An Approach for Speech Enhancement in Low SNR Environments using Granular Speaker Embedding

Jayasree Saha
IIIT-Hyderabad
India
jayashree.saha@research.iiit.ac.in

Rudrabha Mukhopadhyay
IIIT-Hyderabad
India
radrabha.m@research.iiit.ac.in

Aparna Agrawal
IIIT-Hyderabad
India
aparna.agrawal@research.iiit.ac.in

Surabhi Jain
IIIT-Hyderabad
India
surabhi.jain@research.iiit.ac.in

C. V. Jawahar
IIIT-Hyderabad
India
jawahar@iiit.ac.in

ABSTRACT

The proliferation of speech technology applications has led to an unprecedented demand for effective speech enhancement techniques, particularly in low Signal-to-Noise Ratio (SNR) conditions. This research presents a novel approach to speech enhancement, specifically designed for very low SNR scenarios. Our technique focuses on speaker embedding at a granular level and highlights its consistent impact on enhancing speech quality and improving Automatic Speech Recognition (ASR) performance, a significant downstream task. Experimental findings demonstrate competitive speech quality and substantial enhancements in ASR accuracy compared to alternative methods in low SNR situations. The proposed technique offers promising advancements in addressing the challenges posed by low SNR conditions in speech technology applications.

CCS CONCEPTS

• **Computating methodologies** → **Learning paradigm**; *Conformer*; multi-task learning; • **Speech** → Speech enhancement.

KEYWORDS

Speech enhancement, Conformer, Granular speaker embedding

ACM Reference Format:
Jayasree Saha, Rudrabha Mukhopadhyay, Aparna Agrawal, Surabhi Jain, and C. V. Jawahar. 2024. An Approach for Speech Enhancement in Low SNR Environments using Granular Speaker Embedding . In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 04–07, 2024, Bangalore, India.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3632410.3632413

CODS-COMAD 2024, January 04–07, 2024, Bangalore, India

ACM ISBN 979-8-4007-1634-8/24/01…$15.00
https://doi.org/10.1145/3632410.3632413

1 INTRODUCTION

Speech enhancement technique plays a vital role in enhancing speech quality and intelligibility across various domains where environmental noise challenges effective communication. In telecommunication systems like mobile phones and Voice over IP (VoIP) services, background noise can significantly degrade the quality of speech during conversations. The growing popularity of voice-controlled devices and virtual assistants, such as smart speakers and voice-activated systems, has also highlighted the need for accurate recognition and interpretation of user commands in noisy environments. Additionally, individuals with hearing impairments face significant difficulties in understanding speech when surrounded by noise. Communication within vehicles presents another challenge, as road and engine noise and in-car distractions can hamper effective communication. By implementing speech enhancement algorithms, the clarity of speech can be improved in in-car communication systems, including hands-free calling and voice commands, leading to a safer and more pleasant driving experience. Lastly, with the increasing reliance on remote work and virtual meetings, audio and video conferencing platforms have become essential for effective communication. However, background noise often hinders understanding speech during these virtual interactions. The challenges mentioned above become even more pronounced when the Signal-to-Noise Ratio (SNR) is low. In low SNR conditions, where the speech signal is significantly weaker than the background noise, speech quality, and intelligibility degradation becomes more severe. Therefore, developing and implementing robust speech enhancement techniques are crucial in mitigating the detrimental effects of low SNR scenarios, enabling clearer and more intelligible speech communication in challenging acoustic environments.

In recent times, the field of speech enhancement has witnessed the emergence of deep learning-based techniques, including recurrent neural networks (RNNs)[3, 16, 26], Variational Auto-Encoders (VAEs)[12, 13], and Generative Adversarial Networks (GANs)[22, 23]. Audio-specific transformers [8, 14] represent a relatively new approach that has only been explored in limited contexts for speech enhancement tasks [4, 18, 27]. Nevertheless, they offer powerful solutions that leverage large-scale training datasets and high-performance computing to learn complex mappings between noisy and clean speech signals, leading to substantial improvements in speech quality and intelligibility. Despite these advancements, challenges persist in developing effective techniques to handle very

