Personalized Bone-Conduction Bandwidth Extension with Speaker Characteristics

Pan Xu*, Zhongyu Zhang† and Zhonghua Fu*†
* Northwestern Polytechnical University, Xi'an, China
E-mail: hp59497@mail.nwpu.edu.cn, mailfzh@nwpu.edu.cn
†iFLYTEK Xi'an Superbrain Information Technology Co., Ltd., Xi'an, China
E-mail: zyzhang80@iflytek.com

***Abstract*— To address the challenges of high-frequency attenuation and speaker variability in bone-conducted (BC) speech, this paper proposes a personalized feature-integrated bandwidth extension (BWE) method. The proposed approach employs a dual-encoder architecture that extracts time-frequency representations from low-quality BC signals and personalized features from air-conducted (AC) reference speech. These features are fused via an attention mechanism and jointly optimized with a speaker classification task. Experiments on both simulated and real-world datasets demonstrate that the proposed method significantly outperforms existing baseline methods in terms of speech quality, spectral fidelity, and intelligibility. These results validate the effectiveness of the approach in complex acoustic environments and offer a novel direction for BC speech enhancement.**

I. INTRODUCTION

Capturing high-quality speech in noisy or complex environments remains a major challenge in speech enhancement [1]. Traditional air-conducted (AC) microphones often struggle under heavy noise or occlusion. They are easily affected by environmental sounds, echoes, and other interferences [2]. Recently, bone conduction (BC) headphones have attracted growing interest. Unlike conventional AC devices, BC headphones offer an open-ear design, greater comfort, and better environmental awareness, making them ideal for sports, outdoor communication, and hearing assistance [3]. BC devices work by transmitting vibrations through the skull directly to the cochlea, thereby bypassing the outer and middle ears [4]. This unique mechanism ensures stable speech perception, even in cases of middle-ear disorders or high-noise conditions. As a result, BC technology shows promise for hearing-impaired users [5] and robust speech communication in noisy environments [6].

However, bone-conducted (BC) speech signals often suffer from severe high-frequency attenuation (>2 kHz) during transmission due to tissue impedance and mechanical damping, leading to degraded speech intelligibility [7]. Furthermore, nonlinear factors including sensor placement variations, skull structure differences and individual physiological characteristics cause signal instability [8], limiting BC speech applications in high-fidelity communication systems.

To address this issue, researchers have explored bone-conducted speech bandwidth extension (BWE) techniques. These methods aim to restore the missing high-frequency components using data-driven approaches, thereby improving speech clarity and naturalness [9]. Early BWE methods mainly relied on statistical models, such as Gaussian Mixture Models (GMM) and spectral envelope mapping techniques [10], to learn the spectral relationship between BC and air-conducted (AC) speech. However, these methods often showed limited accuracy in spectral modeling and poor generalization across speakers. Recently, end-to-end models based on Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have been applied in this field [11,12,13]. While these models achieved better performance, most of them failed to consider speaker-specific characteristics. As a result, the reconstructed speech lacks naturalness, and the models often have complex structures with limited generalization ability.

For speaker-specific feature representation, existing approaches commonly adopt techniques such as i-vector [14] and x-vector [15] to extract speaker embeddings from speech signals. Although these methods effectively capture speaker identity, the extracted features may not be fully relevant to the BWE task. Moreover, introducing a separate speaker embedding module increases the overall model complexity. U-Net is a classical convolutional neural network based on an encoder–decoder architecture [16]. In this work, we observe that the deep encoder in U-Net can implicitly capture global acoustic patterns, such as pronunciation habits. Its symmetric structure, equipped with skip connections, preserves both high-frequency details and speaker characteristics, making it a suitable framework for personalized BWE. In practical applications, the system obtains speaker-specific prior information by collecting enrollment samples (clean air-conducted speech) from the target user.

To this end, we propose a personalized bone-conducted BWE method (PBC-BWE) based on VoiceFixer(VF) [17], an advanced speech restoration framework. The U-Net structure employed in the analysis stage of VF not only supports

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

effective speech enhancement but also implicitly captures speaker characteristics, without increasing model complexity. The main contributions of this work are as follows:

• Lightweight personalized feature extraction: We propose to reuse the VF encoder to construct a lightweight feature extractor, ensuring model compactness and relevant acoustic features for the bandwidth extension task.

