ClearAI: AI-Driven Speech Enhancement for Hypophonic Speech

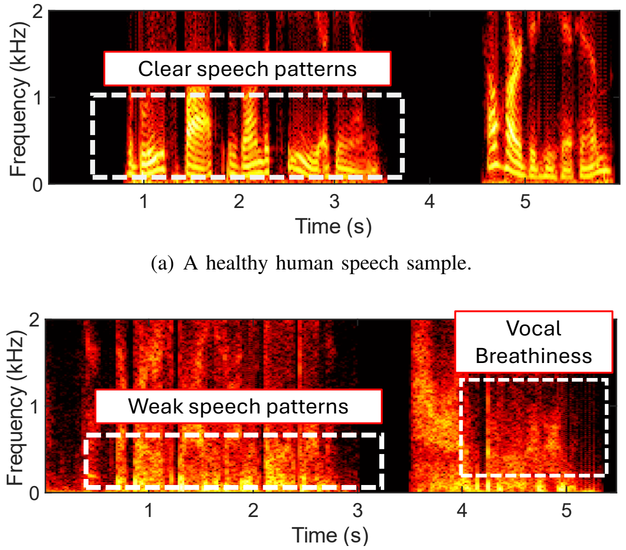
Yuanda Wang*, Qiben Yan*, Thea Knowles*, Daryn Cushnie-Sparrow†
* Michigan State University, East Lansing, USA
† Western University, London, Canada

Abstract*—Hypophonia is a common speech symptom related to Parkinson's disease, affecting human comprehension and effective communication. Unlike dysarthric speech, hypophonic speech is characterized by its low volume and breathy voice, which makes it challenging to be heard and understood by human and voice-controllable systems especially in noisy environments. Conventional speech enhancement techniques, primarily focusing on amplifying audio power or cancelling environmental noise, fall short in improving the intelligibility and perception for hypophonic speech. To enhance hypophonic speech, we present ClearAI, an innovative AI-powered technology to improve speech quality for individuals suffering from hypophonia. ClearAI first leverages voice conversion technology to create a parallel dataset composed of normal and corresponding hypophonic speech samples. Then, ClearAI incorporates a predictive model trained on augmented parallel data to estimate the optimal audio style from hypophonic speech to strengthen the audio intensity and enhance the speech patterns. Next, a speech restoration model is built on the generated parallel speech data to reconstruct clear speech from the style transferred speech. Our experimental results reveal that ClearAI leads to substantial improvements in audio intensity in both digital formats and over-the-air transmission. In addition, ClearAI successfully reduces the hypophonic speech recognition error rate by more than 30% in noisy environments. Our human test results also validate ClearAI enhanced speech has the best human perceptual quality compared with other baseline methods.

***Index Terms*—Hypophonic Speech, Speech Enhancement, Parkinson's Disease**

I. Introduction

Hypophonia (i.e., quiet speech) is a distinctive voice disorder, commonly seen in people with Parkinson's disease (PD). Characteristics of hypophonia reflect compromised laryngeal and respiratory control, which result in reduced speech intensity and breathy-hoarse voice quality. These symptoms are present in up to an estimated 89% of people with PD [1] and are accompanied by a broader set of speech symptoms collectively known as *hypokinetic dysarthria* [2]. The low intensity and vocal breathiness characteristic of hypophonic speech are partially attributed to incomplete vocal fold closure. Additionally, the reduced respiratory support often observed in PD can lead to shorter phrases punctuated by more frequent pauses, as well as compromised control over breathing, resulting in increased respiratory noise during speech. Fig. 1 shows the speech samples collected from a healthy human and a patient with PD using the same linguistic content. While the healthy human speech shows clear speech patterns, the hypophonic speech sample is dominated by vocal breathiness, weak high-frequency spectral amplitude, and a weak harmonic-to-noise structure. These features, coupled with other speech features of PD such as imprecise articulation, often lead to reduced speech clarity and intelligibility. In this study, we specifically define the characteristics of hypophonic speech as a lack of vocal cord vibrational features in speech, accompanied by a predominance of breath-related noise.



(a) A healthy human speech sample.

(b) A hypophonic speech sample.

Fig. 1: Spectrograms of healthy human speech and hypophonic speech from a PD patient.