low SNR conditions. One of the critical issues is the degradation of downstream tasks, such as Automatic Speech Recognition (ASR), due to noise and other distortions in the speech signal. When deploying a task-specific speech enhancement (SE) model, the most straightforward approach is to use a loss function that is directly relevant to the intended outcome. While a naive approach would be to use a measure based solely on the difference in signal level, such as the *L1* or *L2* loss, this may not fully capture the nuances of the desired output and may result in suboptimal performance. Studies have shown that it may not fully align with human auditory perception, intelligibility scores, or ASR accuracy. Despite this, integrating various components to make speech enhancement more realistic remains a challenging problem. Prior research has demonstrated that an SE system can benefit from incorporating additional information beyond just the audio signal. For example, incorporating face/lip images [11] and symbolic sequences for acoustic signals [2] have been explored to improve SE models' performance.

In this work, we investigate a specialized application of an existing speaker embedding network. We aim to harness this network's capabilities to develop speech enhancement models that effectively mitigate noise interference while preserving crucial phoneme characteristics. Through rigorous analysis of the audio data, we have made an intriguing observation: some small granular audio segments, with a duration of approximately 250 milliseconds, may be unaffected by noise and maintain their pristine quality. This finding serves as the basis for our approach, suggesting that the model can learn valuable features from clean audio segments to enhance the denoising process. In summary, our key contributions are:

(1) We present a novel application of a speaker embedding network specifically designed for speech enhancement tasks. Our research showcases the unique and valuable potential of utilizing this network to improve the quality and intelligibility of speech signals.
(2) We introduce a novel conformer-based architecture for speech enhancement, which leverages a multi-task learning framework to learn denoising masks.
(3) To further improve the quality of our denoised audio outputs, we fine-tune the BigGAN vocoder [7]. By leveraging the capabilities of this advanced vocoder, renowned for its ability to generate smooth and high-quality audio, we strive to enhance the overall audio quality and intangibility of the denoised audio signals produced by our model.

2 PROPOSED METHOD

2.1 Model architecture

In this paper, a conformer-based speech enhancement network architecture is proposed and presented in Figure 1. The architecture includes various components such as feature Extraction modules for audio, a speaker verification model, two encoders (Spk-enc and Spec-enc), a decoder, and two output blocks. The encoder and decoder stages comprise N conformer blocks, which outperform previous transformer and convolution neural network (CNN) based architectures by achieving state-of-the-art accuracy. The conformer architecture is parameter-efficient and can learn an audio sequence's local and global dependencies. Spk-enc and Spec-enc encode speaker embedding and linear spectrogram, respectively. To ensure consistency in output dimensions between Spk-enc and Spec-enc, an upsampler is added to the spec-enc. The encoded features from both encoders are concatenated and passed to the decoder, which also comprises $2 \times N$ conformer blocks. After the decoder block, two output blocks are appended to obtain both the linear spectrogram and mel spectrogram. Each output block follows a sequential structure consisting of layer normalization, a feed-forward layer, Swish activation function, dropout, and another feed-forward layer. This design ensures that both output blocks receive consistent treatment and that the resulting spectrograms are of high quality. Our proposed methodology consists of a series of steps for enhancing noisy audio data. First, we utilize the "librosa" Python package to extract mel-spectrogram and linear spectrogram representations from the input audio signal. To simulate prosody-like features, we slice the mel-spectrogram and pass each slice through a speaker verification model. Following this, we feed the linear spectrogram and prosody style feature into the spec-enc and spk-enc modules, respectively. To allow concatenation, we upsample the spk-enc output as its time dimensions differ from that of the spec-enc output. We then pass the concatenated features to the decoder block for further processing. In the final step, we project the decoded feature onto separate output blocks to predict both the linear and mel spectrogram. During inference, we use the Griffin-lim [9] algorithm to generate the speech signal from the linear spectrogram.