• Optimized attention-based fusion mechanism: Dot-product attention module is proposed to dynamically fuse personalized speech features with spectral reconstruction features through learnable attention weights, enhancing the individual adaptability of the reconstructed output.

• Speaker classification with multi-task learning: A parallel speaker classification branch is introduced under a multi-task learning framework. This enhances the discriminability and robustness of speaker representations, ensuring that the personalized features are highly speaker-dependent.

II. METHOD

A. Model Architecture

This study extends the VF model [17] by adding a feature extraction module, a feature fusion module, and a speaker classification module, as shown in Fig. 1. Modules marked with blue boxes are used during training, while those marked in red are used during inference. Personalized features from the adaptation layer are saved offline for direct use in the attention module during inference. Compared to the original VF model, only the attention module is added during inference. The model follows a two-stage processing flow. In the analysis stage, a dual-encoder structure is used: the main encoder processes low-quality Mel spectrogram, and the feature encoder extracts speaker-specific features from clean enrollment speech. Both use the same pretrained backbone to ensure consistent feature mapping. In the synthesis stage, waveform reconstruction is performed using a HiFi-GAN vocoder [18]. Both the VF model and HiFi-GAN are optimized for 16 kHz sampling rate.

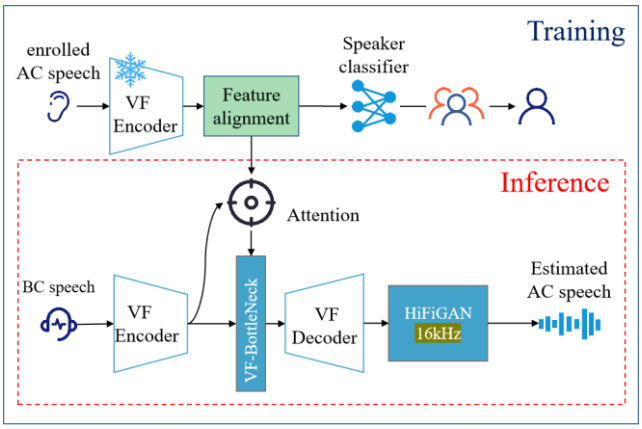

Fig.1 The overview of the proposed method.

B. Personalized Feature Encoder

Based on a pre-trained U-Net architecture, we reuse and freeze its encoder parameters to leverage the pre-trained weights. This encoder extracts personalized features that are stored offline for efficient access during training, thus avoiding additional trainable parameters. Experimental results show that incorporating these features improves the reconstruction quality of high-frequency components.

During the training preparation stage, each target speaker is provided with enrolled AC speech. The signal is processed using Short-Time Fourier Transform (STFT) and a Mel filter bank to generate a Mel spectrogram, which is then fed into the feature extractor to produce speaker embeddings for offline use. During training, the system selects the corresponding offline feature for the current speaker and refines it using a four-layer residual convolution block (as shown in Fig. 2). This produces speaker-adaptive features that retain speaker-specific characteristics and integrate effectively with the main network.

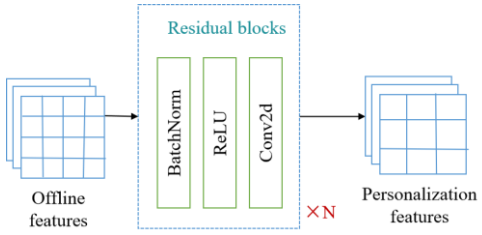

Fig. 2 The structure of the Feature aligning layer.

C. Attention-Based Fusion Module

The Scaled Dot-Product Attention mechanism [19] is commonly used to model dependencies within sequences. Its core idea is to compute attention weights by forming a query matrix $\boldsymbol{Q}$ and a set of key-value pairs $\boldsymbol{K}$ and $\boldsymbol{V}$, as shown in (1). Notably, our method alters the computation order of attention, shifting its goal from weighted summation to feature space alignment. We transform the $\boldsymbol{V}$ matrix using a covariance-style matrix derived from $\boldsymbol{K}$ and $\boldsymbol{Q}$, ensuring the output aligns with the query's feature characteristics and provides a more coherent representation for downstream tasks.

$$\mathrm{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \boldsymbol{V}\cdot \mathrm{softmax}\left(\frac{\boldsymbol{K}^{\boldsymbol{T}}\boldsymbol{Q}}{\sqrt{d_k}}\right), \tag{1}$$

here, $\sqrt{d_k}$ is scaling factor used to stabilize the distribution of attention scores. To enhance subjective quality and speaker consistency in BC BWE, we design a personalized feature fusion module based on attention mechanism, as shown in Fig.3.

