BINAURAL SPEECH ENHANCEMENT USING COMPLEX CONVOLUTIONAL RECURRENT NETWORKS

Vikas Tokala[1], *Eric Grinstein*[1], *Mike Brookes*[1], *Simon Doclo*[2], *Jesper Jensen*[3,4], *Patrick A. Naylor* [1*]

[1]Department of Electrical and Electronic Engineering, Imperial College London, UK
[2] Department of Medical Physics and Acoustics, University of Oldenburg, Germany.
[3] Demant A/S, Smørum, Denmark.
[4] Department of Electronic Systems, Aalborg University, Denmark

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476738

ABSTRACT

From hearing aids to augmented and virtual reality devices, binaural speech enhancement algorithms have been established as state-of-the-art techniques to improve speech intelligibility and listening comfort. In this paper, we present an end-to-end binaural speech enhancement method using a complex recurrent convolutional network with an encoder-decoder architecture and a complex LSTM recurrent block placed between the encoder and decoder. A loss function that focuses on the preservation of spatial information in addition to speech intelligibility improvement and noise reduction is introduced. The network estimates individual complex ratio masks for the left and right-ear channels of a binaural hearing device in the time-frequency domain. We show that, compared to other baseline algorithms, the proposed method significantly improves the estimated speech intelligibility and reduces the noise while preserving the spatial information of the binaural signals in acoustic situations with a single target speaker and isotropic noise of various types.

Index Terms— Binaural speech enhancement, Complex Convolutional Neural Networks, recurrent networks, interaural cues, noise reduction.

1. INTRODUCTION

Binaural speech enhancement has gained significant attention as a state-of-the-art approach for enhancing speech in various applications, including hearing aids and augmented/virtual reality devices [1, 2]. Binaural signals preserve the spatial characteristics of sounds which help listeners in noisy acoustic environments achieve better speech intelligibility and accurate sound source localization [3]. The fundamental spatial cues that help in localizing sounds and improving intelligibility are Interaural Level Differences (ILD) and Interaural Phase Differences (IPD) or Interaural Time Differences (ITD) [4]. Previously proposed methods for binaural speech enhancement include multichannel Wiener filters [5, 6], beamforming-based [1], and mask-based enhancement methods [7, 8]. Binaural speech separation using time-domain Convolutional Encoder-Decoder (CED) was proposed in [9]. Recent advancements in deep learning techniques have led to remarkable improvements in monaural speech enhancement. These methods, whether applied in the time domain [10, 11] or the Time-Frequency (TF) domain [12–14], demonstrate impressive results.

*This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956369 and the UK Engineering and Physical Sciences Research Council [grant number EP/S035842/1]

Spectrograms are commonly used as input to networks in the TF domain for speech enhancement [7, 14, 15]. Many methods in this domain focus on enhancing only the magnitude of the spectrogram while using the noisy phase information for reconstructing the enhanced speech signal [7, 16]. To address optimal phase estimation and signal reconstruction, some approaches jointly estimate both magnitude and phase by utilizing complex-valued spectrograms. These methods have shown promising results and can outperform real-valued networks in monaural speech enhancement [15,16].

The Convolutional Recurrent Network (CRN) introduced in [13] employed a Convolutional Encoder-Decoder (CED) architecture with Long short-term memory (LSTM) blocks placed in between the encoder and decoder. In [15], a deep complex CRN was trained to optimize the Scale Invariant SNR (SI-SNR) for monaural speech signals. However, using a similar approach for binaural signals could be damaging to the interaural cues. More specifically, for the case of binaural signals, phase information is vital for preserving the IPD values and the enhanced signals should preserve level differences as the noisy signal to retain ILD. While the model may effectively reduce noise and enhance speech intelligibility, modifying the level and phase information could potentially impact the spatial information of the target, leading to a compromised ability for localization and spatial awareness [4, 17]. In [18], a complex convolutional attention-based transformer network has been proposed which uses a similar loss function based on interaural cues.

