



A Deformable Convolution GAN Approach for Speech Dereverberation in Cochlear Implant Users

Hsin-Tien Chiang, John H.L. Hansen

Cochlear Implant Processing Laboratory, Center for Robust Speech Systems (CRSS-CILab), The University of Texas at Dallas, USA

(hsin-tien.chiang, john.hansen)@utdallas.edu

Abstract

Speech dereverberation is crucial for enhancing intelligibility and quality, especially for cochlear implant (CI) users, who are highly susceptible to smearing effects induced by reverberation. While conventional and deep learning-based methods have shown promise for normal-hearing (NH) individuals, their effectiveness for CI users remains limited. To bridge this gap, we propose a deformable convolutional GAN architecture for dereverberation for CI users. The deformable convolution layers introduce kernel offset prediction, adaptively adjusting the receptive field based on distortion in reverberant speech. We first evaluate the effectiveness of the proposed method on REVERB challenge dataset. A listening test is conducted with both NH and CI users. Results show that the proposed method markedly improves speech intelligibility for CI users by preserving a more intact envelope structure, enhancing their ability to perceive key transient speech segments for sentence comprehension.

Index Terms: speech dereverberation, cochlear implant, generative adversarial networks, deformable convolution

1. Introduction

In real-world environments, room acoustics significantly impact the transmission of speech signals. During conversations, individuals perceive not only the direct sound but also reflections from surfaces such as walls, ceilings, and furniture. These reflections, collectively termed reverberation, introduce temporal and spectral smearing that interferes with the direct sound, thereby degrading speech quality. This degradation adversely affects listening experiences and poses challenges for tasks like speech recognition and speaker identification, highlighting the importance of effective dereverberation techniques.

Early research on speech dereverberation includes estimating a Wiener-like filter based on factors such as the estimated reverberation time [1], the power spectral density (PSD) of late reverberation [2], or a relative convolutive transfer function model [3]. One of the most widely used approaches is the weighted prediction error (WPE) algorithm [4, 5]. WPE utilizes variance-normalized delayed linear prediction to estimate late reverberation based on past speech frames, which is then subtracted from the current signal to recover the target speech. However, WPE iteratively refines its estimation by updating both the time-varying PSD of the target speech and the linear filter, which is time-consuming.

With the advancement of deep learning (DL), many DL-based methods have been developed to address speech dereverberation, primarily in the time-frequency (T-F) domain, with limited exploration in the time domain [6, 7]. In T-F domain approaches, dereverberation methods are typically categorized into masking-based and mapping-based techniques. Masking-

based approaches include the ideal binary mask (IBM) [8, 9], ideal ratio mask (IRM) [10], and complex ratio mask (CRM) [11], whereas mapping-based methods directly reconstruct the spectral representation of clean speech [12, 13, 14]. There are also research working on generative models for speech dereverberation, including generative adversarial networks (GANs) [15, 16] and diffusion models [17].

However, most of these methods are designed for normal-hearing (NH) individuals, with little application and translation into approaches for improving speech understanding for cochlear implant users. Cochlear implants (CIs) are electronic devices that provide a hearing solution for individuals with severe-to-profound hearing loss. While moderate reverberation generally does not hinder speech understanding for NH individuals, CI users are more susceptible to the negative effects of reverberation, resulting in significant reduction of sentence-level and word-level speech understanding [18, 19, 20]. This is primarily due to the distortion of spectro-temporal cues, blurring of formant transitions, and a reduction of envelope structure (amplitude modulations), all of which collectively and individually impair speech understanding in quiet and noisy environments. Additionally, reverberation has been shown to amplify low-frequency energy and mask higher-frequency speech components thus degrading the representation of the fundamental frequency which also negatively impacts speech understanding [21, 22]. Several algorithms have been proposed to suppress reverberation in CI users. A channel-selection strategy based on the signal-to-reverberant ratio of individual frequency channels resulted in over 60% improvement [18]. In a later study, the approach was modified such that signal was no longer dependent on prior knowledge of the room impulse response or anechoic signal, which demonstrated an increase of performance by an average of 32.21 percentage points [20].

