DUal-NET: A Transformer-Based U-Net Model for Denoising Bone Conduction Speech

Yueyuan Sui
Northwestern University
Evanston, IL, USA
yueyuansui2024@u.northwestern.edu

Minghui Zhao
Columbia University
New York, NY, USA
mz2866@columbia.edu

Junxi Xia
Northwestern University
Evanston, IL, USA
junxixia2024@u.northwestern.edu

Yiting Zhang
Northwestern University
Evanston, IL, USA
yitingzhang2025@u.northwestern.edu

Xiaofan Jiang
Columbia University
New York, NY, USA
jiang@ee.columbia.edu

Stephen Xia
Northwestern University
Evanston, IL, USA
stephen.xia@northwestern.edu

Abstract

We propose "DUal-NET", a novel transformer-based model for enhancing speech capture through bone conduction headsets in human-centered sensing systems. As wearable bone-conduction devices become increasingly important for continuous health monitoring and ambient computing, they face a unique challenge: bone-conduction microphones can receive significant interference from speakers playing audio to the user, this occurs because the headset is directly in contact with the skull and induces vibrations similar to human speech, much like a user speaking, degrading speech recognition accuracy and communication quality. Existing state-of-art speech enhancement and sound source separation methods are 'blind' and assume that the interference noise is not available due to the inherent difficulty in observing clean correlated noise. By contrast, headsets have full knowledge of the sounds they play through their speakers, and DUal-NET takes advantage of this raw signal in its denoising process. We demonstrate that DUal-NET can significantly improve standard speech quality metrics over existing state-of-art methods in realistic scenarios (PESQ: 135%, STOI: 50%, LSD: 66%), enabling more accurate speech sensing for human-centered applications including health monitoring, personalized assistants, and augmented communication.

CCS Concepts

• **Computing methodologies** → **Speech processing**.

Keywords

speech enhancement, U-Net, transformer, bone-conduction, source separation

ACM Reference Format:
Yueyuan Sui, Minghui Zhao, Junxi Xia, Yiting Zhang, Xiaofan Jiang, and Stephen Xia. 2025. DUal-NET: A Transformer-Based U-Net Model for Denoising Bone Conduction Speech. In *The 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems (HumanSys '25), May 6–9, 2025, Irvine, CA, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3722570.3726887

HumanSys '25, Irvine, CA, USA

ACM ISBN 979-8-4007-1609-9/2025/05.
https://doi.org/10.1145/3722570.3726887

1 Introduction

With the rapid progression of intelligent wearable systems and the increasing emphasis on enhancing human well-being through technology, there is a growing demand for sensing technologies that can reliably capture human speech in diverse environments. Speech serves as a fundamental input modality for human-centered cyber-physical systems, enabling applications from health monitoring and activity recognition to personalized assistants and augmented communication.

The most common way to capture speech is through over-the-air (OTA) microphones, which convert changes in air pressure into electrical signals. While OTA microphones are widely used in devices such as earbuds and headphones, they are inherently susceptible to ambient noise. This sensitivity to ambient noise can significantly reduce the clarity and quality of captured speech, creating challenges in maintaining high-fidelity voice communications and command recognition in noisy environments [28].

To mitigate the limitations of OTA microphones' susceptibility to ambient noise, recent works have explored the use of bone conduction microphones (BCMs) and other vibration-based sensors for speech capture [8, 12, 22]. Unlike OTA microphones, BCMs are in direct contact with the head, making them highly sensitive to vibrations from the skull while speaking. This direct contact gives BCMs natural resilience to ambient noise. Additionally, vibration-based sensors such as accelerometers (ACCEL) have been explored to detect speech [9, 13] and facial movements for critical applications such as authentication [21]. The inherent robustness of bone conduction sensing methods to ambient noise provides a significant advantage over traditional OTA microphones [26].