Our proposed method, Binaural Complex Convolutional Recurrent Network (BCCRN), uses a complex-valued CED-based recurrent network for binaural speech enhancement and introduces terms in the loss function to improve speech intelligibility while preserving the interaural cues based on human perception of the target speech signal with a smaller complex recurrent network compared to [18].

2. BCCRN MODEL ARCHITECTURE

The proposed BCCRN is trained to estimate individual Complex Ratio Mask (CRM) for each channel. The block diagram of the model architecture is shown in Fig. 1. The Short Time Fourier Transform (STFT) blocks transform the input signal into the TF domain. The encoder block is made of 6 complex convolutional layers with Parametric Rectified Linear Unit (PReLU) activation and employs batch normalization. Separate encoder and decoder blocks are used for the left and right-ear channels to estimate the individual CRMs. The decoder block consists of 6 complex convolutional layers which are symmetric in design to the layers in the encoder to reconstruct the signal. The encoder extracts high-level features from the input spectrograms and reduces the resolution of the input. The decoder

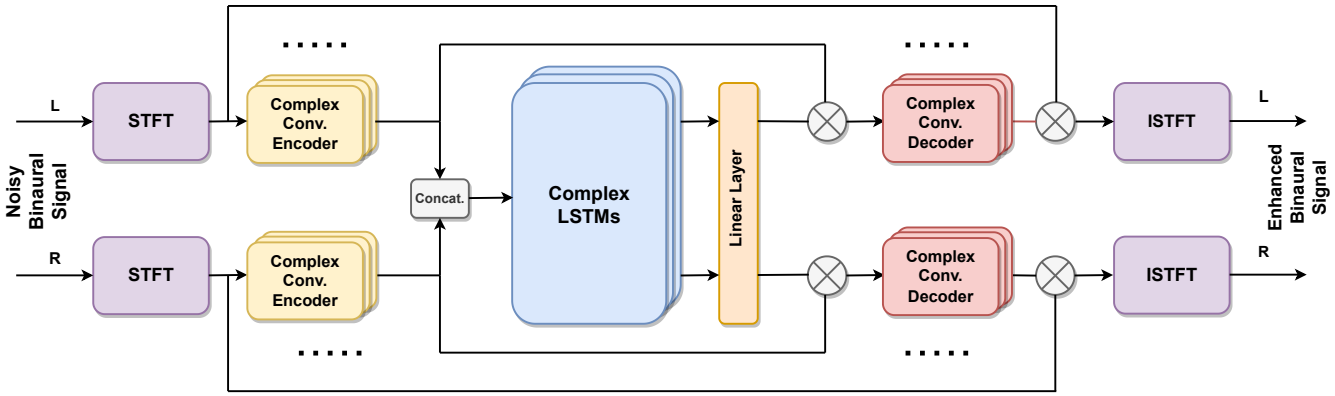

Fig. 1: Model architecture of Binaural Complex Convolutional Recurrent Network (BCCRN).

rebuilds the low-resolution features back to the original size. The encoded information from the left and right encoder blocks is concatenated and provided to the complex LSTM block. The LSTM layers are designed to model the frequency dependencies. Skip connections are placed between the encoder and decoder layers based on the CRN architecture which improves the information flow and facilitates network optimization [13]. The individual decoders output individual CRMs that are applied to the left and right channels of the noisy binaural signal for enhancement. The Inverse STFT (ISTFT) blocks transform the signal back into the time domain.