2.2 Speaker's characteristic extraction

Our research aims to explore the potential benefits of preserving speaker characteristics at the partial utterance level in enhancing speech quality. To achieve this goal, we propose an approach that decouples speaker modeling from speech enhancement by training an independent speaker-discriminative embedding network. The most promising representation of a speaker is one that can distinguish itself from other speakers. We use a speaker verification model to produce such characteristics using only a short adaptation signal, independent of phonetic content and background noise. We adopt the highly scalable and effective neural network framework proposed in [17] for speaker verification. This framework involves a network that maps a sequence of log-mel spectrogram frames generated from a speech utterance of any length to a fixed-dimension embedding vector. It is widely understood that noise affects different sections of a speech signal unequally due to variations in the time domain. As a result, micro-speech segments provide a better opportunity to learn the speaker's characteristics, such as rhythm, stress, and pitch intonation, along the time axis. In this paper, we are exploring the possibility of indirectly enhancing phonetic content by restoring the speaker embedding at a micro-level. Our hypothesis is that the speaker's characteristics are highly correlated with phonetic content in the broader context. Therefore, there is a high probability that restoring such information while attempting to restore the speaker's characteristics in the micro-segments will result in improved phonetic content. In order to capture the speaker's style, which may vary over time, we utilize a speaker verification model to process mel-spectrograms of one second of speech in a sliced manner. Based on the average speaking rate for American English speakers of approximately 250 syllables per minute [21], we