While the implementation of BCMs or ACCELs can substantially mitigate the influence of environmental noise, the internal sounds produced by headphones, earbuds, and bone-conduction speakers can still induce vibrations in facial skin and bones that the BCM or ACCEL can sense. For instance, during virtual meetings, a BCM-equipped user may inadvertently capture not only their own voice but also the voices of other participants emanating from their headphones. This scenario can result in the concurrent recording of multiple voices, thereby affecting the integrity of captured speech [27]. This work focuses on removing vibration-induced noise in BCMs caused by (bone-conduction) speakers in headset wearables, as shown in Figure 1.

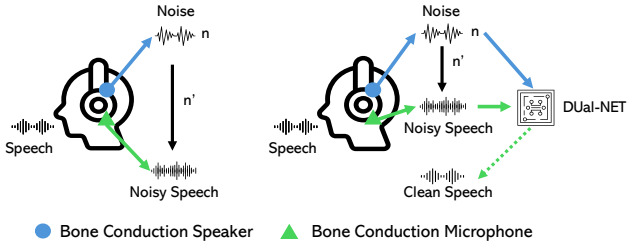

Figure 1: Left - We focus on eliminating interference from headset (bone-conduction) speakers in bone-conduction microphones. Right - We propose a novel deep learning based denoising architecture that, unlike existing sound source separation works, takes advantage of the headset's complete knowledge of the sound it plays through its speaker.

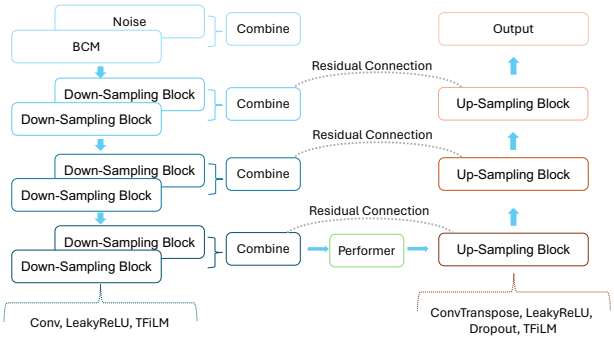

Figure 2: Model Architecture. DUal-NET takes advantage of the additional raw noise that is being played and accessible to headsets to improve denoising.

There have been many works that explore denoising, sound source separation, and speech enhancement methods to separate speech from background noise [1, 2, 6, 11, 14, 15, 19, 23, 25]. A common characteristic of these methods is their operation under the constraint of unknown noise signals, due to the inherent difficulty in capturing environmental noise accurately. In other words, these approaches typically only receive the noisy speech signal as input.

In this work, we take advantage of the fact that while speakers on the headset may induce bone and skin vibrations that interfere with BCM-recorded speech, the headset has complete knowledge of the sound it plays through the speaker. We propose DUal-NET, a novel transformer-based U-Net architecture that adaptively removes interference from (bone-conduction) speakers in bone-conduction speech. Unlike existing deep learning-based blind source separation works, DUal-NET leverages clean noise played through the headset to aid in removing interference observed in bone-conduction speech. This is similar to traditional signal processing-based adaptive filtering methods, such as least means squares (LMS) filtering, that attempt to find filter coefficients that minimize the error between the observed and desired signals. However, we demonstrate that DUal-NET can achieve up to a 79% improvement over these traditional adaptive filtering methods. Moreover, by incorporating knowledge of the interfering noise from the headset, DUal-NET achieves up to a 32% improvement over state-of-art deep learning-based denoising architectures. Our contributions are summarized as follows.

- We propose DUal-NET, a novel transformer-based U-Net architecture that denoises bone-conduction speech by taking advantage of the headset's knowledge of the interfering noise.
- We demonstrate that DUal-NET outperforms traditional adaptive filtering methods by up to 79% and state-of-art deep learning denoising architectures by up to 32%.
- We open-source all code, designs, and datasets[1].

2 Method

Figure 2 illustrates our denoising architecture. Inspired by TUNet[18], our architecture employs an enhanced U-Net framework, incorporating TFiLM in both the downsampling and upsampling layers, and utilizing an improved version of the Transformer, Performer, in the narrow bottleneck layer. To enable our model to simultaneously process both the noise and the noisy speech signals, we modify the downsampling procedure. Specifically, we perform separate downsampling for the noise and the noisy speech signals, then merge the corresponding downsampled layers. This merged information is subsequently used for inference, allowing for a more effective separation of speech from background noise.