3. SIGNAL MODEL AND LOSS FUNCTION

The noisy time-domain input signal for the right channel, y_R, is given by

$$y_R(t) = s_R(t) + v_R(t), \tag{1}$$

where s_R is the anechoic clean speech signal, v_R is the noise and t is the time index. The STFT is used to transform the signals into the TF domain and the respective TF representations are $Y_R(k,\ell)$, $S_R(k,\ell)$ and $V_R(k,\ell)$ with k and ℓ being the frequency and time frame indices respectively. The left channel is described similarly with an L subscript. For brevity, the L and R indices are omitted from the remainder of this paper. The estimated CRM, $M(k,\ell)$ is applied to the noisy signal $Y(k,\ell)$ to obtain the enhanced speech signal $\hat{S}$. Individual channels are enhanced by applying the estimated CRM, $(M_r + jM_i)$ to the complex-valued noisy signal $(Y_r + jY_i)$ in the TF domain (omitting k and ℓ indices),

$$\hat{S}_r + j\hat{S}_i = (M_r + jM_i) \cdot (Y_r + jY_i)\,, \tag{2}$$

where r and i indicate the real and imaginary parts. The computed CRM is given by

$$M_r + jM_i = \frac{\hat{S}_r + j\hat{S}_i}{Y_r + jY_i} = \frac{Y_r\hat{S}_r + Y_i\hat{S}_i}{Y_r^2 + Y_i^2} + j\frac{Y_r\hat{S}_i - Y_i\hat{S}_r}{Y_r^2 + Y_i^2}. \tag{3}$$

3.1. Loss Function

The loss function for model training consists of four terms and optimizes the network for noise reduction, intelligibility improvement, and interaural cue preservation. The loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \alpha\mathcal{L}_{SNR} + \beta\mathcal{L}_{STOI} + \gamma\mathcal{L}_{ILD} + \kappa\mathcal{L}_{IPD}, \tag{4}$$

where $\mathcal{L}_{SNR}$ is the Signal-to-Noise Ratio (SNR) loss, $\mathcal{L}_{STOI}$ is the Short-Time Objective Intelligibility (STOI) [19] loss, and $\mathcal{L}_{ILD}$ and $\mathcal{L}_{IPD}$ are the proposed ILD and IPD error losses which are functions of both $\hat{S}_L$ and $\hat{S}_R$. The parameters α, β, γ, and κ are the weights applied to each term.

$\mathcal{L}_{SNR}$ is defined as the mean of the left and right-ear channel values and append a negative sign to maximize the SNR value, such that $\mathcal{L}_{SNR} = -(\text{SNR}_L + \text{SNR}_R)/2$. The SNR of the enhanced signal, $\hat{\mathbf{s}}$, is defined as

$$\text{SNR}(\mathbf{s}, \hat{\mathbf{s}}) = 10\log_{10}\left(\frac{\|\mathbf{s}\|^2}{\|\mathbf{e}_{noise}\|^2}\right), \tag{5}$$

where $\mathbf{e}_{noise} = \hat{\mathbf{s}} - \mathbf{s}$ with $\mathbf{s}$ and $\hat{\mathbf{s}}$ being the clean and enhanced signal vectors respectively, and $\|.\|$ is the L2 norm.

To optimize the intelligibility of the enhanced speech signals, STOI is computed for left and right channels individually and averaged, and a negative sign is appended to maximize the STOI so that $\mathcal{L}_{STOI} = -(\text{STOI}_L + \text{STOI}_R)/2$ [20]. Including $\mathcal{L}_{STOI}$ helps the network optimize individual CRMs to maximize intelligibility in the enhanced signals.

To ensure the preservation of interaural cues in the enhanced binaural speech using two separate CRMs, minimizing the ILD and IPD errors of the target speech is enforced using the loss function. The ILD and IPD for the clean speech signal are given by

$$ILD_S(k,\ell) = 20\log_{10}\left(\frac{|S_L(k,\ell)|}{|S_R(k,\ell)|}\right), \tag{6}$$

$$IPD_S(k,\ell) = \arctan\left(\frac{S_L(k,\ell)}{S_R(k,\ell)}\right). \tag{7}$$

The ILD and IPD for the enhanced speech, $\hat{S}$, are calculated similarly. In order to restrict the ILD and IPD errors to the speech-active regions, an Ideal Binary Mask (IBM) [21] $\mathcal{M}$ is computed by selecting the TF bins which have energy above a threshold. The energy $E(k,\ell)$ of the clean signal is given by

