U-Net Based Dereverberation using Fusion of Spectral Representation of TEO and Raw Audio

Devdeep Shetranjiwala, Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, Gujarat, India
{202001150@daiict.ac.in, hemant_patil@daiict.ac.in}

***Abstract*—In this paper, we present a comprehensive analysis of natural and reverberated speech signals, focusing on their time-domain characteristics and corresponding Teager Energy Operator (TEO) profiles. Our study highlights significant differences between these two perspectives, offering new insights into the acoustically complex phenomenon of reverberation. We propose a dereverberation method that utilizes the data fusion of spectrogram of raw waveform and its corresponding TEO profile.This fusion approach is integrated into a U-Net-based convolutional neural network (CNN) model, which is well-suited for learning structured representations in speech enhancement tasks. In particular, U-Net performs reverse mapping from the reverberated speech signal to the original audio. For our experiments, we use the Valentini-Botinhao Noisy Speech Dataset, which provides a controlled environment for evaluating dereverberation techniques. Our results demonstrate that the proposed fusion of features, effectively suppresses reverberation, significantly improving speech quality and intelligibility. Specifically, we observe a 1.25% improvement in Root Mean Squared Logarithmic Error (RMSLE), decreasing from 40 to 39.5, indicating the robustness of our approach.**

***Index Terms*—Deep Learning, U-Net CNN, Teager Energy Operator (TEO), Dereverberation.**

I. Introduction

Reverberation is a natural acoustic phenomenon that occurs when sound waves reflect off surfaces multiple times until they gradually lose energy and dissipate. The distinction between echo and reverberation lies in the *number* and *duration* of reflections. Echo involves fewer reflections with a longer delay, whereas reverberation consists of rapid, densely packed reflections with a shorter duration [6]. Echo results from a single, distinct reflection off a distant hard surface, while reverberation occurs due to multiple reflections from nearby surfaces, creating a continuous, blended sound [6]. Reverberation introduces a delay in the speech signal due to reflections that vary based on environmental factors. It converts a monocomponent signal into a multicomponent one, where the spectral proximity of components makes it difficult to distinguish the original speech from its reflections [1]. Recent advancements in speech dereverberation have led to state-of-the-art techniques. Zhang *et al*. proposed an end-to-end framework integrating dereverberation, beamforming, and speech recognition, demonstrating enhanced numerical stability and improved performance in challenging auditory environments such as cocktail party scenarios [11], [12]. Additionally, Lemercier *et al*. introduced a customizable online neural network-based dereverberation system designed for hearing devices, optimizing speech clarity [13]. These contributions collectively enhance speech recognition accuracy and reduce the impact of reverberation. In this paper, we explore speech features for dereverberation. Our focus is on the comparison between natural and reverberated speech signals, specifically examining these signals in the time-domain, and their corresponding Teager Energy Operator (TEO) profiles. Furthermore, we delve into the study of dereverberation, a process aimed at reducing the reverberation effect in a speech signal. For this purpose, we have a model based on the U-Net architecture, which is a type of convolutional neural network known for its effectiveness in reverse mapping from noisy speech spectrum to clean speech spectrum. This is particularly important in telecommunication, automatic speech recognition (ASR), and other speech processing applications, where speech quality is crucial. Therefore, the study of reverberation and the development of effective dereverberation techniques are of great significance [7], [10].

II. Mathematical Model for Reverberation

Reverberation occurs when sound waves continue to reflect off surfaces within an enclosed space, even after the original sound source has ceased [2]. These reflections can be classified based on their order: first-order reflections involve a single deviation, second-order reflections undergo two deviations, and higher-order reflections continue this pattern. As reflections accumulate, their intensity gradually diminishes due to absorption by surfaces and objects within the environment, as depicted in Fig. 1 (a) [1], [2], [5].

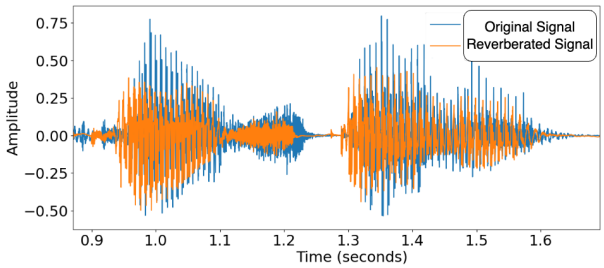

Fig. 1. Reverberation causes alterations in the delay and amplitude of speech, where 'simple' refers to unaltered, clean speech, and 'reverb' denotes the speech post-reverberation. Data taken from WHAMR dataset [16].

