2014.
[21] A. Varga, and H. J. 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] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp. 1026-1034, 2015.
[23] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, pp. 749-752, 2001.
[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.