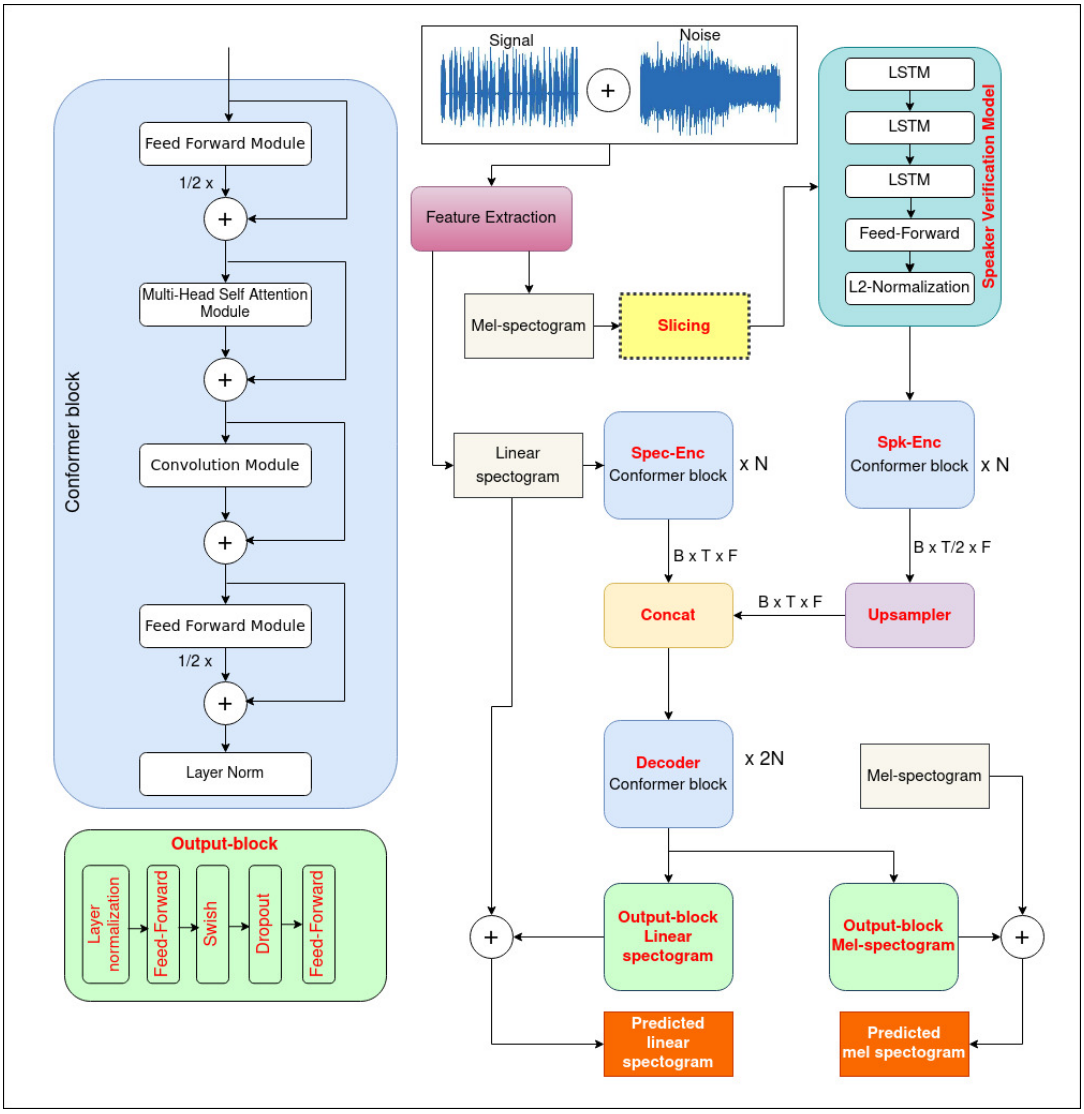


Figure 1: Diagram of the Proposed Speech Enhancement Model: The model's primary objective is to learn a mask, and as part of the process, noisy linear and mel spectrograms are combined with the mask to produce denoised spectrograms

divide the mel-spectrograms into 250ms windows with 210ms stride. After passing the slices through the speaker verification model, we concatenate the resulting slices along the time dimension, assuming that this feature can effectively capture the speaker's style over time. This approach enables us to better understand the correlation between speaker characteristics and phonetic content and could lead to an improved speech-enhancement algorithm.

2.3 Loss function

Deep neural network models for speech enhancement (SE) tasks often use the L1 loss as a loss function to learn a mask (additive or multiplicative) for noisy speech. Recent studies have highlighted a growing interest in SE methods that utilize multi-task learning [1, 15], reflecting the potential benefits of leveraging additional information sources in enhancing speech quality. Koizumi et al. [15] have incorporated speaker identification into a multi-task-learning-based loss function for speech enhancement, as they believe that time-dependent speaker information could better represent dynamic phoneme information. Building on this idea, we take the next step in utilizing speaker embedding as a valuable cue for training the main SE framework. Specifically, we propose including speaker embedding as an important cue for the SE task by training a separate speaker verification network to learn speaker embedding from noisy speech. We then integrate this auxiliary network into the main framework as part of our multi-tasking strategy, ensuring that the output of our SE model can reconstruct speaker embedding similar to the auxiliary network. This approach reduces the chances of denoising steps interfering with the phoneme quality of target

Table 1: Results of different SEs in a low SNR condition.

LRS3 test dataset	−5dB				−10dB			
Metric / Methods	PESQ ↑	STOI ↑	CER ↓	WER↓	PESQ↑	STOI↑	CER↓	WER ↓
Noisy	1.145	0.536	1.423	1.391	1.094	0.396	2.522	2.098
DNS 64	1.462	0.776	1.076	1.173	1.289	0.612	1.836	1.789
SEGAN	1.121	0.525	2.122	1.998	1.089	0.365	2.888	2.362
Baseline-1	1.171	0.611	1.512	1.514	0.107	0.486	1.974	1.869
Baseline-2	1.122	0.514	1.694	1.577	1.090	0.395	2.525	2.194
Our Model	1.252	0.623	**0.550**	**0.743**	1.163	0.532	**0.781**	**1.006**
Our model + BigVGAN	1.256	0.642	0.584	0.790	1.157	0.544	0.833	1.056

Table 2: Ablation study on the order of λ_1, λ_2, and λ_3

LRS3 test dataset	-5dB				-10dB			
Metric / Order	PESQ↑	STOI↑	CER↓	WER↓	PESQ↑	STOI↑	CER↓	WER↓
$\lambda_1 > \lambda_2 > \lambda_3$	1.252	0.623	0.550	0.743	1.163	0.532	0.781	1.006
$\lambda_1 > \lambda_3 > \lambda_2$	1.107	0.518	0.550	0.716	1.098	0.421	1.423	1.410
$\lambda_2 > \lambda_3 > \lambda_1$	1.102	0.514	0.592	0.755	1.092	0.416	1.458	1.443
$\lambda_2 > \lambda_1 > \lambda_3$	1.110	0.519	0.579	0.733	1.098	0.518	1.340	1.354
$\lambda_3 > \lambda_1 > \lambda_2$	1.098	0.513	0.601	0.752	1.103	0.421	1.259	1.251
$\lambda_3 > \lambda_2 > \lambda_1$	1.100	0.520	0.546	0.715	1.104	0.424	1.396	1.388