The reflections introduce a delay and variation in amplitude relative to the original speech signal, causing temporal and

spectral distortions, as shown in Fig. 1. The speech signal in a reverberant environment can be mathematically modeled as the convolution of the clean speech signal, $s(t)$, with the room impulse response (RIR), $h(t)$, as depicted in Fig. 2 [1]:

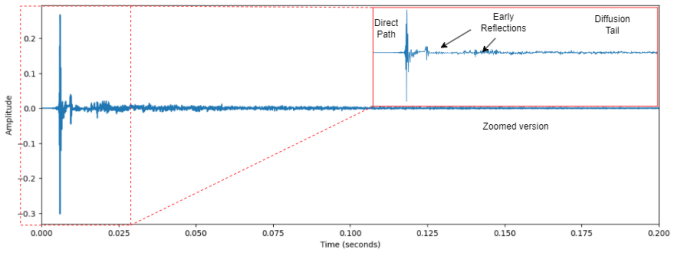

Fig. 2. Impulse response and its characteristics. Impulse response (IR) taken from ICASSP DNS 2023 Challenge Dataset. After [6].

$$y(t) = s(t) * h(t), \tag{1}$$

The observed reverberant speech is denoted as $y(t)$, with $*$ representing the convolution operation. Fig. 2 depicts the impulse response (IR) of a reverberant environment, highlighting a distinct structure within the initial 100 ms, followed by a *diffuse tail*. The early segment comprises discrete reflections, primarily first- or second-order, which then transition into a densely packed diffuse region. These characteristics play a crucial role in defining the acoustic properties of various environments. The first peak in the reverberant signal corresponds to the direct-path signal, which travels the shortest distance from the source to the microphone. Subsequent peaks arise due to reflections, each associated with a unique propagation path. As the reflections increase in density, they begin to overlap over time [2]. Since surfaces and air absorb energy at each reflection, longer propagation paths result in lower amplitudes, creating a gradually decaying tail in the impulse response. Under the Linear Time-Invariant (LTI) assumption, the impulse response of an LTI system describes the acoustic environment in which the recording takes place [1], [2].

Reverberation time (RT60) is a key parameter for assessing reverberation, measuring how long it takes for sound energy to decrease by 60 dB after the source stops emitting. It is defined as the time interval $t_2 - t_1$ during which the sound pressure level drops by 60 dB [3]. RT60 plays a vital role in characterizing a room's acoustical properties. When speech signals are recorded with distant microphones, they inevitably include both noise and reverberation, which degrade speech quality and impact automatic speech recognition (ASR) performance. A reverberant speech signal at time t, represented as $y(t)$, can be formulated as:

$$y(t) = h(t) * s(t) + n(t), \tag{2}$$

where $h(t)$ is the room impulse response between the speaker and the microphone, $s(t)$ is the clean speech signal, $n(t)$ represents background noise, and $*$ denotes convolution. While various signal processing techniques, such as TEO, effectively suppress additive noise (i.e., $n(t)$ in Eq. (2)), handling reverberation (i.e., $h(t)$ in Eq. (2)) remains a challenging problem in speech enhancement research [14].

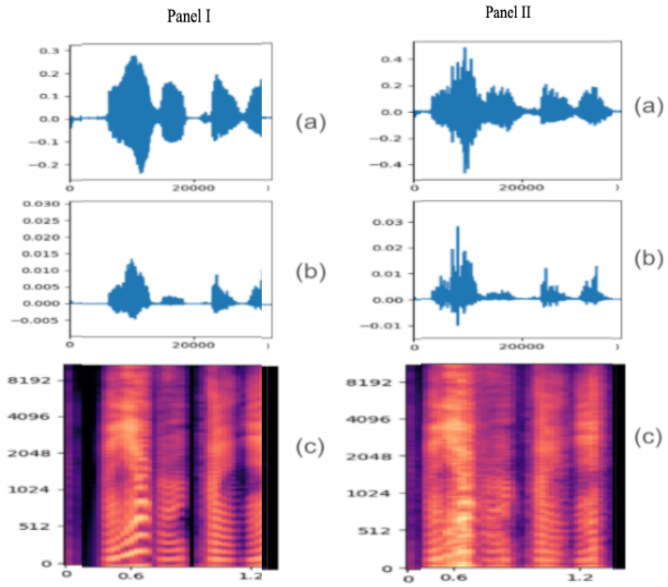

Fig. 3. TEO Analysis: (a) original speech signal, (b) corresponding TEO profile, (c) TEO-based spectrogram. Panels I and II show clean and reverberated speech, respectively. Considerations include RT60 of 1.4 in live spaces, such as a small theatre, with specific materials on the theatre walls affecting acoustics.