$$E(k,\ell) = 10\log_{10}|S(k,\ell)|^2. \tag{8}$$

The IBM $\mathcal{M}(k,\ell)$ that defines the speech active TF tiles is then defined as,

$$\mathcal{M}(k,\ell) = \begin{cases} 1 & E(k,\ell) > \max\limits_{\ell}(E(k,\ell)) - \mathcal{T} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where $\max_l(E(k,\ell))$ is the maximum energy computed for each frequency bin, k. Individual IBMs, $\mathcal{M}_L$ and $\mathcal{M}_R$ are computed for the left and right-ear channels. The final mask $\mathcal{M}$ is obtained by choosing the bins that have energy above the threshold, $\max_\ell(E(k,\ell)) - \mathcal{T}$, in both channels and is given by

$$\mathcal{M}(k,\ell) = \mathcal{M}_L(k,\ell) \odot \mathcal{M}_R(k,\ell), \tag{10}$$

where $\odot$ denotes the Hadamard product. For training and evaluation $\mathcal{T} = 20$ dB was used [21]. The human auditory system interprets ILDs and IPDs differently based on the frequency range. Human spatial hearing relies primarily on interaural phase difference cues for frequencies below 1500 Hz, while interaural level difference cues play a crucial role for frequencies above 1500 Hz [4]. From the estimated IBM $\mathcal{M}(k,\ell)$, separate masks for ILD and IPD cues are computed based on human spatial hearing. For choosing the IPD cues in the speech active bins below $f_p = 1500$ Hz, $\mathcal{M}(k,\ell)$ for $k \leqslant K_p$ is selected where $K_p = f_p \times N_{fft}/f_s$ with N_{fft} and f_s being the FFT length and sampling frequency respectively. Similarly, for choosing the ILD cues $\mathcal{M}(k,\ell)$ for $k > K_p$ is selected. The $\mathcal{L}_{ILD}$ and $\mathcal{L}_{IPD}$ terms are given by

$$\mathcal{L}_{ILD} = \frac{1}{N_{ld}} \sum_{k>K_p,\ell} \mathcal{M}(k,\ell)\left(|ILD_S(k,\ell) - ILD_{\hat{S}}(k,\ell)|\right), \tag{11}$$

$$\mathcal{L}_{IPD} = \frac{1}{N_{pd}} \sum_{k\leqslant K_p,\ell} \mathcal{M}(k,\ell)|IPD_S(k,\ell) - IPD_{\hat{S}}(k,\ell)| \tag{12}$$

where $N_{ld} = \sum_{k>K_p,\ell} \mathcal{M}(k,\ell)$ and $N_{pd} = \sum_{k\leqslant K_p,\ell} \mathcal{M}(k,\ell)$ are the total number of speech-active frequency and time bins determined from the mask for ILD and IPD cues respectively. To preserve the interaural cues of the target speaker, the network optimization is guided by using masks based on the target speech. The errors in ILD and IPD are calculated in the time-frequency domain, while losses related to SNR and STOI are computed in the time domain through waveform synthesis using the ISTFT.

