Narrative Review

Speaker Adaptation Using Deep Neural Networks For Speech Enhancement

Chidambar B, D Hanumanth Rao Naidu

[1]Department of Mathematical and Computational Sciences, Sri Sathya Sai University for Human Excellence, Karnataka, India.

Abstract: The performance of Deep Neural Networks (DNN) based speech enhancement techniques often degrades when encountering a speaker mismatch between the training and testing conditions. The acoustic properties of speech vary significantly across different speakers. Speaker adaptation in DNNs aims to address this challenge by adapting the networks to individual speaker characteristics using smaller amounts of training data. Recent research has demonstrated that the performance of DNN-based speech enhancement techniques improves when they are adapted to specific speakers' spectral characteristics. This paper presents an overview of the current methodologies and advancements in speaker adaptation for DNNs. It also contributes to the understanding of how various adaptation strategies are employed to different DNN architectures. We highlight the strengths and weaknesses of each adaptation strategy and provide recommendations for achieving optimal performance.

Key Words: speaker adaptation; deep neural networks; speech enhancement.

Correspondence: chidambar.b@sssuhe.ac.in; hanumanth.naidu@sssuhe.ac.in

Citation: Chidambar B. Naidu D. H. R.; *Awareness* 2 (1): 42-50

Academic Editor: Dr. H. A. Semetko

Received: July 28, 2024

Revised: Dec. 13,2024

Accepted: Dec. 23,2024

Published: Jan. 14, 2025

1. Introduction

Model based speech enhancement approaches, such as the hidden Markov model (HMM), codebooks and DNNs, which rely on a trained model of speech data, have demonstrated superior results compared to traditional techniques like spectral subtraction and statistically-based methods, particularly in the presence of non-stationary noise conditions [1]. Among these, DNNs have gained prominence due to their exceptional ability to handle complex noise environments and superior acoustical modeling of speech. Despite the successes of DNN-based speech enhancement techniques, they often perform poorly when encountering speakers not seen during training [2]. The "one size fits all" approach of DNN-based speech enhancement falls short when confronted with diverse acoustical speech patterns, leading to suboptimal performance.

Speaker adaptation is a technique used in speech processing to improve performance by adapting a speech model to a target speaker's characteristics. This approach is crucial since the variability in speech among different speakers significantly impacts the accuracy of speech processing systems, such as speech recognition, speaker verification, speaker identification, and speech enhancement.

Speaker adaptation in DNNs aims to adapt DNNs to individual speakers' spectral characteristics, thereby providing personalized speech enhancement. The motivation for speaker adaptation in speech enhancement is driven by its potential impact on a wide range of applications such as mobile phones, teleconferencing systems [3], and hearing assistive devices [4], enhancing the quality and intelligibility of speech in a personalized manner, thereby improving the user experience.

The scope of this review is centered on exploring various speaker adaptation techniques in DNNs used for speech enhancement applications. This paper will assess the methodologies used for speaker adaptation, and highlight the improved speech enhancement performance results in recent research.

The remainder of the paper is organized as follows: Section 2 provides a discussion on general speaker adaptation techniques. Section 3 describes the recent advances in speaker adaptation techniques using DNNs for speech enhancement. The challenges are presented in Section 4. Finally, Section 5 presents the conclusions.

2. Traditional Speaker Adaptation Algorithms

Speaker adaptation of speech data models has been a topic of research for several decades, with speaker verification, speaker recognition, and speech recognition being the primary fields of application. The initial methods of speaker adaptation focused on Vector Quantization (VQ) models [5]. Later, the focus shifted to Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) methods. The early spectral models laid the groundwork for MLLR [6], which adjusts acoustical model parameters based on adaptation data to better match the spectral characteristics of the target speaker. MLLR adapts the mean vectors and covariance matrices of Gaussian components as follows:

$$\hat{\mu}_s = A_s \mu + b_s, \tag{1}$$

$$\hat{\Sigma}_s = H_s \Sigma H_s^T. \tag{2}$$

MLLR and similar approaches [7, 8] are prevalent in speaker recognition. Gauvain and Lee introduced MAP estimation for adapting HMM/Gaussian Mixture Models (GMM), merging prior knowledge from speaker independent models with new speaker data to optimize model parameters [9]. MAP estimation maximizes:

$$\mathrm{P}(\vartheta|X) \propto \mathrm{p}(X|\vartheta)p(\theta)^{\mathrm{r}}, \tag{3}$$

where p(ϑ) is the prior distribution, and r is a weighting factor. In contrast, [10] involves learning transformations specific to a speaker during training and testing phases. Cluster adaptive training, which groups speakers by acoustic characteristics to adapt models, was introduced in [11].