Previous solutions for dereverberation for CIs are primary signal-processing-based approaches. This study utilizes a DL architecture that integrates deformable convolution networks with a GAN framework for speech dereverberation in CI users. While deformable convolution networks have proven effective in computer vision [23, 24], they have not been applied within a GAN framework for speech dereverberation in CI users. The proposed study represents an application of how to leverage DL techniques to improve speech understanding in reverberation conditions for CI users. The contributions of this study are as follows:

- We apply a deformable convolution-based GAN for reverberation suppression. To the best of our knowledge, this is the first approach that utilizes a GAN framework for CI users. Our subjective results demonstrate the superiority of the proposed method in recovering a clear and intact envelope structure, which enhances the perception of important transient

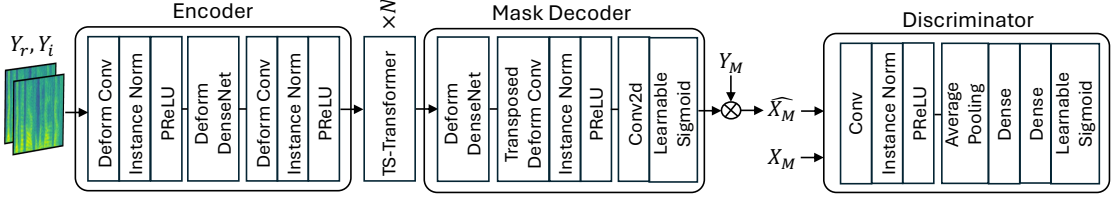


Figure 1: The framework of the proposed method.

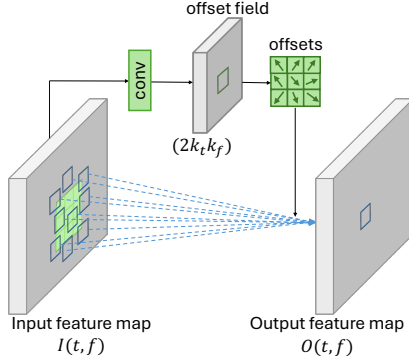


Figure 2: Illustration of deformable convolution.

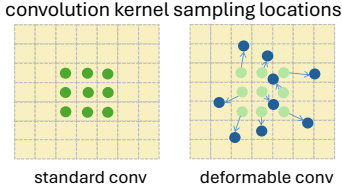


Figure 3: The sampling locations for standard and deformable convolutions utilized a 3×3 kernel to generate the output.

speech segments essential for speech intelligibility.

- We replace standard convolution layers with deformable convolution networks, which introduce an additional convolutional layer to predict kernel offsets, thereby enabling our model to dynamically identify optimal T-F regions and effectively handle varying levels of reverberation distortion in the degraded signal.

2. Method

2.1. Deformable convolution network

A standard convolution operation applies on a fixed grid of neighboring locations to extract useful features, which can be expressed as:

$$O(t, f) = \sum_{(x, y) \in R} K(x, y) \cdot I(t + x, f + y),$$

$$R \in \{(-\lfloor k_t/2 \rfloor, -\lfloor k_f/2 \rfloor), \dots, (\lfloor k_t/2 \rfloor, \lfloor k_f/2 \rfloor)\} \quad (1)$$

where I is the 2-D intermediate feature map from the previous layer with dimensions $(T \times F)$, and K is the convolutional kernel of size $(k_t \times k_f)$. The receptive field R defines the grid of locations over which the convolution is performed, with (t, f)

representing the specific position. However, as can be seen in (1), standard convolution layers generate output by sampling from fixed neighboring locations on the input.