speech. This speaker verification model takes a mel-spectrogram as input to produce speaker embedding. Since the speaker verification model takes a mel-spectrogram as input to produce speaker embedding, joint prediction of both features within the SE network may facilitate co-learning and lead to faster convergence. To incorporate speaker embedding as a potential member of the multi-objective loss function, we consider reconstructing three potential features and use the following equation as the loss function

$$\mathcal{L} = \lambda_1 \mathcal{L}_{spec} + \lambda_2 \mathcal{L}_{mel} + \lambda_3 \mathcal{L}_{spk} \tag{1}$$

where, $\mathcal{L}_{spec} = \sum_t ||x^t_{spec} - \hat{x}^t_{spec}||_1$, $\mathcal{L}_{mel} = \sum_t ||x^t_{mel} - \hat{x}^t_{mel}||_1$, and $\mathcal{L}_{spk} = \sum_t ||x^t_{spk} - \hat{x}^t_{spk}||_1$. $\hat{x}$ and x are the corresponding features of the enhanced and clean speech signal, respectively, and *spec*, *mel*, *spk* determines the linear spectrogram, mel-spectrogram, and speaker embedding, respectively. Also, $\lambda_1, \lambda_2, \lambda_3$ are the scaling parameters that control the effect of the subtask. Moreover, we keep $\lambda_1 > \lambda_2 > \lambda_3$ for our task.

2.4 Finetune BigVGAN

Lee et al. [7] recently proposed BigVGAN, a universal vocoder capable of effectively handling various out-of-distribution scenarios without requiring fine-tuning. In their work, they introduced a periodic activation function and anti-aliased representation into the GAN generator. These additions provide the desired inductive bias for audio synthesis and result in a significant enhancement in audio quality. The researchers trained the model on clean speech data from LibriTTS and achieved state-of-the-art performance in zero-shot conditions such as unseen speakers, languages, recording environments, singing voices, music, and instrumental audio. In light of these promising results, we aim to further enhance the audio quality by fine-tuning the BigVGAN model using denoised audio generated by our method.

3 DATASET & EVALUATION METRICS

We conducted experiments using the "pre-train" sets of the LRS3 dataset [24], a publicly available collection of spoken sentences from TED videos. It consists of approximately 400 hours of video data and 118,516 utterances, for training purposes. To generate the noisy data for our experiments, we used the VGG-Sound [10] dataset. This dataset consists of 500 hours of diverse audio data, spanning 310 distinct classes of challenging acoustic environments and noise characteristics encountered in real-life applications. We excluded audio samples from the VGG-Sound dataset in which people were speaking. During the training process, we introduced a random signal-to-noise ratio (SNR) ranging from -10dB to 5dB to make the training data more challenging. We also resampled all speech waveforms to a sampling rate of 16 kHz and transformed the signal to linear and Mel-spectrograms using the Hann window function with a frame length of 400 and hop length of 160, followed by a 512-bin fast Fourier Transform (FFT).

To assess the performance of our method on a diverse range of speakers and speech patterns, we selected the LRS3 test dataset, which includes a total of 412 speaker utterances. We utilize two standard metrics (higher is better) for speech quality and intelligibility: wideband Perceptual Evaluation of Speech Quality (PESQ) [20] (-0.5 to 4.5) and Short-Time Objective Intelligibility (STOI) [25] (0 to 1). We have also shown the word error rate (WER) and character error rate (CER) (lower is better) for evaluating the performance of ASR on denoised data.

4 EXPERIMENTS

We followed the procedure outlined in the original paper [17] to train our speaker verification network. For trainning SE model, we utilized the Adam optimizer with a learning rate of 1×10^{-3} for every training step. We conducted 100,000 training steps for our model, using a batch size of 64 and 8 workers. We employed 2 Nvidia GeForce GPUs to optimize training efficiency, which allowed us to process large amounts of data in parallel and achieve faster training times. We enhanced noisy speech in low SNR conditions by utilizing this training approach. Using multiple GPUs and workers further expedited the training process, allowing us to efficiently process large amounts of data and optimize our model for improved performance. The primary focus of our experiment is the sliced speaker embedding strategy. We have compared our method with dns64 [5] and SEGAN [19]. According to the experimental findings, the use of dns64 significantly enhances speech intelligibility in comparison to other techniques. Nevertheless, our approach yielded comparable outcomes while also surpassing the performance of ASR in low SNR conditions.