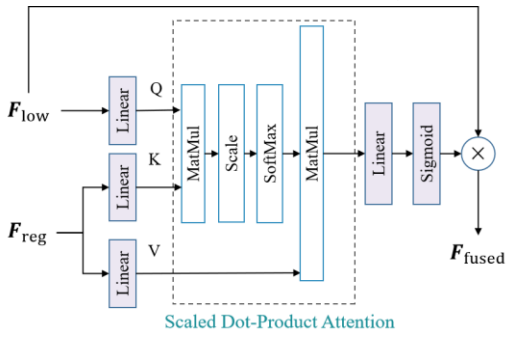

Fig.3 Attention-based feature fusion module.

Specifically, the local time-frequency features of the low-quality speech are linearly transformed to form the query matrix $\boldsymbol{Q}$. Similarly, the personalized features are used to construct the key-value pair matrices, $\boldsymbol{K}$ and $\boldsymbol{V}$, respectively. The scaled dot-product attention mechanism is then applied to dynamically fuse these two feature types. Let the time-frequency representation of input low-quality speech be denoted as $\boldsymbol{F}_{\text{low}} \in \mathbb{R}^{\text{B}\times\text{C}\times\text{T}_\text{q}\times\text{F}}$, where B is the batch size, C is the original channel dimension, T_q and F are the temporal and frequency dimensions, respectively. After linear mapping, it becomes $\boldsymbol{F}_\text{l} \in \mathbb{R}^{\text{B}\times\text{C}_1\times\text{T}_\text{q}\times\text{F}}$. Likewise, the personalized feature is denoted as $\boldsymbol{F}_{\text{reg}} \in \mathbb{R}^{\text{B}\times\text{C}\times\text{T}_\text{k}\times\text{F}}$, and is mapped to $\boldsymbol{F}_\text{r} \in \mathbb{R}^{\text{B}\times\text{C}_1\times\text{T}_\text{k}\times\text{F}}$, where C_1 is the transformed channel dimension and T_k is the temporal length of the personalized feature.The fused features are passed through a fully connected layer for further transformation, as described in (1) and (2):

$$\boldsymbol{F}_1 = \mathbf{W} * \text{A}(\boldsymbol{F}_\text{l}, \boldsymbol{F}_\text{r}, \boldsymbol{F}_\text{r}) + \mathbf{b}, \tag{2}$$

where $\mathbf{W}$ denotes the weight matrix, $\mathbf{b}$ represents the bias term, $\boldsymbol{F}_1$ denotes the fused feature, with the same shape as F_{low}. The fused feature $\boldsymbol{F}_1$ is passed through an activation function to generate a weight vector, which is then applied to $\boldsymbol{F}_{\text{low}}$ for weighted output, as defined in (3), where $\sigma(\cdot)$ is the sigmoid activation function.

$$\boldsymbol{F}_{\text{fused}} = \sigma(\boldsymbol{F}_\text{l}) * \boldsymbol{F}_{\text{low}}. \tag{3}$$

This design achieves feature fusion via an attention mechanism, where the activation function projects the extracted correlation information onto the main-path features as a score matrix. To better preserve the original speech characteristics, a residual connection is formed by element-wise addition between the fused feature $\boldsymbol{F}_{\text{fused}}$ and the original input feature $\boldsymbol{F}_{\text{low}}$, enhancing representational stability and robustness.

D. Speech classifier

This paper employs a multi-class classification task [20] for speaker identification. We innovatively integrate a speaker classification module with a BC bandwidth extension (BWE) system. The classifier architecture is shown in Fig. 4.