III. TEO-Based Features

TEO (Teager Energy Operator) is a non-linear operator used to estimate a signal's running energy by capturing the interaction between its amplitude and frequency. Unlike the conventional L^2 norm, TEO provides an alternative measure of energy in signal processing [18]. It is defined as:

$$\Phi_d\{z(n)\} = z^2(n) - z(n-1)z(n+1) \approx A^2\omega^2. \tag{3}$$

TEO generates high-energy peaks around Glottal Closure Instants (GCIs) due to the abrupt excitation of the vocal tract during glottal closure. This sudden closure results in an energy spike, which is captured by TEO. Additionally, smaller fluctuations, or bumps, appear alongside these peaks, highlighting the non-linear characteristics of natural speech production [1]. The energy of $z(n)$ is estimated using TEO, as proposed by Kaiser (1990) [18], leading to Eq. (3). Since TEO is primarily designed for monocomponent signals, energy separation algorithms (ESA) were introduced to extract individual contributions of amplitude A and frequency ω. A real-valued, continuous-time AM-FM signal is expressed as:

$$z(t) = a(t) \cdot cos(\phi(t)) = a(t) \cdot cos(\omega_c t + \omega_m \int_0^t p(\lambda) d\lambda + \theta), \tag{4}$$

where $a(t)$ represents the time-varying amplitude signal modulated by the high frequency signal $cos(\cdot)$, which results

in AM. The examination of reverberation through the lens of the Short-Time Fourier Transform (STFT) domain, the study of reverberation via Teager Energy features and noise spectrum comparison of TEO as conducted in the papers [1] and [17], inspired us to undertake an analysis using the TEO as shown in Fig.3.

IV. Model Examination

A. Fully-Convolutional Networks U-Net

1) Architecture: The U-Net, a fully-convolutional network architecture, is particularly beneficial for speech dereverberation tasks due to its unique *structure* and *capabilities* [19]. Moreover, U-Net to handle inputs of varying sizes, making it adaptable to different speech signals. Its end-to-end training allows for the direct optimization of the dereverberation task, leading to improved performance [8]. Additionally, U-Net has been successfully used in a Late Reverberation Suppression (LS) setting, demonstrating its effectiveness in reducing reverberation. Furthermore, the Tiny Recurrent U-Net (TRU-Net) variant has been shown to be efficient for online speech enhancement, providing real-time processing capabilities [9].

TABLE I
Parameters and Layers of the U-Net Model After [19].

Parameter	Value
Window Length	512
FFT Length	512
Number of Features	256
Number of Segments	256
Filter Height	6
Filter Width	6
Number of Channels	1
Number of Filters	[64,128,256,512,512,512,512,512]

The U-Net model consists of an input layer, a squeezing path, and an expanding path. The squeezing path includes convolutional 2D layers with leaky ReLU activation functions and batch normalization. The expanding path includes transposed convolutional 2D layers, batch normalization, dropout layers, ReLU activation functions, and a $\tanh$ layer. The model concludes with a regression layer. Skip connections are defined between the leaky ReLU layers and the concatenation layers in the expanding path.

2) Preprocessing and Feature Extraction: In the preprocessing stage, we employ the STFT in order to obtain a *time-frequency representation* of the input speech signal. Given that speech is inherently *non-stationary*, STFT provides a structured way to analyze its spectral evolution over time. However, conventional STFT-based representations often fail to fully capture the intricate energy dynamics present in reverberant conditions [19].

To enhance the spectral representation, we introduce a novel *TEO Fusion* mechanism. The process follows these key steps:

1) **STFT Extraction:** The spectrogram is computed using STFT, providing a 2D time-frequency representation of the reverberant signal.
2) **TEO Spectrum Enhancement:** The *TEO* is applied to the raw signal in order to capture localized instantaneous energy variations and emphasize *high-energy transient components* [18].
3) **Fusion Strategy:** The STFT and TEO-derived spectrum are *fused* using an adaptive weighted averaging method, which integrates localized energy fluctuations into the standard spectrogram.
4) **Normalization and Reshaping:** The fused spectrogram is *scaled to [-1,1]*, ensuring consistency across training samples.