2.1 Down-Sampling Block

Our model comprises a total of three layers, each containing two downsampling blocks. Each downsampling block includes a 1D convolutional layer that feeds into a layer of LeakyReLU activations, similar to other U-Net models. Downsampling block layers $b = 1, 2, 3$ contains 2^{5b} convolutional filters and a stride of 4, with convolution kernel sizes of 66, 18, and 8 respectively. Temporal Feature-Wise Linear Modulation (TFiLM) has been used to aid convolutional layers in grasping long-range dependencies. Acting as a normalization layer, TFiLM fuses max pooling and Long Short-Term Memory (LSTM) networks. The max pooling operation segments the temporal dimension into small parts, while the LSTMS capture extended dependencies.

To merge the corresponding downsampled layers, we employ a method akin to residual connections. However, instead of adding the noisy speech signal to the noise signal, we subtract the noise signal from the noisy speech signal. This subtraction yields the merged downsampled layer information, which is then utilized in the subsequent inference process.

2.2 Bottleneck

We integrate the Transformer architecture into the bottleneck of our U-Net structure. We posit that leveraging the Transformer architecture can significantly enhance the model's ability to capture global contextual information, thereby improving the overall understanding of the global features. Moreover, the bottleneck layer in the U-Net structure, characterized by the smallest feature maps and the most concentrated information, is also the layer most susceptible to detail loss. By incorporating the Transformer, we aim to mitigate

[1] https://github.com/IMEC-Northwestern/DUal-NET

this information loss and thus boost the overall performance of the model.

However, integrating the Transformer does introduce additional complexity and computational overhead. To address this, we opt for the improved version of the Transformer architecture, Performer, which is particularly advantageous due to its linear time complexity self-attention mechanism. This choice allows us to harness the benefits of the Transformer while maintaining computational efficiency.

2.3 Up-Sampling Block

As with the downsampling blocks, our model also comprises three upsampling blocks, each of which performs Transposed Convolutional Layer, LeakyReLU, dropout, and TFiLM. The upsampling block preceding the final output only performs Transposed Convolutional Layer.

Upsampling blocks contain transposed convolution filters of size 128, 64, and 1, with a stride of 4. The convolution kernel sizes of each convolution layer are 8, 18, and 66, respectively.

3 Evaluation

3.1 Synthesized Noisy Speech Dataset

We utilize the VCTK dataset [29], which comprises 44 hours of clean speech from 109 native English speakers. To create noisy speech files, we sourced royalty-free music of equivalent duration from YouTube and manually mixed these tracks with the VCTK dataset, resulting in a total of 44 hours of mixed audio. The average signal-to-noise ratio (SNR) of the synthesized noisy speech is -13.19dB, and the average short-time objective intelligibility (STOI) score is 0.3884.

3.2 Real-World Dataset

Similar to the previous scenario, we utilized the VCTK dataset for speech data and royalty-free music for noise. We employed an OTA speaker to simulate a human face. To replicate typical headset usage, we attached a bone-conduction speaker and microphone to the OTA speaker, as shown in Figure 3. VCTK speech data was played through the OTA speaker, while royalty-free music was transmitted via the bone-conduction speaker. We recorded approximately 9 hours of noisy speech data across four distinct noise volume levels, corresponding to SNRs of -57.17 dB, -21.42 dB, -17.69 dB, and -13.09 dB, and the average STOI score is 0.6397.

3.3 Preprocessing and Training

For training and validation, we divided the dataset into training and validation sets using a 9:1 ratio for all models requiring training. This ensures a robust evaluation while maintaining sufficient data for training the models effectively. Training data was divided into smaller segments with a window size of 512ms and 50% overlap. For all other methods, we applied the preprocessing steps mentioned in the respective papers.

For the synthesized noisy speech dataset, we train the models for 800 epochs. For the real-world dataset, due to the reduced size of the training dataset, we train the models for 30 epochs. After these time periods, we notice that the training, testing, and validation loss change very little. We trained all other methods using their default training parameters specified in their respective papers and codes.