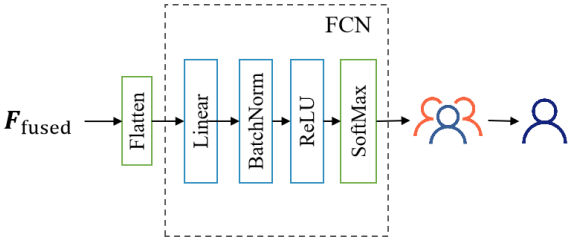

Fig. 4 Speaker classifier.

The classifier takes the deep features extracted by the encoder as input and outputs the probability distribution over speaker classes. Specifically, for a given input sample x, the predicted probability of belonging to class k is computed as:

$$p_k = \frac{e^{z_k}}{\sum_{j=1}^{\text{K}} e^{z_j}}, \tag{4}$$

where z_k denotes the output score for class k, and K is the total number of classes. The final predicted label corresponds to the class with the highest probability.

Furthermore, the system employs a joint training strategy. The extracted personalized features are processed in parallel through two branches:

1) Spectrogram Reconstruction Branch: A cross-attention mechanism enables interaction between the personalized features and the main encoder outputs. The resulting fused features replace the original encoder outputs and are fed into the decoder to reconstruct the Mel spectrogram. The L1 loss is applied to constrain the spectral reconstruction accuracy:

$$\mathcal{L}_1 = \left\| \widehat{\boldsymbol{S}}_{\text{mel}} - \boldsymbol{S}_{\text{mel}} \right\|_1, \tag{5}$$

where $\widehat{\boldsymbol{S}}_{\text{mel}}$ and $\boldsymbol{S}_{\text{mel}}$ denote the estimated and target Mel spectrogram, respectively.

2) Speaker Classification Branch: the personalized features are projected through a linear layer followed by a classifier to predict the speaker identity. The cross-entropy loss is used to enhance the discriminability of speaker features:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{\text{N}} y_i \log(\hat{y}_i), \tag{6}$$

where y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability for the i-th speaker class. To optimize model training, we propose a multi-task loss function, as shown in (7), where the weighting factor λ is set to 0.05.

$$\mathcal{L} = \mathcal{L}_1 + \lambda * \mathcal{L}_{cls}. \tag{7}$$

III. EXPERIMENTAL SETUP

A. Datasets

In this study, we primarily employ the ABCS Mandarin speech corpus [21] as the experimental dataset, which contains 47,182 paired recordings of air-conducted (AC) and bone-conducted (BC) speech collected from 100 speakers (50 male and 50 female). The speech content originates from two sources: the RASC863 corpus provided by the Chinese Academy of Social Sciences and a set of 30,000 daily conversational utterances provided by iFLYTEK. Each speaker recorded approximately 25 minutes of speech, resulting in a total duration of 42 hours. All recordings were originally sampled at 44.1 kHz and later downsampled to 16 kHz for processing.

The dataset was divided into training, validation, and test subsets. The training set consists of daily conversational recordings from all speakers (38 hours). For evaluation, recordings from the RASC863 corpus were exclusively used, ensuring no overlap in content with the training data. Specifically, recordings from 8 speakers (2 hours) formed the validation set, while another 8 speakers' recordings (2 hours) comprised the test set. To support personalized speech processing, a 30-second air-conducted registration utterance was randomly extracted from each speaker and used to derive offline speaker embeddings.

To systematically evaluate the generalization capability of the proposed method, we conducted comparative experiments using the VCTK multi-speaker corpus [22], focusing on evaluating the feature extraction and reconstruction performance in the Mel-spectral domain. The dataset comprises 96 kHz recordings from 110 English speakers with diverse accents. All audio was standardized to 16-bit PCM format and downsampled to 48 kHz. For data partitioning, the training set contains the first 300 utterances from 108 speakers (excluding speakers P280 and P315 due to recording issues), while the test set retains the last 225 utterances from 8 independent speakers (P360-P364, P374, P376, and S5) for rigorous cross-speaker generalization evaluation. In accordance with the data simulation strategy of the VF model, the test data were downsampled to simulate extreme bandwidth-limited conditions at 1 kHz, 2 kHz, and 4 kHz. For training data augmentation, random low-pass filtering with cutoff frequencies varying from 0.5 kHz to 44.1 kHz was applied, along with controlled additive noise. The noise samples were sourced from the TUT2018 Urban Acoustic Scenes dataset [23], which contains high-fidelity recordings (44.1 kHz sampling rate, 24-bit resolution) from a variety of acoustic environments. The signal-to-noise ratio (SNR) was dynamically adjusted during training within a range of -5 dB to 40 dB to enhance model robustness.