The fused spectrum is computed using a weighted combination of the standard STFT spectrum and the TEO-derived spectrum:

$$S_{\text{fusion}}(t,f) = \alpha S_{\text{STFT}}(t,f) + (1-\alpha) S_{\text{TEO}}(t,f), \quad (5)$$

where $S_{\text{STFT}}(t,f)$ represents the STFT magnitude spectrum, and $S_{\text{TEO}}(t,f)$ corresponds to the spectral representation obtained from TEO profile. The parameter α controls the relative contribution of each component. For this experiment, we set $\alpha = 0.5$. This simple averaging ensures that transient energy variations captured by TEO are integrated into the standard spectrogram while maintaining spectral consistency.

The enhanced STFT-TEO fusion is then *fed into a U-Net-based dereverberation model*, as shown in Fig. 4. This approach builds upon the model proposed in [19], where we retain the same U-Net architecture but introduce the *TEO-enhanced STFT as input*. The U-Net model is trained to map the fused spectrogram to a *clean STFT spectrum*, which is subsequently inverted to reconstruct a high quality, dereverberated waveform. The key advantage of our approach is that the *additional feature space introduced by TEO fusion* enriches the spectral representation, allowing the U-Net to learn a more *robust mapping between reverberant and clean speech signals*.

Our results indicate that this *simple yet effective modification* leads to a **marginal but consistent improvement** over the standard STFT input, reinforcing the hypothesis that **TEO-derived energy cues enhance speech dereverberation models**.

B. Database Used

We used the noisy speech database from the University of Edinburgh's School of Informatics [15] for training and evaluating speech enhancement and TTS models. Recorded at 48 kHz, it includes 23,000 training samples from 56 speakers, 11,000 additional samples from 28 speakers, and 824 test samples. Bit resolution is 16-bit per speech sample. The dataset's diverse noisy conditions make it a valuable resource for developing robust models [15].

C. Performance Metrics

To evaluate the model's performance, we used three key metrics: Cepstrum Distance (CD), Log-Likelihood Ratio (LLR), and Root Mean Squared Logarithmic Error (RMSLE) for assessing speech dereverberation quality.

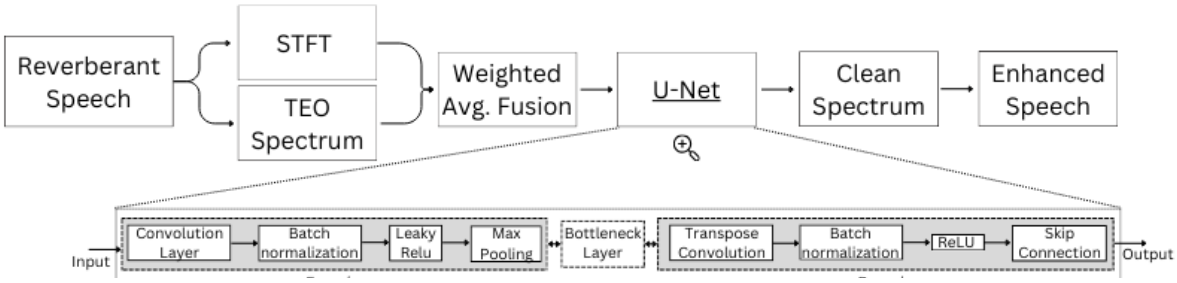

Fig. 4. Functional block diagram of the proposed TEO-STFT fusion-based speech enhancement model.

TABLE II
TRAINING PARAMETERS. AFTER [19].

Parameter	Value
Initial Learning Rate	8e-4
Mini Batch Size	8
Max Epochs	50
Learn Rate Drop Period	15
Execution Environment	GPU
Validation Data	valReverb,valClean

1) Cepstrum Distance (CD): It quantifies the spectral difference between the predicted and clean speech signals in the cepstral-domain [19]. It measures how closely the predicted signal's spectral envelope matches that of the clean speech. A lower CD value indicates a better approximation to the clean signal [19].

2) Log-Likelihood Ratio (LLR): LLR is a linear prediction-based measure that evaluates the spectral distortion between two signals [19]. It compares the prediction residuals of the clean and estimated speech signals, with lower values indicating reduced distortion and better dereverberation quality [19].