4. EXPERIMENTS

4.1. Datasets

To generate binaural speech data, monaural clean speech signals were acquired from the CSTR VCTK corpus [22] and then spatialized using Head Related Impulse Responses (HRIRs) from [23]. The selected speech corpus contains approximately 13 hours of spoken English data recorded by 110 speakers with diverse accents. From this dataset, each two-second utterance was converted into binaural form with distinct left and right-ear channels. The dataset consisted of 20,000 speech utterances that were divided into training, validation, and testing sets. Diffuse isotropic noise was generated using noise signals from the NOISEX - 92 database [24]. Uncorrelated noise sources were evenly placed at intervals of 5° in the azimuthal plane to create isotropic noise [8] using HRIRs from [23]. For generating binaural signals, the target speech was placed randomly within the frontal plane (-90° to $+90^\circ$), utilizing HRIRs recorded with a HATS [23]. For training, isotropic noise was added to the VCTK corpus [22] so that $(SNR_L + SNR_R)/2$ lies between -7 dB and 16 dB. The noise types used for training are White Gaussian Noise (WGN), Speech Shaped Noise (SSN), factory noise, and office noise. An unseen engine noise type was included in the evaluation set. The datasets were generated in the anechoic condition. The evaluation set comprises speech signals from both the VCTK corpus [22] and the TIMIT corpus [25]. In this set, random target azimuths in the frontal azimuthal plane are assigned, and isotropic noise is introduced at varying SNRs ranging from -6 dB to 15 dB. The speaker was positioned at a distance of either 0.8 m or 3 m randomly at a fixed elevation of 0°.

4.2. Training setup

To compute the STFT, an FFT length of 512, a window length of 25 ms, and a hop length of 6.25 ms were employed. A sampling rate of 16 kHz was utilized for all signals.

Binaural Complex Convolutional Recurrent Network (BCCRN): The number of channels used in the model's convolutional layers for the complex-valued encoder and decoder block layers are $\{32, 64, 128, 256, 256, 256\}$, with a stride of 2 in the frequency and 1 in the time dimension with a kernel size of (5,1) and all the convolutions in these layers are causal. 8 layers of bidirectional complex-valued LSTMs with a hidden size of 128 was used. The model was implemented with Pytorch which provides native complex data support for most of the functions. The linear layer placed after the recurrent block has an input and output feature size of 1024. The Pytorch model was trained using the Adam optimizer, an initial learning rate of 0.001, and a multi-step learning rate scheduler to modify the learning rate with the validation loss. The model has around 5.7 million parameters and was trained for 100 epochs with an additional early stopping condition of no improvement in the validation loss for three consecutive epochs. The weights for the loss function $\alpha, \beta, \gamma, \kappa$ in (4) were assigned as {1,10,1,10} respectively. These weight values were selected to standardize the units of each individual loss function term. The terms involving SNR and ILD are computed in dB, IPD is calculated in radians and STOI is a bounded score ranging from 0 to 1. The model was trained with the proposed loss function described in (4), named BCCRN-SILP, and for comparison, the model was trained to maximize the SNR, named BCCRN-S from (5).

4.3. Baselines

Binaural STOI-Optimal Masking (BSOBM): A binaural speech enhancement method using STOI-optimal masks proposed in [7]. Here a feed-forward Deep Neural Network (DNN) was trained to estimate a STOI-optimal continuous-valued mask to enhance binaural signals using dynamically programmed High-resolution Stochastic WSTOI-optimal Binary Mask (HSWOBM) as the training target [7]. To preserve the ILDs, a better-ear mask was computed by choosing the maximum of the two masks. The mask is used to supply Speech Presence Probability (SPP) to an Optimally-modified Log Spectral Amplitude (OM-LSA) enhancer. The model was trained and evaluated on the same dataset as the proposed model.

Binaural TasNet (BiTasNet): A time-domain CED-based network for binaural speech separation which was introduced in [9]. The best-performing version of the model, the parallel encoder with mask and sum, was modified and retrained for single-speaker binaural speech enhancement. The network was trained to maximize SNR [9]. The encoder and decoders in the model had a size of 128, a feature dimension of 128, kernel size of 3 and 12 layers. All other parameters were adapted from the original article and the model has a size of 7 million parameters. The model was trained and evaluated on the same dataset used for the proposed method.