B. Experimental Configuration

To ensure consistent model training, we standardized the model configuration and training parameters. All window functions were implemented using the Hann window. For the ABCS bone-conducted speech dataset, the window length, frame shift, and Mel filterbank size were set to 512, 128, and 80, respectively, while for the VCTK dataset the corresponding values were 2048, 441, and 128.

For training, the model was trained using the Adam optimizer ($\beta1$=0.5, $\beta2$=0.999, ϵ=1e-9) with an initial learning rate of 5×10^{-4}. A batch size of 24 was used, and the training was conducted for a total of 5×10^{5} steps. The learning rate was decayed by a factor of 0.9 every 40k steps, with 100 warm-up steps at the beginning of training. Each input segment had a duration of 1.0 s, and approximately 300 hours of data were used per epoch.

C. Ablation Study

To evaluate the effectiveness of the speaker classification module, we conducted an ablation study by removing both the feature adaptation layer and the speaker classification branch from the original PBC-BWE framework. In this variant, the output of the feature encoder is directly fused with the speech time-frequency features without joint optimization. We refer to this simplified model as PF-BWE. Compared to PBC-BWE, the PF-BWE model eliminates the need for joint training and contains fewer parameters, while still relying on pre-extracted offline speaker features. The training strategy remains consistent with the original setup to ensure fair comparison. In addition, we designed a speaker embedding experiment based on ShuffleNet (Spk-BWE) [24]. To ensure the extracted features were directly comparable and relevant to our BWE task, the model architecture and training procedure for Spk-BWE were deliberately kept consistent with those of the PBC-BWE model, which leverages the shared encoder of our main U-Net for feature extraction.

IV. EXPERIMENTS RESULTS

A. Evaluation Metrics

Given the maturity of current vocoder technologies, this study focuses on the analysis stage, specifically on personalized feature modeling and the model's ability to restore low-quality speech representations. To evaluate this, we adopt several objective metrics in the Mel domain, such as log-spectral distance (Mel-LSD) [25], spectral-domain scale-invariant signal-to-noise ratio (Mel-SISPNR) [26], and Structural similarity index (Mel-SSIM) [27]. Additionally, we incorporate commonly used intelligibility and perceptual quality metrics such as PESQ [28], LSD [25], and STOI [29] to provide a comprehensive evaluation of the reconstructed speech. These metrics are chosen for their sensitivity to both spectral structure and perceived quality, offering an effective assessment of the model's feature-level restoration performance.

B. Experimental Results Analysis

Through systematic experiments involving both simulated data and real bone conduction data, this study comprehensively validated the performance of the proposed PBC-BWE model.

PBC-BWE achieved consistent advantages across 2kHz, 4kHz, and 8kHz sampling rates. At the challenging 2kHz condition, it achieved a Mel-LSD of 0.52, a 16.1% reduction versus the VF baseline, while improving Mel-SISPNR by 8.7%. The optimal performance was achieved at 8 kHz, with a Mel-SSIM of 0.88. Experimental results are presented in TABLE I.

TABLE I: Evaluation Results on the VCTK Corpus.

Sampling Rate	Model	Mel-LSD	Mel-SISPNR	Mel-SSIM
2kHz	VoiceFixer	0.62	11.39	0.67
	PF-BWE	0.63	11.30	0.67
	PBC-BWE	**0.52**	**12.38**	**0.73**
4kHz	VoiceFixer	0.53	13.11	0.75
	PF-BWE	0.56	12.92	0.75
	PBC-BWE	**0.42**	**14.43**	**0.81**
8kHz	VoiceFixer	0.42	15.06	0.84
	PF-BWE	0.45	14.82	0.84
	PBC-BWE	**0.33**	**16.54**	**0.88**

On the real bone conduction test set (ABCS corpus), PBC-BWE also demonstrated excellent performance. TABLE II shows that PBC-BWE outperforms all baselines across evaluation metrics. It achieves a PESQ score of 1.97, a 38.7% improvement over the original data and better than VF (1.67). Spectrally, it reduces time-domain LSD by 8.4% and Mel-LSD by 29.5% compared to VF, indicating superior reconstruction

TABLE II: Objective Evaluation Results on the ABCS Corpus. The best performance numbers are shown in boldface.