D. Root Mean Squared Logarithmic Error (RMSLE):

RMSLE measures the difference between the logarithms of predicted and actual values, focusing on relative errors. It is defined as:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(1+\hat{y}_i) - \log(1+y_i))^2}, \qquad (6)$$

where $\hat{y}_i$ and y_i are the predicted and actual values, respectively, and N is the total samples. The logarithm prevents issues with zero values, and RMSLE is beneficial for speech enhancement as it penalizes overestimations more than underestimations.

Table III presents the comparative performance metrics for speech reconstruction and reverberation using the STFT and STFT + TEO methods.Both performance metrics indicate that the model's output closely matches the clean, dereverberated signal. The STFT + TEO method demonstrates marginal improvements over STFT alone, particularly in CD and LLR values.

TABLE III
COMPARATIVE PERFORMANCE METRICS OF SPEECH RECONSTRUCTION AND REVERBERATION FOR STFT AND STFT + TEO METHODS

	Reverberated	STFT	STFT + TEO
Average CD Mean	4.25	3.84	**3.83**
Average CD Median	3.63	3.35	3.37
Average LLR Mean	0.97	0.94	**0.91**
Average LLR Median	0.87	0.80	0.82

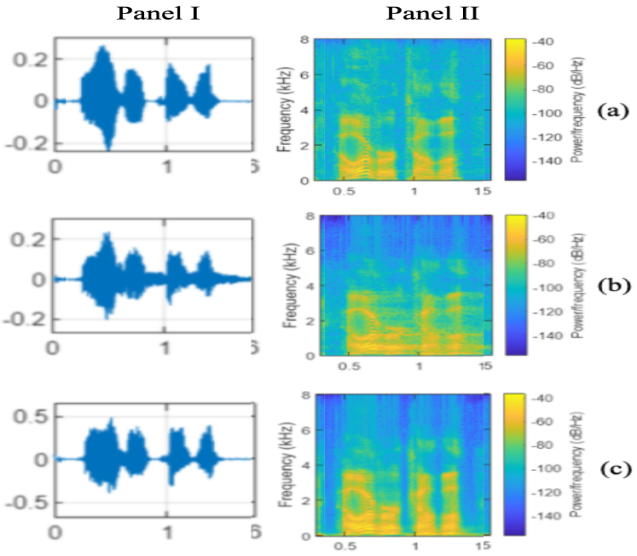

Fig. 5. Comparative analysis of time-domain waveforms (Panel I) and spectrograms (Panel II) for (a) clean, (b) reverberated, and (c) predicted dereverberated speech signals.

E. Experimental Results

The U-Net demonstrated strong performance in dereverberation. Building upon prior work [19], we adopted the U-Net architecture while introducing modifications to the input spectrogram representation. These refinements enabled improved dereverberation quality, as reflected in key performance metrics. The model was trained using the parameters outlined in Table III, with a steady reduction in training loss, confirming effective learning. Notably, the RMSLE improved by 1.25%, decreasing from 40 to 39.5. The training progression, characterized by a consistent decline in loss and improved RMSLE over successive epochs, highlights the model's ability

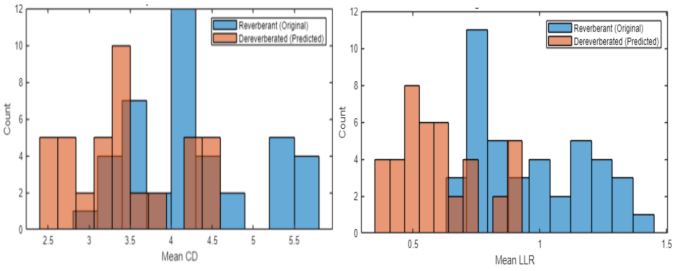

Fig. 6. Cepstrum Distance (CD) estimation and Log-Likelihood Ratio (LLR) between reverberated and dereverberated signals.

to generalize well to reverberant speech.

Figs. 6 illustrate the comparative analysis of speech signals in both time and frequency-domains. The dereverberated output exhibits improved alignment with the clean speech signal, demonstrating the model's capacity to recover speech quality from reverberant input. Furthermore, the STFT + TEO method yielded consistent improvements over the STFT-only approach, particularly in terms of Cepstrum Distance (CD) and Log-Likelihood Ratio (LLR). While direct numerical comparisons with prior work [19] are not applicable due to dataset differences, our results reaffirm the effectiveness of the U-Net framework. The proposed modifications to the input spectrogram contribute to improved performance, demonstrating that spectral domain refinements can enhance dereverberation quality across diverse datasets.