4.1 Ablation Study

To fully understand the importance of this approach, we implemented several ablation setups in our evaluation. In the first setup (Baseline-1), we removed the speaker embedding and mel-projection block from the architecture to assess their impact on the model's overall performance. In the second setup (Baseline-2), we incorporated CNN blocks in the encoder and decoder architecture, as used by Hedge et al. [11], to evaluate the efficacy of the conformer blocks over CNN blocks. We have conducted a series of experiments to explore the optimal time window size for granular speaker embedding , their significance within the model, the ordering of values for λ_1, λ_2, and λ_3, as well as the overall generality of our model.

Table 3: Ablation study to understand slicing speaker embedding

	-5dB				-10dB			
Metric / words/s	PESQ↑	STOI↑	CER↓	WER↓	PESQ↑	STOI↑	CER↓	WER↓
1	1.106	0.514	0.569	0.747	1.106	0.424	1.290	1.296
3	1.105	0.517	0.617	0.775	1.094	0.423	1.318	1.359
4	1.252	0.623	0.550	0.743	1.163	0.532	0.781	1.006
4 (alt_voice)	1.101	0.520	0.534	0.700	1.101	0.420	1.302	1.329
4 (alt_spk)	1.101	0.513	0.570	0.742	1.093	0.420	1.354	1.294

Table 4: Performance of comparable methods on TIMIT dataset

	-5dB				-10dB			
Metric / Methods	PESQ↑	STOI↑	CER↓	WER↓	PESQ↑	STOI↑	CER↓	WER↓
Our_model	1.291	0.579	0.670	0.978	1.193	0.505	0.941	1.208
SEGAN	1.091	0.413	0.529	0.795	1.108	0.324	4.059	3.052
DNS64	1.461	0.756	0.569	0.856	1.280	0.607	1.155	1.306

Through our ablation experiments, we were able to gain a deeper understanding of the impact of each component on the overall performance of our sliced speaker embedding strategy. These findings are critical in the development of more effective speech enhancement algorithms and provide valuable insights for future research in this field.

4.2 Results Analysis

Based on the findings in Table 1, our model demonstrated comparable results to dns64 [5] in the low SNR condition, the best-performing method with respect to speech enhancement metrics. The most remarkable outcome was the superior performance of our approach compared to other methods in terms of Word Error Rate (WER) under low SNR conditions. We also observed that BigVGAN improves audio quality in terms of STOI. In Figures 2 and 3, we showcase the noisy spectrogram, as well as the spectrograms for the reconstructed speech signal using our methods, DNS 64, and the clean version. The presence of background noise heavily affects the speech, but our method successfully suppresses the noise-dominant region. Moreover, BIgVGAN demonstrates further improvements in the final result.

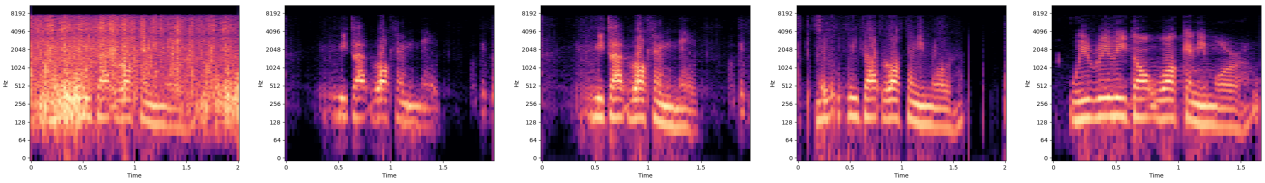
Speaker Embedding To evaluate the efficacy of granular speaker embedding, we conducted a study in which we plotted the embeddings of three speakers using two different methods. In the first method, we collected nine different speech contents from each speaker, with each sample being one second long. We then passed the mel-spectrogram of each sample through the speaker verification network and applied t-SNE to plot the embeddings on a scatter plot. In the second method, we collected slices of the mel-spectrogram for each speaker's content and passed them through the same network before plotting the resulting embeddings. The plots generated by these two methods are depicted in Figure 4. The scatter plot produced by the first method showed distinct clusters for each of the three individual speakers, indicating that the speaker embeddings for the one-second sample can capture speaker-specific features. However, the second method did not produce such distinct clusters, suggesting that it may not be as effective at capturing speaker-specific features across different speeches of a speaker. Nonetheless, we noticed distinct small clusters from a speaker sample that were closely spaced together upon closer inspection. This observation indicates that the granular speaker embedding has the potential to capture various variations in speech. To assess its effectiveness within our model, we conducted an ablation study on the length of slices (varying words/s) and have summarized the results in Table 3, which suggests that a 250 ms time window (4 words/second) is the optimal choice for our algorithm. Additionally, we conducted another experiment involving speaker embeddings of 250 ms time slices from the same speaker but different speech (alt_voice in Table 3), as well as speaker embeddings from a different speaker (alt_spk in Table 3). The results displayed in Table 3 demonstrate that opting for distinct speech data from the same speaker or speech data from a different speaker results in a performance decline. This decline is even more significant than the baseline-1 performance, which does not include the speaker embedding module.