3. Recent Advances of Speaker Adaptation Techniques using DNN for Speech Enhancement

Traditional speaker adaptation techniques, outlined in the previous section, mainly focused on model parameter adjustments and transformations of models whose parameters were interpretable. However, these methods cannot be directly extended to DNN architectures, as DNN lack parameter interpretability. Thus, DNN offer new opportunities and challenges to address speaker adaptation more robustly.

Speaker adaptation in DNN refers to modifying a pre-trained DNN to enhance performance for a new speaker without the need for extensive retraining from scratch. Studies have shown that this retraining approach outperforms training DNNs from scratch for speaker adaptation tasks [12]. Various speaker adaptation techniques for DNN include fine-tuning, learning hidden unit contributions, speaker adaptive training, I-vector adaptation and embedding layer adaptation. In this section, we will explore the various speaker adaptation strategies using DNNs for speech enhancement and they are given below.

3.1. Speaker and Noise Embedding

This technique involves embedding distinctive features of a speaker's clean speech vectors and background noise vectors into the neural network. In [13], the authors develop a novel training framework that performs joint optimization of the speech enhancement model and the speaker embedding extraction process. Generally, traditional approaches treat speaker embedding and speech enhancement modules as separate processes, leading to suboptimal speech enhancement performances. This paper introduces a joint learning mechanism where the speaker verification is pretrained and further adapted through a combined enhancement loss and speaker verification loss. This dynamically adapts to the speaker's identity and acoustic environment, offering personalized speech enhancement performance. In [14], another novel Speaker-Aware Speech Enhancement (SASE) method extracts speaker information from a clean reference using LSTM layers and then employs a Convolutional Recurrent Neural Network (CRN) to embed the extracted speaker information. Even in this method, a joint learning framework is utilized, which optimizes speaker extraction through LSTM and speech enhancement through CRN simultaneously. With the extended self-attention mechanism, the proposed framework performs adaptation using just a few seconds of clean reference speech.

While previous works have used speaker information to guide speech enhancement, the authors in [15] utilize additional noise conditions to guide speech enhancement for various types of noise environments. The work incorporates Noise-Aware Training (NAT), which employs voice activity detection to differentiate between speech and non-speech frames and uses these non-speech frames as additional noise information. These non-speech frames represent the characteristics of the background noise, which are termed as Dynamic Noise Embedding (DNE). The extracted DNE is concatenated with features of the input speech signal and fed into the speech enhancement module to obtain the final enhanced speech.

The work in [16] investigated the utilization of speaker embeddings, like d-vectors, to selectively enhance the speech of target speakers while suppressing the background noise and other interfering speakers. The work proposes two new models: Personalized Deep Complex Convolution Recurrent Neural Network (DCCRN) that integrates the speaker d-vector as an additional input to personalize the enhancement, and Personalized Deep Convolution Attention U-Net (PDCAttUNet) that incorporates a self-attention mechanism to capture temporal dependencies and fine-tune the enhancement based on the target speaker's characteristics. The attention mechanisms allow the network to focus more precisely on the relevant aspects of the target speaker. The work introduces a new metric for assessing the Target Speaker OverSuppression (TSOS), thereby ensuring that enhancement does not negatively impact the intelligibility of the target speaker. Building on the speaker extraction strategy, the authors in [17] employ a real-time, low-computational PercepNet framework and adapt the PercepNet by utilizing a speaker embedding. The uniqueness of this paper is that, despite the integration of advanced features and personalization, the personalized PercepNet maintains low complexity, making it deployable to edge devices.