Deformable convolution extends standard convolution by introducing trainable offsets, allowing the receptive field to adapt dynamically to the input, unlike the fixed positions in standard convolutions. The deformable convolution is defined as:

$$O(t, f) = \sum_{(x, y) \in R} K(x, y) \cdot I(t + x + \delta_x, f + y + \delta_y) \quad (2)$$

where (δ_x, δ_y) are learned offsets for each (x, y) location in the sampling grid R . As shown in Figure 2, these offsets are computed through an additional convolutional layer, generating a deformable feature map that adapts to the spatial characteristics of the input. This allows dynamic adjustment of sampling locations to optimal T-F regions in the input reverberant spectrogram. Figure 3 compares the sampling locations used by standard and deformable convolutions to generate the output. It can be seen that compared to standard convolution, deformable convolution can dynamically adjust and expand the receptive field, offering flexibility for feature modeling.

2.2. Model structure

The architecture of the proposed method is shown in Figure 1. The model converts a reverberant speech waveform y into real and imaginary complex spectrogram using short-time Fourier transform (STFT). The real and imaginary parts, Y_r and Y_i , are then concatenated as input and sent into the generator. The generator comprises of encoder, two-stage transformer blocks (TS-Transformer) and a mask decoder. The encoder is composed of a deformable convolutional block, a deformable DenseNet (Deform DenseNet), followed by another deformable convolutional block. Each block integrates a deformable convolutional layer (Deform Conv), instance normalization, and PReLU activation. Building on [25], Deform DenseNet replaces the four standard convolutional layers with deformable convolutions. Then, in each TS-Transformer, the first transformer along time dependencies and the second one models along frequency. The amount of TS-Transformer is 4. The mask decoder aims to predict a magnitude mask, which is then element-wise to the reverberant magnitude Y_M to estimate the dereverberated magnitude spectrum \hat{X}_M . The mask decoder comprises a Deform DenseNet, followed by a transposed deformable convolution layer, instance normalization, PReLU activation, and a final convolutional layer with a learnable sigmoid function. The enhanced magnitude spectrum \hat{X}_M is combined with the noisy phase and transformed back to the enhanced signal \hat{x} using the inverse STFT. The generator is trained based on the magnitude loss and complex loss between the clean and dereverberated spectrogram [26], and adversarial loss which encourages the

Table 1: Performance comparison on SimData and RealData of the REVERB challenge evaluation set. “-” denotes that the result is not provided in the original paper.

	SimData				RealData	
	CD	SRMR	LLR	FWSegSNR	PESQ	SRMR
No Processing	3.975	3.687	0.574	3.617	1.503	3.180
WPE [5]	3.748	4.220	0.514	4.864	1.722	3.978
BSW [28]	4.325	4.072	0.54	7.971	1.633	3.455
SkipConvNet [14]	2.328	4.852	0.261	10.746	2.154	7.06
SkipConvGAN [15]	2.318	5.887	0.234	11.896	2.911	6.355
CMGAN [16]	2.25	5.47	0.31	11.74	-	6.55
Proposed method	2.329	5.136	0.211	12.259	3.015	7.13

generator to produce dereverberated speech with a perceptual evaluation of speech quality (PESQ) score close to that of clean speech.

Since objective functions in speech enhancement are often not directly correlated with evaluation metrics, we follow the approach in [16] that incorporates a metric discriminator to approximate the target metric (i.e. PESQ). We adopt the architecture in [27], which includes four convolutional blocks, each with a convolutional layer, instance normalization, and PReLU activation. Then global average pooling, feed-forward layers, and a learnable sigmoid function are applied to predict the PESQ score. The discriminator is trained by first estimating the maximum normalized PESQ score from clean speech, and second, by predicting the PESQ score from the dereverberated speech.