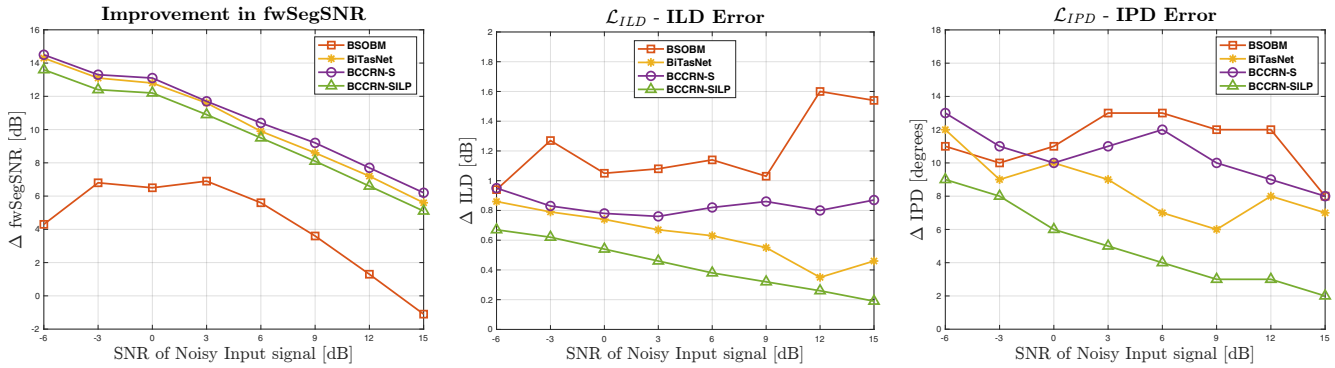

Fig. 2: Comparison of (a) improvement in frequency-weighted Segmental SNR (fwSegSNR) (b) ILD error (11) and (c) IPD error (12) for speech signals with isotropic noise averaged over all frames, frequency bins, and utterances.

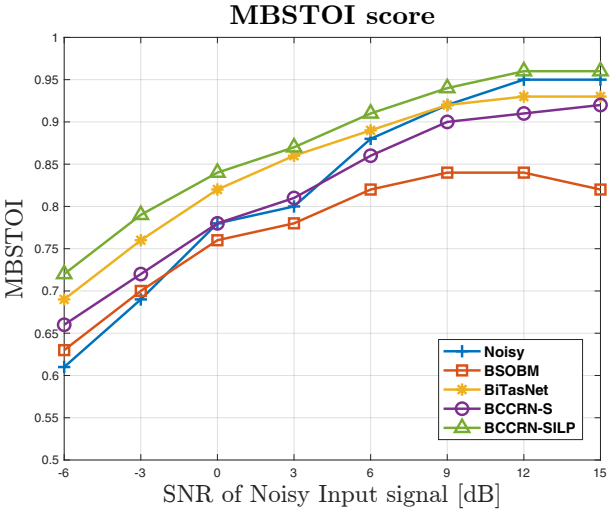

Fig. 3: Comparison of the MBSTOI score for speech signals with isotropic noise after enhancement averaged over all utterances.