A similar real-time DNN based personalized speech enhancement technique is proposed in [18]. Tencent Ethereal Audio Personalized Speech Enhancement (TEA-PSE) incorporates a two stage speech enhancement framework combined with the state-of-the-art speaker verification architecture, ECAPA-TDNN speaker encoder, to achieve personalized speech enhancement. The Enhanced Context-Aware Predictive Attention Delay Neural Network (ECAPA-TDNN) is used as an encoder to generate the speaker embedding from a target speaker's enrollment, and these extracted embeddings are then used to condition the two-stage speech enhancement framework. This conditioning fine-tunes network weights to match the target speaker's characteristics. The Tencent-Ethereal-Audio-Lab Personalized Speech Enhancement (TEA-PSE) system ranked 1st in the ICASSP 2022 Deep Noise Suppression (DNS2022) challenge. The refined version of TEA-PSE, introduced in [19], TEA-PSE 2.0, introduces a subband processing approach that reduces computational complexity, complemented with improved speech enhancement performance. TEA-PSE 2.0 brings a 0.102 OVRL personalized DNSMOS improvement with only 21.9% multiply-accumulate operations compared with the previous TEA-PSE. An extended version of TEA-PSE 2.0, introduced in [20] as TEA-PSE 3.0, integrates key novel approaches with TEA-PSE 2.0. the novel approaches of TEA-PSE 3.0 are:

- Subband Processing: TEA-PSE 3.0 employs subband processing using Squeezed Temporal Convolution Networks (TCN).
- ResidualSTM: A residual layer is added after each S-TNC layer to improve sequence modeling capabilities.
- Local-Global Representation (LGR): It combines local and global speech features to extract speaker-specific information.

TEA-PSE 3.0 ranked 1st in both ICASSP 2023 Deep Noise Suppression Challenge DNS-Challenge Track 1 and Track 2. Authors in [21] propose a novel multistage, multi-loss training framework that performs both full-band personalized and non-personalized speech enhancement. The framework utilizes a speaker extraction network that employs speaker embeddings to adapt the DNN process for a particular speaker. The speaker extraction network consists of a 1-D convolutional layer processing the time-domain input signal. The network uses several GRU layers that are effective in modeling time-series data, capturing the temporal dynamics and dependencies within the speech signal. It also uses multiple feedforward networks and classification layers to fine-tune the network to differentiate between different speakers. Once the speaker embedding is generated, it is used by the speaker extraction network to filter out non-target speech and background noise and to enhance the target speaker.

3.2. Meta-Learning

Often referred to as "learning to learn", meta-learning enables a model to quickly adapt to new tasks or speakers with minimal training data of the target speaker. A novel speaker adaptation approach termed One Shot Speaker Adaptive Meta-Learning (OSSEM) is introduced in [22]. OSSEM combines a modified transformer with a Speaker-Specific Masking (SSM) to achieve real-time personalized speech enhancement. OSSEM has mainly two components:

- Modified Transformer: It is responsible for the speech enhancement task, which uses a convolutional encoder in the transformer instead of the traditional positional encoder [23].
- Speaker-Specific Masking: This network employs speaker embeddings to create masks that adjust the input features to the speech enhancement model, making the enhancement customized to the target speaker's spectral characteristics.

OSSEM uses a meta-learning framework for rapid adaptation to new speakers. MetaLearning Framework: This is achieved by structuring the training process around support and query sets:

- Support Set: Used to adapt the model to the target speaker's characteristics.
- Query Set: Used to evaluate the model's performance post-speaker adaptation.

In the real time application phase, the OSSEM model adapts to a new speaker using a single utterance. OSSEM has shown competitive performance with other state-of-the-art real-time causal speech enhancement techniques. Another work in [24] utilized a meta-learning framework for speaker adaptation. In this work, a U-Net architecture is used as the meta-learner. Results from the work show that speech enhancement through meta-learning outperforms traditional speech enhancement techniques in few-shot learning scenarios where only a small amount of training data is available for adaptation.

3.3. Multi-Task Learning

The multi-task learning approach trains the neural network on multiple related tasks simultaneously, such as speech enhancement and speaker identification. By learning these related tasks together, the network can improve its generalization capabilities and performance on individual tasks. In [3], a framework for unified real-time personalized and non-personalized speech enhancement is introduced that integrates both personalized and non-personalized speech enhancement within a single DNN architecture. This framework consists of two components:

- Speaker Embedding Network: This network extracts the target speaker's features to facilitate personalized enhancement [25]. ECAPA-TDNN is used as the speaker embedder as it achieves state-of-the-art results in several speaker recognition tasks.
- Enhancement Network: This network performs the actual speech enhancement. It is conditioned on the output of the speaker embedding network and a frame-wise control input that directs whether the enhancement should be personalized or non-personalized. This is accomplished using a binary input that dynamically switches between personalized and non-personalized models across the frames. This multitasking unified framework provides better speech enhancement results when compared to state-of-the-art architectures.