3. Experiments

3.1. Dataset and experimental setup

We use the REVERB challenge dataset [29], which provides single-channel, two-channel and eight-channel configurations at a 16 kHz sampling rate. We consider the single channel configuration in this work. The dataset is divided into training, development and evaluation sets. The training set consists of 7,861 clean utterances from the WSJCAM0 corpus [30]. Reverberant speech is simulated by convolving these clean utterances with 24 measured room impulse responses and adding noise at an signal-to-noise ratio (SNR) of 20 dB. Reverberation times range from 0.2 to 0.8 seconds. Both the development and evaluation sets include simulated data (*SimData*) and real recordings (*RealData*). *SimData* comprises six conditions: three rooms with reverberation times of 0.3, 0.6, and 0.7 seconds, and two distances between the speaker and microphone: near at 0.5 m and far at 2 m. *RealData* is taken from the MC-WSJ-AV corpus [31] and it includes one room with two distances: near at 1 m and far at 2.5 m, and a reverberation time of 0.7 seconds. The development set contains 1,484 utterances for *SimData* and 179 for *RealData*, while the evaluation set includes 2,176 and 372, respectively.

During training, the utterances in the training set were segmented into 4-second slices, whereas the test set retained variable-length utterances without slicing. The FFT point number, Hanning window size, and hop size were set to 400, 400, and 100. The model was trained for 200 epochs using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 8. Early stopping is applied to finish training if there is no improvement in development set for 20 consecutive epochs.

Table 2: Ablation study on the performance of the proposed method in different room sizes. The table shows improvements in average PESQ and SRMR.

Room size	PESQ			SRMR		
	Small	Medium	Large	Small	Medium	Large
No Processing	1.91	1.314	1.285	4.542	3.364	3.155
Proposed method	3.403	2.825	2.815	5.317	5.33	4.762
- Deform Conv	3.378	2.749	2.727	5.025	4.957	4.562
- Discriminator	3.292	2.682	2.527	4.927	4.851	4.31

3.2. Evaluation metrics

To evaluate performance, we utilized five objective measures: cepstral distance (CD), signal-to-reverberation modulation energy ratio (SRMR), log-likelihood ratio (LLR), frequency-weighted segmental SNR (FWSegSNR) and PESQ, all of which are provided by the REVERB challenge corpus. Except for SRMR, these metrics require a clean speech reference to score the reverberant or enhanced speech. As a result, improvements in all metrics are assessed for SimData, while only SRMR improvements are evaluated for RealData.

4. Results

4.1. Comparison with baselines

We compared the proposed method with six methods, including two signal-processing-based and four DL-based approaches. The signal-processing-based methods include WPE [4, 5] and blind spectral weighting (BSW) [28]. BSW suppresses late reverberation without the need for prior knowledge of the anechoic signal or RIR and operates in real time without the iterative processing needed by WPE. We select BSW because its real-time operation is more practical for CI applications. The DL-based methods include the magnitude-mapping U-Net model SkipConvNet [14] and two GAN-based approaches: SkipConvGAN [15] and CMGAN [16]. SkipConvNet incorporates convolutional layers between the U-Net encoder and decoder, SkipConvGAN is derived from SkipConvNet by incorporating a discriminator to form a GAN architecture, and CMGAN includes a conformer-based generator and a metric discriminator.

As shown in Table 1, WPE and BSW demonstrate inferior performance across all objective metrics compared to DL-based methods. Among DL-based methods, SkipConvNet performs slightly lower on simulated data but achieves better SRMR scores on real recordings compared to SkipConvGAN and CMGAN. The two GAN-based methods, SkipConvGAN and CMGAN, demonstrate comparable performance, with the former achieving the highest SRMR and the latter achieving the lowest CD score on simulated data. Overall, the proposed method outperforms baselines on most objective metrics. In addition, it yields superior SRMR on real recordings, demonstrating better robustness and generalization ability. Our demos are available at ¹.