Dataset	Method	PESQ ↑	LSD ↓	STOI ↑	Mel-LSD ↓	Mel-SISPNR↑	Mel-SSIM↑
ABCS corpus	Raw Data	1.42	1.36	0.69	1.22	8.85	0.62
	VoiceFixer	1.67	1.09	0.83	0.61	15.09	0.82
	PF-BWE	1.65	1.11	0.83	0.62	14.19	0.82
	PBC-BWE	**1.97**	**1.00**	**0.89**	**0.43**	**17.78**	**0.93**
	Spk-BWE	1.85	1.18	0.83	0.60	13.41	0.90

accuracy. For intelligibility, PBC-BWE attains an STOI of 0.89, a 30.0% gain over the original, and its Mel-SISPNR improves by 100.9%, confirming its effectiveness in preserving clarity and Mel-domain features. The Mel-SSIM score of 0.90, a 9.8% increase over the baseline, further validates its strength in maintaining speech structure.

The experiment on Spk-BWE demonstrated that using speaker embedding features alone yielded limited performance gains, which strongly underscores the effectiveness of the U-Net backbone network in its own right. In addition, the ablation studies reveal that the PF-BWE model (without the speaker classification module) maintains comparable baseline performance to VF, confirming that the fundamental reconstruction ability remains unaffected. The complete PBC-BWE model shows 30.6% and 25.3% improvements over PF-BWE in Mel-LSD and Mel-SISPNR metrics respectively, validating the speaker module's critical role in personalized high-frequency reconstruction.

As shown in Fig. 5, PBC-BWE generates Mel spectrogram with more complete high-frequency structures and superior energy continuity compared to VF's spectral deficiencies. The enhancement is particularly pronounced in the 2-4kHz intelligibility-critical band as indicated by the orange box.

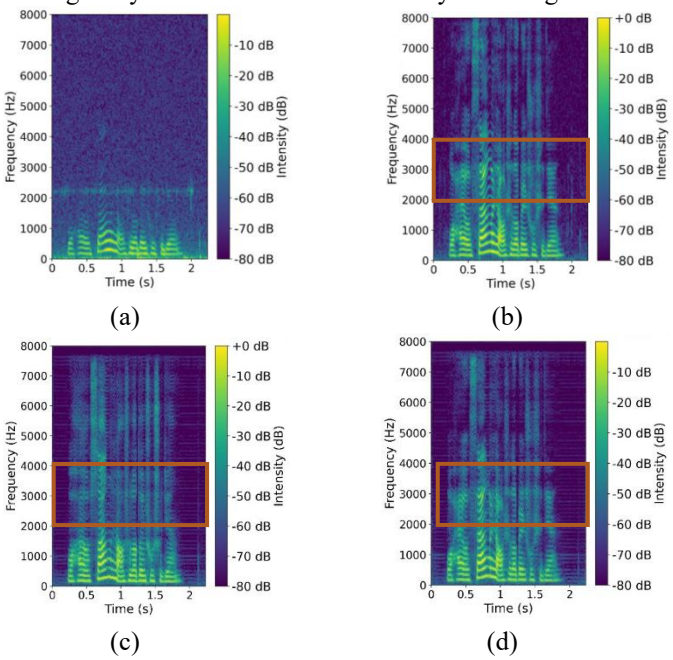

Fig. 5 Spectrograms of BC BWE task results. (a) BC speech; (b)AC speech; (c) VoiceFixer; (d) PBC-BWE.

V. CONCLUSION

This paper presents PBC-BWE, a personalized bandwidth extension model for bone-conducted speech. By integrating speaker-specific features, attention-based fusion, and an auxiliary speaker classification task, the model effectively restores high-frequency components while enhancing speaker consistency. Experimental results show consistent improvements over baseline methods across objective metrics such as PESQ, STOI, and mel-domain scores. The ablation studies confirm the individual contributions of each component. Future work will focus on further reducing model complexity for real-time applications and enhancing robustness in real-world and cross-lingual deployment scenarios.

REFERENCES

[1] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.