3.4. Zero-Shot Learning

In [26], authors introduce a speech enhancement method based on a denoising autoencoder with multi-branched encoders (DAEME), that combines the robustness of deep learning with the flexibility of ensemble methods. The DAEME model has the following components:

- Multi-branched Encoder: Each branch is trained on a subset of the data characterized by certain noise types and speaker characteristics. This allows the encoder to handle different noise types and speaker characteristics effectively.
- Dynamically Sized Decision Tree (DSDT) Framework: The decision tree aids in partitioning the training data effectively.
- Decoder: After processing by the encoder branches, the outputs are fused by a decoder that consolidates the branched outputs into a single enhanced speech signal.

The DAEME model presents a novel speaker adaptation strategy where it adapts the DNN to new speakers and varying noise conditions without requiring retraining. In the work described in [27], authors perform personalized speech enhancement using zero-shot learning with a knowledge distillation framework. This approach allows for speaker adaptation without requiring speaker-specific clean speech data. In the knowledge distillation approach, a large, well-trained teacher model imparts knowledge to a smaller student model. The well-trained teacher model processes noisy speech to obtain enhanced speech outputs, which serve as pseudo-targets for training the compact student model in real-time during deployment. The authors in [28] introduce an ensemble of specialist modules through zero-shot learning to perform personalized speech enhancement. In this proposed method, an ensemble of specialist models is employed, where each specialist model is trained to handle a specific subset of training data characterized by speaker characteristics or noise types. A gating module estimates the relevance of each specialist for a given input during inference and selects the most appropriate specialist model to process the input. This approach adapts to new speaker and noise conditions without requiring retraining of the entire model. A Siamese network generates discriminative, speaker-specific embeddings that maximize or minimize the similarity of the output vectors depending on whether the input utterances are from the same speaker. These generated embeddings are used to train the set of speakers into different clusters through k-means clustering. The gating module is trained to predict to which cluster a new input belongs based on the speaker's characteristics.

After individual training, the entire ensemble method is fine-tuned to optimize coordination between the gating module and the specialist modules. This framework reduces the computational complexity compared to the baseline while maintaining the same speech enhancement performance.

A similar work was proposed in [29], presenting a novel approach to speech enhancement (SE) using zero-shot learning combined with a Quality-Net based model selection strategy.

A central element of the proposed system is Quality-Net, a deep learning based non-intrusive quality assessment model trained to predict speech quality scores. Quality-Net is utilized in two key capacities:

- Model Training: During the offline phase, it assesses the quality of noisy speech samples and assists in clustering the training data to effectively train specialized SE models.

- Model Selection: In the online phase, it evaluates incoming noisy speech to select the optimal SE model for real time enhancement.

Two strategies are employed for zero shot model selection:

- ZMOS-QS (Quality Score based): Utilizes the quality scores predicted by the Quality-Net to cluster the training data and select the appropriate model at test time.
- ZMOS-QE (Quality Embedding based): Uses embeddings from the Quality-Net for clustering and model selection, focusing on latent representations that capture more nuanced aspects of speech quality.selection, focusing on latent representations that capture more nuanced aspects of speech quality.

The proposed ZMOS approach, incorporating both the QS and QE strategies, provides superior speech enhancement performance compared to traditional and baseline systems.

In [30], authors present a self-supervised learning method to adapt speech enhancement models. The proposed method extracts speech features from unlabeled, in-the-wild noisy recordings of the target use without access to corresponding clean speech data. This addresses the privacy concerns and practical challenges associated with collecting clean speech data. The paper utilizes contrastive learning techniques that aid in refining the model to distinguish between different sounds within the noisy data. A data purification technique introduced in the paper identifies and prioritizes the cleaner parts of the noisy data during training. A novel personalization technique is proposed in [31] that uses Neural Speech Synthesis (NSS) for data augmentation. This study explores how synthetic data generated by advanced Text-To-Speech (TTS) systems can be utilized to train personalized speech enhancement when direct recordings of the target speaker are limited. In this work, two types of NSS systems are evaluated for the generation of target synthetic data:

- YourTTS: a multi-speaker, multilingual TTS system.
- AudioLM: an autoregressive speech synthesis system that does not rely on textual input.