4.2. Ablation study

Starting with the deformable convolutional GAN architecture, we progressively made two modifications: (1) replacing the deformable convolution layers with standard convolutional layers (- Deform Conv), and (2) removing the discriminator from the

¹<https://doi.org/10.5281/zenodo.15565914>

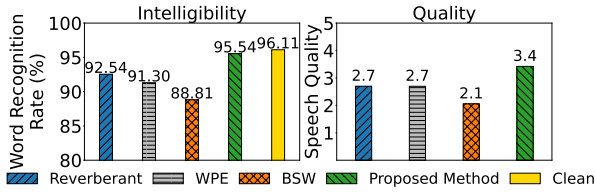


Figure 4: Mean word recognition rate and speech quality score for NH participants.

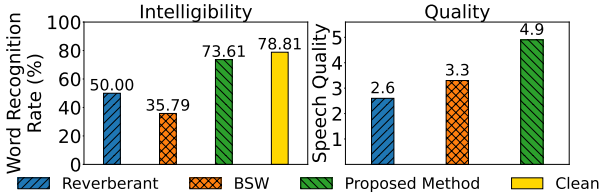


Figure 5: Mean word recognition rate and speech quality score for CI participants.

model after standard convolution was applied (– Discriminator). As shown in Table 2, both PESQ and SRMR scores decline when each component is removed, confirming that the deformable convolution layers and discriminator contribute to reverberation suppression. While both PESQ and SRMR improve, replacing deformable convolution with standard convolution causes a greater drop in SRMR scores, whereas removing the discriminator from the standard convolution configuration results in a larger decrease in PESQ scores, given that the discriminator optimizes based on the PESQ scores.

4.3. Subjective listening test

For the subjective listening tests, we utilized SimData from the REVERB challenge evaluation set. The tests assessed intelligibility and quality with seven NH and two CI subjects. We compared the proposed method with WPE and BSW, with BSW chosen for its real-time practicality in CI applications. Since CI users experience greater listening efforts and listening fatigue more easily compared to NH users, we remove WPE for CI subject testing.

For the intelligibility test, participants were asked to type the words they perceived. The intelligibility is evaluated using word recognition rate (WRR) measures from the test samples. NH participants assessed a total of 100 utterances across five conditions (clean, reverberant, WPE, BSW, and the proposed method), with 20 utterances per condition; CI participants evaluated 80 samples across four conditions (clean, reverberant, BSW, and the proposed method), with 20 samples per condition. For the quality test, participants first listened to a clean speech sample as a reference. Then test audio samples, including both reverberant and dereverberated, were played in random order. Participants rated the quality of each sample on a 5-point scale, where 1 indicated poor quality and 5 represented excellent quality. NH participants assessed a total of 20 utterances (4 conditions \times 5 samples per condition); CI participants evaluated 15 utterances (3 conditions \times 5 samples per condition).

Figure 4 and 5 illustrate the intelligibility and quality results for NH and CI users. First, for reverberant speech, CI users show a significant intelligibility drop, with WRR falling to 50%, while NH users maintain a high WRR of 92.54%. This con-

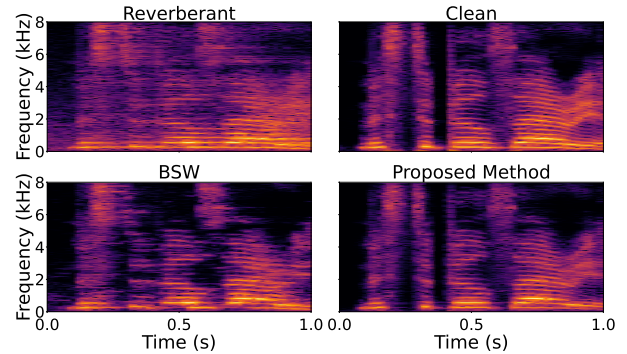


Figure 6: Comparison of spectrograms.