[2] H. Wang, X. Zhang and D. Wang, "Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 3134-3143, 2022, doi: 10.1109/TASLP.2022.3209943.

[3] Ellsperman SE, Nairn EM, Stucken EZ. Review of Bone Conduction Hearing Devices. Audiol Res. 2021 May 18;11(2):207-219. doi: 10.3390/audiolres11020019. PMID: 34069846; PMCID: PMC8161441.

[4] Stenfelt S. Acoustic and physiologic aspects of bone conduction hearing. Adv Otorhinolaryngol. 2011;71:10-21. doi: 10.1159/000323574. Epub 2011 Mar 8. PMID: 21389700.

[5] Tjellström A, Håkansson B, Granström G. Bone-anchored hearing aids: current status in adults and children. Otolaryngol Clin North Am. 2001 Apr;34(2):337-64. doi: 10.1016/s0030-6665(05)70335-2. PMID: 11382574.

[6] Li Mingzi, Cohen Israel, Mousazadeh Saman . Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones[C]//IEEE China Summit International Conference on Signal and Information Processing.2014.

[7] M. S. Rahman, A. Saha and T. Shimamura, "Low-frequency band noise suppression using bone conducted speech," Proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, 2011, pp. 520-525, doi: 10.1109/PACRIM.2011.6032948.

[8] Y. Pan, J. Zhou, H. Wang, W. Zheng, L. Tao, and H. K. Kwan, "Enhancing bone-conducted speech with spectrum similarity metric in adversarial learning," Speech Communication, vol. 170, p. 103223, 2025, doi: 10.1016/j.specom.2025.103223.

[9] C. Li, F. Yang and J. Yang, "A Two-Stage Approach to Quality Restoration of Bone-Conducted Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 818-829, 2024, doi: 10.1109/TASLP.2023.3337988.

[10] Trung P. N., Unoki M., Akagi M. "A study on restoration of bone conducted speech in noisy environments with LP-based model and Gaussian mixture model," Journal of Signal Processing, 2012: 409–417.

[11] F. Gao, M. Fu, and H. Meng, "Speech enhancement with bone-conducted and air-conducted microphone signals using deep autoencoder," IEEE Signal Processing Letters, vol. 26, no. 5, pp. 774–778, 2019.

[12] C. Zheng, J. Yang, X. Zhang, T. Cao, M. Sun, and L. Zheng, "Bandwidth extension WaveNet for bone-conducted speech enhancement," in Proc. 7th Conf. Sound Music Technol., 2020, pp. 3–14.

[13] C. Zheng, T. Cao, J. Yang, X. Zhang, and M. Sun, "Spectra restoration of bone-conducted speech via attention-based contextual information and spectro-temporal structure constraint," IEICE Trans. Fundam., vol. E102.A, no. 12, pp. 2001–2007, Dec. 2019.

[14] N. Dehak, P. J. Kenny et al. "Front-End Factor Analysis for Speaker Verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, May 2011, doi: 10.1109/TASL.2010.2064307.

[15] D. Snyder, P. Ghahremani et al. "Deep neural network-based speaker embeddings for end-to-end speaker verification," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 2016, pp. 165-170, doi: 10.1109/SLT.2016.7846260.

[16] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich, Germany: Springer, 2015: 234–241.

[17] Liu, Haohe et al. "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration." ArXiv abs/2204.05841 (2022): n. pag.

[18] Kong, Jungil, Jaehyeon Kim and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." ArXiv abs/2010.05646 (2020): n. pag.

[19] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[20] Bishop C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[21] M. Wang, J. Chen, X. -L. Zhang and S. Rahardja, "End-to-End Multi-Modal Speech Recognition on an Air and Bone Conducted Speech Corpus," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 513-524, 2023, doi: 10.1109/TASLP.2022.3224305.

[22] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.

[23] T. Heittola, A. Mesaros, and T. Virtanen, "TUT Urban Acoustic Scenes 2018, Development dataset," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1228142.

[24] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6848–6856).

[25] Heming Wang and DeLiang Wang. Towards robust speech super-resolution. IEEE/ACM Transac-tions on Audio, Speech, and Language Processing, 2021.

[26] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–half-baked or welldone? In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, pp.626–630, 2019.

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, pp. 600–612, 2004.

[28] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P. 862, 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.