Models are trained using a combination of real speech recordings and synthetic speech generated by YourTTS and AudioLM. The research findings suggest that better personalization of speech enhancement models is possible with the generation of high-quality synthetic data.

4. Challenges

DNNs lack direct interpretation of their parameters, making speaker adaptation intricate and complex. The multilayered architecture coupled with a complex framework of parameters, makes it difficult to understand how speaker specific changes in the network or input data affect the output.

DNNs are fixed at training time [32] and cannot undergo any structural alterations during actual deployment. Acoustical conditions change in real-time applications, which complicates the adaptation of DNNs. In [33], the authors introduced a novel approach for speaker adaptation using codebook integrated deep neural networks for speech enhancement. This method is advantageous as it can be applied to any DNN architecture and facilitates the adaptation of DNNs to multiple speakers without necessitating any modifications to the network itself.

Although the data augmentation technique decreases computational complexity by training the DNN with the target speaker's synthetic data, offering some advantages, it often poses a limitation. Transitioning to another speaker requires retraining the entire DNN with the other speaker's synthetic data, presenting a challenge of excessive retraining. Speaker adaptation techniques, by introducing a speaker extraction/auxiliary network to generate speaker embeddings, add to the computational complexity and processing time of the DNN. This can be a concern for their deployment on devices like hearing aids and mobile phones.

5. Conclusions

In this work, we have presented an overview of the current methodologies and techniques in speaker adaptation using DNN for speech enhancement. The study highlights the strengths of various adaptation techniques, including the use of speaker embeddings, zero-shot adaptation, few-shot learning, meta-learning approaches, and data augmentation. These methods have proven effective in adapting DNN to individual speaker characteristics. However, there are challenges related to DNN retraining when transitioning between different speakers and increased computational complexity. Despite the advancements in speaker adaptation, these challenges highlight the need for more efficient personalized models that provide better speech enhancement performance. These models should adapt to different speakers without the need to retrain the DNN for each transition and manage computational demands effectively.

Supplementary Materials: None.

Author Contributions: The entire article was conceived and authored by Dr. Chidambar, with guidance from Dr. Hanumanth Rao Naidu.

Funding: None.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I am deeply grateful to Sadguru Sri Madhusudan Sai for guiding me to pursue work in the field of speech enhancement and speech processing.

Conflicts of Interest: None.

References

[1] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," IEEE Transactions on Speech Audio Processing, vol. 15, no. 2, pp. 441–452, 2007.

[2] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," IEEE Open Journal of Signal Processing, vol. 2, pp. 33–66, 2021, doi:10.1109/OJSP.2020.3045349

[3] Z. Wang, R. Giri, D. Shah, J.-M. Valin, M. M. Goodwin, and P. Smaragdis, "A framework for unified real-time personalized and non-personalized speech enhancement," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5, doi:10.1109/ICASSP49357.2023.10097247

[4] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speaker adaptation for enhancement of bone-conducted speech," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 10456–10460, doi:10.1109/ICASSP48485.2024.10447322

[5] S. Furui, "Vector-quantization-based speech recognition and speaker recognition techniques," in Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems Computers, vol. 2, 1991, pp. 954–958.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Computer Speech & Language, vol. 9, no. 2, pp. 171–185, 1995, doi:10.1006/csla.1995.0010

[7] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," Computer Speech & Language, vol. 12, no. 2, pp. 75–98, 1998, doi10.1006/csla.1998.0043

[8] L. Neumeyer, A. Sankar, and V. Digalakis, "A comparative study of speaker adaptation techniques," in Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 1995), 1995, pp. 1127–1130, doi:10.21437/Eurospeech.1995-282

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291–298, 1994, doi:10.1109/89.279278

[10]T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in Proceedings of the 4th International Conference on Spoken Language Processing, 1996, pp. 3–35, doi:10.1109/ICSLP.1996.607807