firms that reverberation severely degrades speech recognition for CI users. Second, while WPE and BSW slightly improve objective quality measures over reverberant speech, both methods lead to minor intelligibility declines for NH users, lowering WRR by 1.24% and 3.74%, respectively. For CI users, this phenomenon is more severe, with WRR decreases from 50% (reverberant speech) to 35.79% with BSW. Third, the proposed method consistently improves intelligibility for both NH and CI users, achieving WRRs of 95.54% and 73.61%, significantly narrowing the gap to clean speech. For CI users, it provides a substantial intelligibility boost, increasing WRR by 23.61% over reverberant speech. Subjective quality scores also confirm its effectiveness, reaching 3.4 and 4.9 for NH and CI users, respectively. Fourth, the quality and intelligibility for NH users show a positive correlation, where higher quality scores correspond to higher intelligibility. However, for CI users, while BSW results in a lower WRR than reverberant speech, it has a higher quality score of 3.3 compared to 2.6 for reverberation.

Figure 6 visualizes the spectrograms of reverberant, clean, and dereverberated speech processed by BSW and the proposed method. While BSW slightly reduces reverberation, the residual spectral smearing (blurring) contributes to decreased intelligibility. This effect is particularly detrimental for CI users with poor spectro-temporal resolution, as spectral smearing negatively leads to word boundary identification [32]. In contrast, the envelope structure of the proposed method appears to be more intact than BSW, which increases the potential for CI users to perceive important transient segments of speech necessary to achieve high performance for sentence intelligibility.

5. Conclusion

The proposed method integrates deformable convolution networks within a GAN framework for speech dereverberation for CI users. To the best of our knowledge, this is the first GAN-based approach applied for CI users. The deformable convolution networks dynamically identify optimal T-F regions, adapting effectively to varying levels of reverberation distortion in degraded speech. Subjective tests confirm the effectiveness of proposed method in recovering clear and intact speech envelopes, enhancing the perception of transient segments critical for CI users. Future work will explore causal or compressed architectures to improve real-time efficiency and reduce computational complexity, making the model more practical for CI applications.

6. References

- [1] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.
- [2] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, 2018.
- [3] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.
- [4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [5] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.
- [6] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Interspeech*, 2018, pp. 342–346.
- [7] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020, pp. 4506–4510.
- [8] N. Roman and J. Woodruff, "Intelligibility of reverberant speech with ideal binary masking," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2153–2161, 2011.
- [9] Y.-T. Chen, T.-H. Chen, M.-C. Huang, and T.-S. Chi, "Interaural coherence induced ideal binary mask for binaural speech separation and dereverberation," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.
- [10] X. Li, J. Li, and Y. Yan, "Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions," in *Interspeech*, 2017, pp. 1203–1207.
- [11] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5590–5594.
- [12] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.
- [13] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 390–394.
- [14] V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, "Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping," in *Interspeech*, 2020, pp. 3935–3939.
- [15] V. Kothapally and J. H. Hansen, "Skipconvgan: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1600–1613, 2022.
- [16] R. Cao, S. Abdulatif, and B. Yang, "Cmgan: Conformer-based metric gan for speech enhancement," in *Interspeech*, 2022, pp. 936–940.
- [17] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [18] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3221–3232, 2011.
- [19] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1607–1614, 2013.
- [20] O. Hazrati and P. C. Loizou, "Reverberation suppression in cochlear implants using a blind channel-selection strategy," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 4188–4196, 2013.
- [21] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *Journal of Speech and Hearing Research*, vol. 17, no. 4, pp. 724–739, 1974.
- [22] S. Greenberg, W. A. Ainsworth, A. N. Popper, R. R. Fay, P. Assmann, and Q. Summerfield, "The perception of speech under adverse conditions," *Speech processing in the auditory system*, pp. 231–308, 2004.
- [23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.
- [24] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.
- [25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [26] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2021, pp. 72–76.
- [27] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Interspeech*, 2023, pp. 3834–3838.
- [28] S. O. Sadjadi and J. H. Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 937–945, 2014.
- [29] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.
- [30] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsj-camo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 81–84.
- [31] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005. IEEE, 2005, pp. 357–362.
- [32] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *International journal of audiology*, vol. 51, no. 6, pp. 437–443, 2012.