It is evident that simply amplifying the intensity of a hypophonic speech signal would not improve the speech intelligibility [3]. However, intelligibility gains do often occur when people with hypophonia are cued to speak more effortfully, such as using a loud or clear speaking style [4]. This is likely because improvements in intelligibility are driven by a combination of prosodic, articulatory, and voice quality changes [5], rather than solely enhancing the audibility of the speech signal. For individuals who struggle to maintain these natural adjustments to their speech style, e.g., due to the progression of the disease, the use of a speech amplification device is recommended [6], which involves wearing a loudspeaker to boost their voice's audibility. Speech amplification devices, however, only serve to increase the amplitude of the incoming signal, and as a result do not improve speech clarity. Ideally, speech enhancement would be integrated into amplification technology to improve both audibility and intelligibility. While some speech reconstruction models have shown effectiveness in enhancing dysarthric speech intelligibility [7], their application is generally limited to speech with strong intensity and clear pronunciation. The inherent weakness of hypophonic speech requires a more sophisticated enhancement method beyond simple signal amplification or noise cancellation.

To address the limitations of existing methods, we introduce ClearAI, an AI-driven speech enhancement model designed for hypophonic speech. In order to augment the limited training data, ClearAI includes a data augmentation solution based on voice conversion to generate parallel *normal* and *hypophonic* speech samples, effectively improving the capacity of hypophonic speech enhancement. Subsequently, ClearAI applies audio style transfer to process the hypophonic speech. It does not merely amplify the speech signal intensity but also intelligently adjust the speech's frequency distribution, thereby improving overall quality and intelligibility of hypophonic speech. Finally, we utilize the augmented data to optimize a speech restoration model to rebuild the speech patterns. The experimental results demonstrate that ClearAI outperforms all other baseline methods in digital and physical communication scenarios. Additionally, ClearAI significantly improves hypophonic speech recognition accuracy by more than 30% in noisy scenarios. Furthermore, our human study results indicate that ClearAI not only improves the audio intensity of hypophonic speech, but also the intelligibility and perceptual quality. This paper makes the following contributions:

- We introduce a novel voice conversion approach that generates high-quality hypophonic speech from clear speech samples, effectively augmenting limited hypophonic speech data.
- Leveraging the augmented data, we optimize audio style transfer and speech restoration models to reconstruct clear speech from hypophonic speech with improved accuracy.
- Our experimental results demonstrate the potential of ClearAI to address the significant challenges faced by individuals with speech disabilities.

II. Methodology

Fig. 2 shows ClearAI's system architecture, which includes three modules: data augmentation, style prediction, and speech restoration. In the training phase, we use data augmentation to generate a large dataset consisting of hypophonic speech samples and normal speech samples with the same speech content. The dataset is used to optimize both the style prediction and speech restoration models. In the inference phase, the style prediction model estimates a target audio style for the hypophonic speech, and applies audio style transfer to enhance its intensity, as well as to balance signal power across various frequencies. The style-transferred audio will work as the input for the speech restoration model to reconstruct clear and intelligible speech.

A. Data augmentation

Generally speaking, during the training phase of speech enhancement models, various types of noise or distortions are artificially introduced to a dataset of clear speech [8], [9]. The noisy or distorted data will become the input of the enhancement model, and the initial clear audio is set as the target to train the speech enhancement model. However, existing methods are incapable of recovering healthy speech from hypophonic speech. Moreover, the vocal breathiness and weak harmonic structure of hypophonia speech make it impossible to convert hypophonic speech to healthy speech via signal transformation or noise injection.

To achieve effective hypophonic speech enhancement, it is essential to train a model on parallel normal and hypophonic speech data. The parallel speech samples should contain identical speech content to ensure content consistency, and they must be from the same speaker to maintain voice characteristics. However, collecting a large amount of parallel speech data from the same speaker is extremely challenging, if not impossible. Here, we propose to use voice conversion [10] to generate parallel data by replacing the voice while preserving the original speech content.

Our analysis of hypophonic speech data reveals variability in the severity of hypophonia depending on specific speech segments. For instance, Fig. 3 is a speech spectrogram sample from a PD patient. The patient's effort affects speech intensity and clarity. At the beginning of the speech, the patient can exert more efforts to enhance speech intensity. However, the speaker struggles to maintain high levels of attention and effort while speaking. By the end of the speech, the intensity degrades, and hypophonic speech patterns become evident. This insight guides us to selectively clip and patch speech samples, creating examples that reflect varying degrees of hypophonia. Through voice conversion, we obtain speech sample pairs where one sample exhibits the clarity of healthy speech, while the other retains hypophonic characteristics. The converted speech pairs share the same speech content from the source speech. Since the voice is from the same speaker but under different speaking conditions, the speech enhancement model can learn to reconstruct clearer speech signals without altering the speaker's unique voice characteristics.

In addition to the low speech intensity, a breathy voice is another characteristic feature of hypophonic speech. To more accurately replicate authentic hypophonic speech characteristics, we utilize human breath sounds as a reference to further process the generated hypophonic speech. Finally, the generated speech exhibits both low intensity and strong respiratory noise which resembles a natural hypophonic speech. These augmented speech samples are employed for training audio style prediction and speech restoration models in ClearAI to improve the clarity and quality of hypophonia speech.