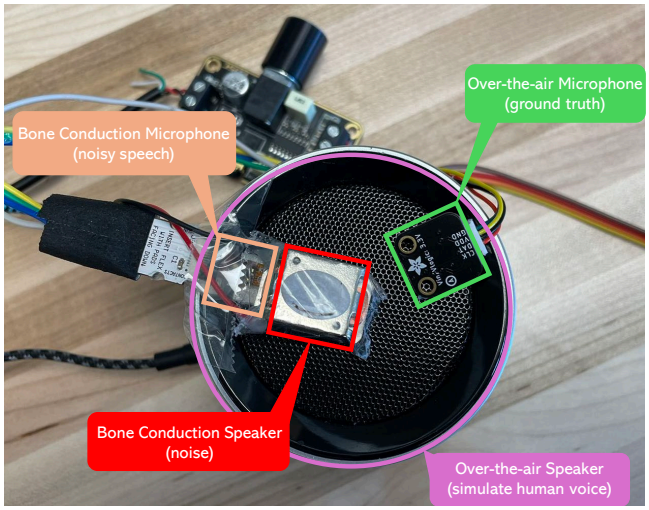

Figure 3: Real-world dataset capture setup.

3.4 Loss Function

In our study, we observed that employing mean squared error (MSE) as the loss function resulted in audio quality that was inferior to that produced by mean absolute error (MAE). Consequently, we opted for MAE as our primary loss function. Nonetheless, relying on MAE or any sample-by-sample distance metric, such as MSE, does not inherently ensure perceptual quality [17]. Therefore, we also integrated a multi-resolution STFT loss [30]. We utilized the same multi-resolution STFT loss parameters as outlined in [4, 16]: FFT bins set to $\in \{512, 1024, 2048\}$, hop lengths of $\in \{50, 120, 240\}$, and window sizes of $\in \{240, 600, 1200\}$.

3.5 Performance Metrics

Perceptual Evaluation of Speech Quality (PESQ). In the context of denoising, PESQ [20] serves as an objective measurement to evaluate the quality of denoised speech signals as perceived by human listeners. This metric compares the generated audio against a high-quality reference, with scores typically ranging from -0.5 to 4.5. A higher PESQ score indicates a closer distance to the original audio's quality.

Short-Time Objective Intelligibility (STOI). STOI [24] compares the clarity and intelligibility of speech enhanced from a lower to a higher resolution against a clean reference, with scores ranging from 0 to 1. A higher STOI score, closer to 1, signifies better intelligibility, indicating that the enhanced speech is easier to understand.

Log-Spectral Distance (LSD). This metric measures the distance between the spectrum of the generated speech and the clean speech on the log scale. The difference is measured in the log spectra because human hearing follows this scale [7]. Scores are calculated over all frequency bins, with lower values indicating a closer match to the reference spectrum.

	PESQ	STOI	LSD
LMS	1.2390	0.2596	3.3878
DCUNet-10	1.3852	0.5703	1.6150
DCUNet-20	1.5641	0.6329	1.4197
DCCRNet	1.4034	0.5718	1.5493
DUal-NET	**3.6827**	**0.9535**	**0.4802**

Table 1: Evaluation on synthesized noisy speech

	PESQ	STOI	LSD
LMS	1.2338	0.2914	4.9009
DCUNet-10	1.4620	0.7608	1.7883
DCUNet-20	1.6878	0.8150	1.5189
DCCRNet	1.7798	0.8199	1.5198
DUal-NET	**1.7489**	**0.8343**	**1.0268**

Table 2: Evaluation on recorded BCM Noisy Speech

$$\mathrm{LSD}x,y = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log Xt,k - \log \hat{X}t,k\right)^2}$$

Xt,k denotes the magnitude spectrum of the original signal at time frame t and frequency k. $\hat{X}t,k$ represents the magnitude spectrum of the super-resolved signal at the same time frame and frequency. T indicates the total number of time frames, and K is the total number of frequency bins.