5. RESULTS AND DISCUSSION

The performance of the methods was assessed by evaluating 750 utterances from both VCTK [22] and TIMIT [25] datasets for each of the 8 input SNRs. The noise reduction capability of the methods was demonstrated through improvement in fwSegSNR [26]. Objective binaural speech intelligibility of the enhanced signals was measured using the Modified Binaural STOI (MBSTOI) [27] metric. The preservation of interaural cues was evaluated by calculating the error in ILD and IPD after processing, using equations (11) and (12), respectively. Figure 2a shows the improvement in fwSegSNR [26] for different input SNRs. BCCRN-S has the best performance for almost all SNRs, with noise reduction measured by the improvement in the fwSegSNR, while BSOBM has the lowest improvement. Nevertheless, the proposed model exhibits similar effectiveness to the BiTasNet and BCCRN-S in reducing noise. The model has better performance when trained to optimize the SNR compared to the proposed loss function and provides an additional 1 dB of improvement in fwSegSNR on average. A maximum improvement of about 14 dB fwSegSNR is observed in the noisy input SNRs and the amount of improvement observed tends to decrease as the input SNR of the noisy signal improves for both the BCCRN versions. Figures 2b and 2c show the ILD and IPD errors after enhancement computed using (11) and (12). The proposed model and loss function have the lowest error for both cues. The suggested model utilizing the SNR loss function demonstrates comparable performance to the proposed loss function in reducing noise, but it does not prioritize preserving the interaural differences. The inclusion of additional terms in the loss function aids the network in better maintaining interaural differences. The observed ILD error was under 1 dB and the IPD error was under 10° for all input SNRs of the noisy signal. Also, the ILD and IPD errors for the proposed method tend to decrease with increasing input SNR while this is not observed in the model with SNR loss function and other methods. Figure 3 shows the MBSTOI of the enhanced signals. The proposed loss function had the best performance for all SNRs and provided an average of 0.15 to 0.25 improvement in the MBSTOI score. The BCCRN-S has a lower intelligibility score even though it has the best noise reduction performance. Despite its improved ability to reduce noise, the BiTasNet model demonstrates a lower binaural intelligibility score as measured by MBSTOI shown in Fig. 3. Informal listening tests revealed that the BiTasNet produced more artefacts and reduced intelligibility of the speech. Similar to fwSegSNR and the error in interaural cues, BSOBM has the lowest MBSTOI score for enhanced signals. A common trend observed in binaural speech enhancement methods is the degradation of the MBSTOI score at high input SNRs due to processing [7] as the input signals inherently have a higher MBSTOI score. However, the proposed loss function does not deteriorate the MBSTOI score, preserves the intelligibility, and provides an improvement at all SNRs. Audio examples of all the methods can be found online [1].

6. CONCLUSION

In this paper, an end-to-end binaural speech enhancement method using a complex convolutional recurrent network is proposed. A loss function that optimizes the network for noise reduction, speech intelligibility, and human perception-based interaural cue preservation is proposed. The results of the experiments indicate that the suggested technique successfully reduced noise while preserving ILD and IPD information in the enhanced output. Additionally, the proposed method yielded better estimated binaural speech intelligibility compared to the baseline methods.

[1] https://vikastokala.github.io/bccrn/

7. REFERENCES

[1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. John Wiley & Sons, Inc., 2008.

[2] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An Introduction to the Speech Enhancement for Augmented Reality (Spear) Challenge," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[3] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, 2004.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: The MIT Press, 1997.

[5] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Binaural multichannel Wiener filter with directional interference rejection," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015.

[6] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural multi-channel Wiener filtering for hearing aids: Preserving interaural time and level differences," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 5, 2006, pp. V–V.

[7] V. Tokala, M. Brookes, and P. A. Naylor, "Binaural Speech Enhancement Using STOI-optimal Masks," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[8] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 461–465.

[9] C. Han, Y. Luo, and N. Mesgarani, "Real-Time Binaural Speech Separation with Preserved Spatial Cues," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 6404–6408.

[10] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[11] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Sep. 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, 2015.

[13] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 2018, 2018, pp. 3229–3233.

[14] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[15] 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 *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, Sep. 2020, pp. 2472–2476.

[16] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 6649–6653.

[17] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, pp. 331–342, 2006.

[18] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Speech Enhancement using Deep Complex Convolutional Transformer Networks," in *Submitted to ICASSP*, Seoul, South Korea, 2024.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar. 2010, pp. 4214–4217.

[20] P. Manuel, "Mpariente/pytorch_stoi," Feb. 2023. [Online]. Available: https://github.com/mpariente/pytorch_stoi

[21] D. Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Springer US, 2005, pp. 181–197.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[23] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-Ear head-related and binaural room impulse responses," *EURASIP J. on Advances in Signal Process.*, vol. 2009, no. 1, p. 298605, Jul. 2009.

[24] 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 Commun.*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, USA, Corpus LDC93S1, 1993.

[26] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[27] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, Sep. 2018.