V. Summary and Conclusions

In sum, our model for speech dereverberation, which fuses spectrograms with a U-Net, has shown promising results but also exhibits limitations. The current method relies heavily on the power of convolutional networks in the frequency-domain, which may not consistently provide the most accurate outcomes across diverse acoustic environments or for different types of reverberation. Future directions for this field involve exploring alternative neural network architectures and machine learning algorithms that can offer more accurate or efficient dereverberation and separation. Additionally, further research is needed to improve the robustness of the current method in various acoustic conditions by developing more efficient algorithms for real-time processing, exploring techniques for handling non-linear or complex reverberations, and creating comprehensive evaluation metrics to assess performance. Moreover, open research problems remain, such as enhancing model robustness in highly variable acoustic environments, effectively addressing non-linear reverberation challenges, and developing standardized evaluation metrics that comprehensively capture performance across diverse scenarios. Our work is intended to inspire further research in these areas and contribute to the ongoing advancements in speech dereverberation and separation, paving the way for innovative solutions in this exciting field of study.

Acknowledgments

This work was supported by MeitY, Govt. of India (Project Grant ID: 11(1)2022-HCC (TDIL)).

We thank Dr. Rishabh Gupta and Dr. Raj Narayana Gadde from the Samsung R&D Institute, Bangalore (SRI-B), India, for their support as part of the Samsung PRISM Program.

References

[1] Madhu R. Kamble and Hemant A. Patil, "Detection of replay spoof speech using Teager energy feature cues," in Special issue on Advances in Automatic Speaker Verification Anti-spoofing, in Computer, Speech and Language, Elsevier, vol. 65, 101140, pp. 1-19, 2021.

[2] H. Kuttruff, *Room Acoustics*, 6^{th} ed. CRC Press, 2016.

[3] W. C. Sabine, *Collected Papers in Acoustics*. 1922.

[4] W. H. Sabine, *Collected Papers in Acoustics*. Dover, New York, 1964.

[5] J. Traer and J. H. McDermott, *Statistics of natural reverberation enable perceptual separation of sound and space*, Proceedings of the National Academy of Sciences (PNAS) 113, 48 (November 2016): E7856–E7865.

[6] E. Indenbom, N. C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, "DeepVQE: Real-time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-3, 2023.

[7] C. Chen, W. Sun, D. Harwath, and K. Grauman, "Learning audio-visual dereverberation," arXiv preprint arXiv:2106.07732, 2023 {Last Accessed on December 25, 2023}.

[8] D. León, F. Tobar, *Late reverberation suppression using U-nets*, arXiv:2110.02144 [eess.AS], 2021 {Last Accessed on January 5, 2024}.

[9] Hyeong-Seok Choi, Sungjin Park, Jie Hwan Lee, Hoon Heo, Dongsuk Jeon, Kyogu Lee *Real-time denoising and dereverberation with tiny recurrent U-Net*, arXiv:2102.03207 [cs.SD], 2021 {Last Accessed on January 5, 2024}.

[10] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with A reverberation time shortening target," arXiv preprint arXiv:2204.08765, 2022. {Last Accessed on December 25, 2023}.

[11] W. Zhang *et al*., *End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp-2, 2021.

[12] W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe and Y. Qian, *End-to-end dereverberation, beamforming, and speech recognition in a cocktail party*, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1-16, 2022.

[13] J. M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, *Customizable end-to-end optimization of online neural network-supported dereverberation for hearing devices*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp-1, 2022.

[14] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, *A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research*, *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no.7, pp. 5-6, 2016.

[15] C. Valentini-Botinhao, *Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]*, University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017, pp. 1–2. Available: https://datashare.ed.ac.uk/handle/10283/2031. {Last Accessed on January 5, 2024}.

[16] M. Maciejewski, G. Wichern, and J. Le Roux, *WHAMR!: Noisy and reverberant single-channel speech separation*, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, Barcelona, pp.2-3.

[17] K. Khoria, M. R. Kamble, and H. A. Patil, "Teager energy cepstral coefficients for classification of normal *vs.* whisper speech," in 28^{th} *European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, pp. 1-3, 2023.

[18] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA, 1990, pp. 381-384, vol.1.

[19] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, *Speech dereverberation using fully convolutional networks*, Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel, 2023.