3.6 Models Compared

- Least Mean Squares (LMS) [5] is an adaptive filtering algorithm designed to adapt to non-stationary signal environments by minimizing the mean square error between the output signal and the desired signal.
- DCUNet [3] is a state-of-art deep learning U-Net architecture that operates on the short-time frequency domain of noisy speech, without methods to leverage correlated noise while denoising speech. We leverage two versions: the smaller DCUNet-10 and the larger DCUNet-20.
- DCCRNet [10] is an advanced speech enhancement model that integrates a Densely-Connected Convolutional Recurrent Network (DCCRN) with a Correlated Noise Reduction (CNR) module to suppress both stationary and non-stationary noise in speech signals. The DCCRN component learns spectral-temporal features, while the CNR module leverages the correlation between noise components to further improve denoising performance.

DCUNet and DCCRNet are both single input denoising architectures that only take noisy signals as input.

3.7 Performance

Summary. Tables 1 and 2 summarize the testing performance of all methods on the synthesized dataset and the real-world dataset.

	Model Size	Inference Time	Memory Usage
DCUNet-10	1.42M	2.64ms	5.42 MB
DCUNet-20	3.52M	5.36ms	13.46 MB
DCCRNet	3.67M	6.87ms	14.00 MB
DUal-NET	**1.65M**	**3.15ms**	**6.31 MB**

Table 3: Model Resource Usage Evaluation

On the synthesized dataset, our approach (highlighted) demonstrates the best performance across all metrics. On the real-world dataset, DUal-NET outperforms all methods on all metrics except for DCCRNet, which only slightly outperforms DUal-NET in PESQ, which evaluates sound quality rather than intelligibility. DUal-NET has a higher STOI score, indicating that the intelligibility of the denoised speech is higher than that of DCCRNet. Additionally, DUal-NET significantly reduces the LSD, which means it is able to more accurately restore the speech spectrum, reducing signal distortion. Thus, DUal-NET still largely outperforms DCCRNet in terms of overall speech quality. This demonstrates that incorporating knowledge of the noise into the denoising process can greatly improve speech enhancement and denoising. Moreover, our transformer-based architecture outperforms traditional signal processing based methods, such as LMS, that also incorporate knowledge of the noise into the denoising process.

Model Size and Inference Time. We compared the model resource usage of four models on an NVIDIA L40 GPU using a 512ms window. As shown in Table 3, only DCUNet-10 has a slightly smaller model size and latency than DUal-NET, while DUal-NET generates significantly more intelligible and higher quality speech. DCCRNet, the model that generated speech with the most similar quality to DUal-NET, sees more than twice the memory usage and inference time as DUal-NET.These results demonstrate that DUal-NET achieves a balance between model complexity and computational efficiency, with an inference time comparable to state-of-the-art deep learning U-Net models.

Performance Gap. We observe that DUal-NET exhibits a performance gap between the synthesized and real-world datasets. This discrepancy is due to two key factors.

First, it is difficult to collect a dataset with sounds mixed non-artificially in the real world, which is a challenge for all sound source separation works. It is not possible to obtain ground truth for sound sources mixed in the air. As such, all source separation works collect clean sources, used as ground truth, and artificially mix them to use as model inputs. In our scenario, the noise source is being played through a bone conduction speaker, which in theory should not propagate very far in the air and impact the ground truth collected by the over-the-air microphone (highlighted in green in Figure 3). However, we found that significant amounts of residual noise still can be heard on the ground truth recordings.

Second, the transfer function caused by the impact of the skin and bones on the noise propagating from the bone conduction speaker to the bone conduction microphone is more complex than what we can explicitly model. As such, the transfer function between the speaker and microphone is easier for DUal-NET to learn and account for

in the synthesized scenario. On the real-world dataset, DUal-NET sees a much greater performance drop compared to other methods because, unlike other state-of-art deep learning methods, DUal-NET directly uses the raw noise for denoising. However, DUal-NET still provides significant speech quality improvements with greater efficiency (smaller model size and latency) by exploiting knowledge of the noise that is present aboard headphones and earphones.