Generalizability: We executed our model on the test datasets of TIMIT [6]. The results in Table 4 indicate a similar level of performance as observed on the LRS3 test dataset.

5 CONCLUSION & FUTURE WORK

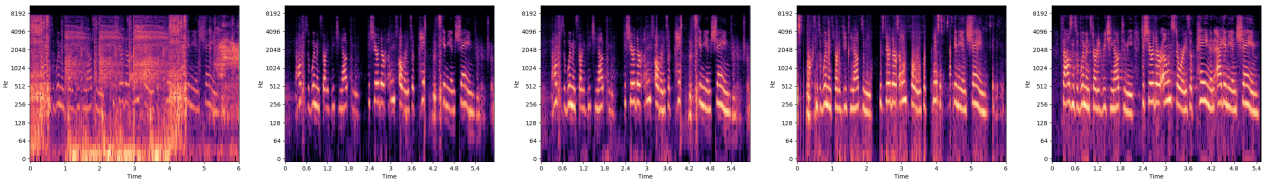
This work introduced a novel speech enhancement technique that utilizes granular speaker embedding within a multi-task learning framework. Our findings highlight the substantial influence of speaker embedding on the design of speech enhancement models. Experimental results demonstrate that our approach enables the Automatic Speech Recognition (ASR) system to accurately recognize words in low Signal-to-Noise Ratio (SNR) conditions, outperforming alternative methods that struggle in such scenarios.

One potential future direction for exploration is incorporating phoneme information to influence ASR outcomes in low SNR conditions directly. Integrating phoneme-level details makes it possible to enhance ASR performance further and address the challenges posed by low SNR environments.



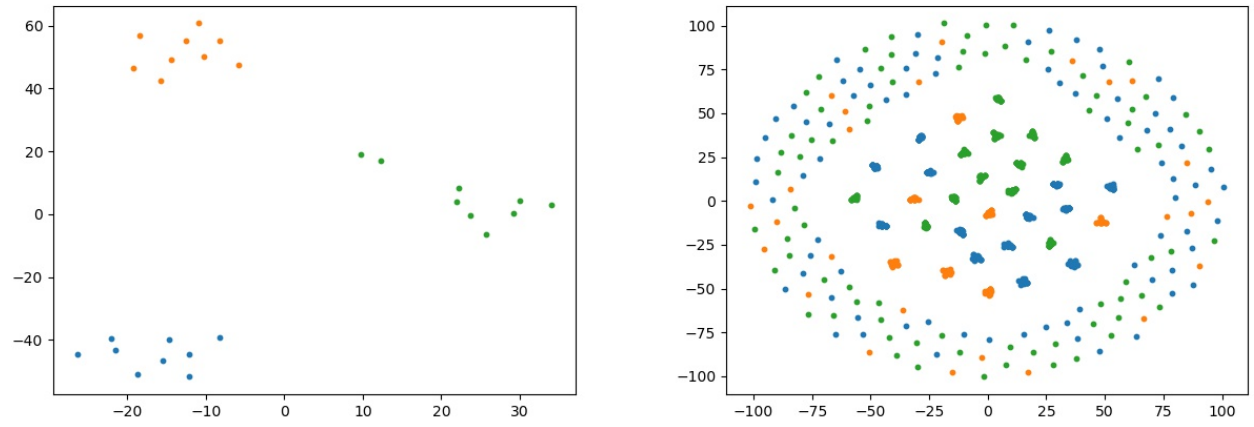
(a) noisy (b) Our model (c) our model + BigVGAN (d) DNS64 (e) clean

Figure 2: Samples generated from LRS3 test set and VGGSound noise and we set SNR=-5dB. It can be seen that BigGAN surely improves our model's outcome further.