[11]M. J. F. Gales, "Cluster adaptive training of hidden markov models," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 4, pp. 417–428, 2000, doi:10.1109/89.848223

[12] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4535–4539, 2015, doi:10.1109/ICASSP.2015.7178829

[13] S. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "A comparative study of time and frequency domain approaches to deep learning based speech enhancement," 07 2020, pp. 1–8, doi:10.1109/IJCNN48605.2020.9206928

[14] J. Lin, A. J. van Wijngaarden, M. C. Smith, and K.-C. Wang, "Speaker-aware speech enhancement with self-attention," 2021 29th European Signal Processing Conference (EUSIPCO), pp. 486–490, 2021, doi:10.23919/EUSIPCO54536.2021. 9616282

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2670–2674, 01 2014, doi:10.21437/Interspeech.2014-571

[16] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in ICASSP 2022. IEEE, May 2022, doi:10.1109/ICASSP43922.2022.9746962

[17] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, "Personalized percepnet: Real-time, low-complexity target voice separation and enhancement," 08 2021, pp. 1124–1128, doi:10.21437/Interspeech.2021-694

[18] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, Y. Wang, L. Xie, and S. Shang, "Tea-pse: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2022 dns challenge," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 9291–9295, doi:10.1109/ICASSP43922.2022.9747765

[19] Y. Ju, S. Zhang, W. Rao, Y. Wang, T. Yu, L. Xie, and S. Shang, "Tea-pse 2.0: Sub-band network for real-time personalized speech enhancement," in 2022 IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 472–479, doi:10.1109/SLT54892.2023.10023174

[20] Y. Ju, J. Chen, S. Zhang, S. He, W. Rao, W. Zhu, Y. Wang, T. Yu, and S. Shang, "Tea-pse 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2023 dns-challenge," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–2, doi:10.1109/ICASSP49357.2023.10096838

[21] L. Chen, C. Xu, X. Zhang, X. Ren, X. Zheng, C. Zhang, L. Guo, and B. Yu, "Multi-stage and multi-loss training for fullband non-personalized and personalized speech enhancement," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 9296–9300, doi:10.1109/ICASSP43922.2022.9746981

[22] C. Yu, S.-W. Fu, T.-A. Hsieh, Y. Tsao, and M. Ravanelli, "Ossem: one-shot speaker adaptive speech enhancement using meta learning," 2021, doi:10.21437/Interspeech.2022-10283

[23] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, K.-H. Hung, S.-S. Wang, C. Yu, H.-C. Kuo, R. E. Zezario, Y.-J. Li, and S.-Y. Chuang, "Boosting objective scores of a speech enhancement model by metricgan post-processing," in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020.

[24] W. Zhou, M. Lu, and R. Ji, "Meta-se: A meta-learning framework for few-shot speech enhancement," IEEE Access, vol. 9, pp. 46068–46078, 2021, doi:10.1109/ACCESS.2021.3066609

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," 10 2020, doi:10.21437/Interspeech.2020-2650

[26] C. Yu, R. E. Zezario, S.-S. Wang, J. Sherman, Y.-M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2756–2769, 2020.

[27] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021, doi:10.1109/WASPAA52581.2021.9632771

[28] A. Sivaraman and M. Kim, "Zero-shot personalized speech enhancement through speaker-informed model selection," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021, pp. 171–175, doi:10.1109/WASPAA52581.2021.9632752

[29] R. E. Zezario, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Speech enhancement with zero-shot model selection," in 2021 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 491–495, doi:10.23919/EUSIPCO54536.2021.9616163

[30] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1342–1356, 2022, doi:10.1109/JSTSP.2022.3181782

[31] A. Kuznetsova, A. Sivaraman, and M. Kim, "The potential of neural speech synthesis-based data augmentation for personalized speech enhancement," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5, doi:10.1109/ICASSP49357.2023.10096601

[32] C. Zheng, Y. Zhou, X. Peng, Y. Zhang, and Y. Lu, "Real-time speech enhancement with dynamic attention span," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2023, pp. 1–5, doi:10.1109/ICASSP49357.2023.10095821

[33] B. Chidambar and D. Naidu, "Speaker adaptation using codebook integrated deep neural networks for speech enhancement," JASA Express Letters, vol. 4, no. 11, p. 115203, 2024.