One qualitative observation we noticed is that all methods, including DUal-NET removed a significant portion of speech on the real-world dataset, as speech and music often overlap in frequency. We plan to explore architectural improvements to address this commonplace problem in future work.

4 Conclusion

We present DUal-NET, a transformer U-Net model for denoising bone-conduction speech from interfering noises from a headset's speakers. Unlike state-of-art speech enhancement architectures, DUal-NET takes advantage of the headset's access to the raw interference signal to improve denoising. We demonstrate that incorporating this knowledge can improve performance by up to 32% in realistic scenarios.

In future work, we plan to validate our method on bone-conduction headsets collected on real humans. While we achieved significant improvements in removing interference, a significant portion of user speech was attenuated, which is common in many adaptive filtering methods. As such, we plan to explore further architectural improvements to preserve user speech while removing interference.

References

[1] Sherif Abdulatif, Ruizhe Cao, and Bin Yang. 2024. CMGAN: Conformer-based metric-GAN for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[2] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. 2017. Monoaural audio source separation using deep convolutional neural networks. In *Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13*. Springer, 258–266.

[3] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. 2018. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*.

[4] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847* (2020).

[5] Paul L Feintuch. 1976. An adaptive recursive LMS filter. *Proc. IEEE* 64, 11 (1976), 1622–1624.

[6] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*. PMLR, 2031–2041.

[7] Augustine Gray and John Markel. 1976. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 5 (1976), 380–391.

[8] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 14–27.

[9] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 14–27.

[10] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264* (2020).

[11] Zili Huang, Shinji Watanabe, Shu-wen Yang, Paola García, and Sanjeev Khudanpur. 2022. Investigating self-supervised learning for speech enhancement and separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6837–6841.

[12] Yuang Li, Yuntao Wang, Xin Liu, Yuanchun Shi, Shwetak Patel, and Shao-Fu Shih. 2022. Enabling Real-Time On-Chip Audio Super Resolution for Bone-Conduction Microphones. *Sensors* 23, 1 (2022), 35.

[13] Yunji Liang, Yuchen Qin, Qi Li, Xiaokai Yan, Zhiwen Yu, Bin Guo, Sagar Samtani, and Yanyong Zhang. 2022. AccMyrinx: Speech Synthesis with Non-Acoustic Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–24.

[14] Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 696–700.

[15] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[16] Moshe Mandel, Or Tal, and Yossi Adi. 2023. Aero: Audio super resolution in the spectral domain. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[17] Juan Manuel Martin-Donas, Angel Manuel Gomez, Jose A Gonzalez, and Antonio M Peinado. 2018. A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal processing letters* 25, 11 (2018), 1680–1684.

[18] Viet-Anh Nguyen, Anh HT Nguyen, and Andy WH Khong. 2022. Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 161–165.

[19] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. 2016. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 9 (2016), 1652–1664.

[20] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.

[21] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. Muteit: Jaw motion based unvoiced command recognition using earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[22] Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen Xia. 2024. TRAMBA: A Hybrid Transformer and Mamba Architecture for Practical Audio and Bone Conduction Speech Super Resolution and Enhancement on Mobile and Wearable Platforms. *arXiv preprint arXiv:2405.01242* (2024).

[23] Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen Xia. 2024. TraMSR: Transformer and Mamba based Practical Speech Super-Resolution for Mobile Wearables. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 1686–1688.

[24] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 4214–4217.

[25] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. 2018. *Audio source separation and speech enhancement*. John Wiley & Sons.

[26] Stephen Xia and Xiaofan Jiang. 2020. Pams: Improving privacy in audio-based mobile systems. In *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*. 41–47.

[27] Stephen Xia and Xiaofan Jiang. 2022. Ava: An adaptive audio filtering architecture for enhancing mobile, embedded, and cyber-physical systems. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 118–131.

[28] Stephen Xia, Jingping Nie, and Xiaofan Jiang. 2021. Csafe: An intelligent audio wearable platform for improving construction worker safety in urban environments. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)*. 207–221.

[29] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. https://doi.org/10.7488/ds/2645.

[30] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203.