(a) noisy (b) Our model (c) our model + BigVGAN (d) DNS64 (e) clean

Figure 3: Samples generated from LRS3 test set and VGGSound noise and we set SNR=-10dB. It can be seen that BigGAN surely improves our model's outcome further.



(a) speaker embedding (b) slices of speaker embedding

Figure 4: t-SNE analysis for three speakers' embedding. Each color represents separate speaker

REFERENCES

[1] Yoshiaki Bando, Kouhei Sekiguchi, and Kazuyoshi Yoshii. 2020. Adaptive Neural Speech Enhancement with a Denoising Variational Autoencoder. In *Proc. Interspeech 2020*. 2437–2441.

[2] X. Lu C.-F. Liao, Y. Tsao and H. Kawai. 2019. Incorporating Symbolic Sequential Modeling for Speech Enhancement. In *Interspeech*.

[3] S. Takaki C. Valentini-Botinhao, X. Wang and J. Yamagishi. 2016. Investigating rnn-based speech enhancement methods for noise robust text-to-speech. In *Proceedings of Speech Synthesis Work- shop (SSW)*.

[4] Timo Gerkmann Danilo de Oliveira, Tal Peer. 2022. Efficient Transformer-based Speech Enhancement Using Long Frames and STFT Magnitudes. In *Proceedings of Interspeech*. 2948–2952.

[5] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*.

[6] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. 1992. TIMIT Acoustic-phonetic Continuous Speech Corpus. *Linguistic Data Consortium* (11 1992).

[7] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=iTtGCMDEzS_

[8] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. 2022. SSAST: Self-Supervised Audio Spectrogram Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10699–10709.

[9] D. Griffin and J.S. Lim. 1984. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics Speech and Signal Processing* (1984).

[10] A. Vedaldi H. Chen, W. Xie and A. Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 721–725.

[11] Sindhu B. Hegde, K.R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2021. Visual Speech Enhancement Without a Real Visual Stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1926–1935.

[12] T. V. Ho and M. Akagi. 2020. Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*. 140–144.

[13] Tuan Vu Ho, Quoc Huy Nguyen, Masato Akagi, and Masashi Unoki. 2022. Vector-quantized Variational Autoencoder for Phase-aware Speech Enhancement. In *Proc. Interspeech 2022*. 176–180.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (2021), 3451–3460.

[15] Yuma Koizumi, Kohei Yatabe, Marc Delcroix, Yoshiki Masuyama, and Daiki Takeuchi. 2020. Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 181–185.

[16] X. Li and R. Horaud. 2020. Online monaural speech enhancement using delayed subband LSTM. In *Proceedings of Interspeech*. 2462–2466.

[17] Alan Papir Li Wan, Quan Wang and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[18] 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.

[19] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv preprint arXiv:1703.09452* (2017).

[20] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Vol. 2. 749–752.

[21] Michael P. Robb, Margaret A. Maclagan, and Yang Chen. 2004. Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics* 18, 1 (2004), 1–15.

[22] Y. Tsao S.-W. Fu, C.-F. Liao and S.-D. Lin. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning. PMLR*. 2031–2041.

[23] M. H. Soni, N. Shah, and H. A. Patil. 2018. Time-frequency masking-based speech enhancement using generative adversarial network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5039–5043.

[24] J. S. Chung T. Afouras and A. Zisserman. 2018. Lrs3-ted: a large- scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496, 2018* (2018).

[25] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (2011), 2125–2136.

[26] N. L. Westhausen and B. T. Meyer. 2020. Dual-signal transformation LSTM network for real-time noise suppression. In *Proceedings of Interspeech*. 2477–2481.

[27] Wang H. et al Yu W., Zhou J. 2022. SETransformer: Speech Enhancement Transformer. *Cogn Comput* 14 (2022), 1152–1158.