B. Audio style transfer

Audio style transfer [11] is an emerging technology that is able to change the "texture" of the audio according to the reference audio. Initially, audio style transfer is widely applied to change the timbre in the music or natural sounds. In this work, we use audio style transfer to process hypophonic speech.

1) Preliminary study: For hypophonic speech samples, one of the primary issues impacting their clarity and intelligibility

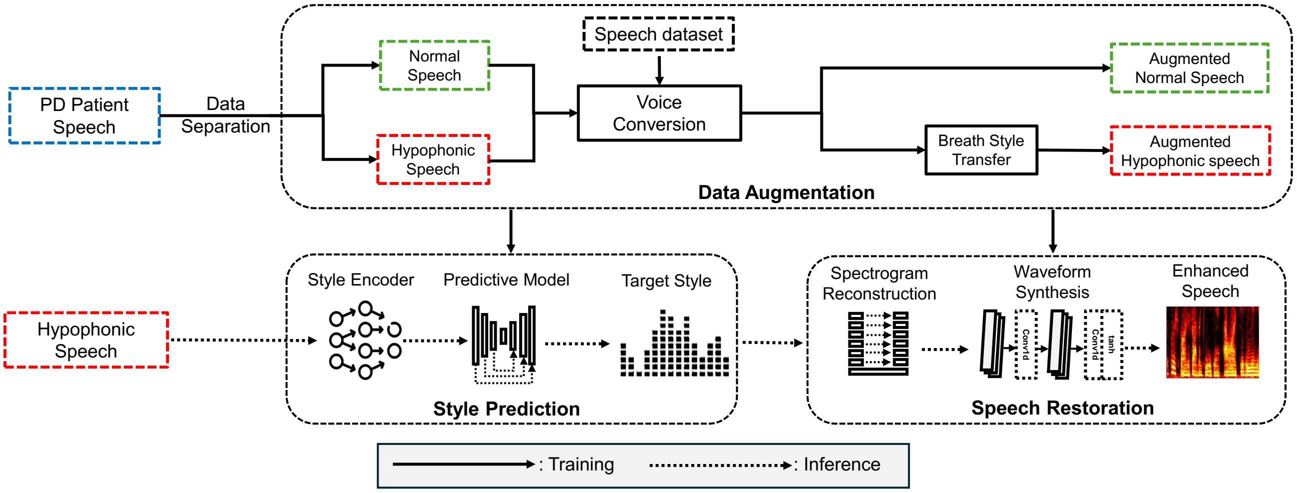


Fig. 2: ClearAI system overview.

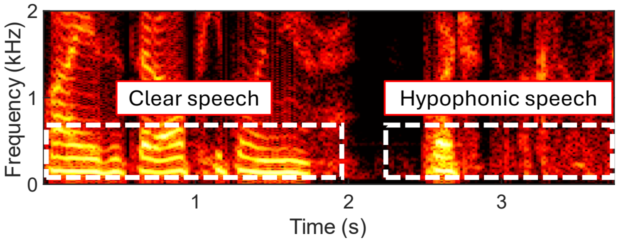


Fig. 3: A speech example from PD patient displays both clear and hypophonic speech patterns.

is their low vocal intensity. Since audio style transfer can be used to modify the "texture" of the audio, we can enhance the intensity of the hypophonic speech by embedding an audio style from a high-intensity audio, e.g., non-disordered human speech. In this work, we select DeepAFx-ST [12], a state-of-the-art model, to conduct audio style transfer in ClearAI. The style transfer process in DeepAFx-ST involves two primary inputs: the raw audio x and a style reference audio y. The audio style transfer can be formulated as:

$$x' = D(x, E(y)), \tag{1}$$

where E is the pre-trained style encoding module, D is a decoder to embed the target style $E(y)$ on the input speech. The transferred audio output x' retains the original audio pattern of x but adopts the audio style from y.

To intuitively demonstrate the effect of audio style transfer, we set a healthy speech as the style reference audio to enhance a hypophonic speech sample from a male PD patient. Fig. 4(a) and 4(b) show the speech spectrograms before and after style transfer. The transferred audio yields significant intensity improvement, especially at the end of the speech where the acoustic signals are extremely weak.

Meanwhile, the style transfer introduces differential signal processing method, enabling varied audio effects across different frequency bands of the input audio signal. In the audio style transfer process, the differential signal operator adjusts the distribution of frequency energy, such that the style transfer can be regarded as a sophisticated audio equalizer. Therefore, it can selectively reduce the intensity of frequencies with strong breathing noise and amplify the attenuated spectral frequencies associated with a weak voice. Fig. 4(c) shows the Fast Fourier Transformation (FFT) results of the a short pronunciation in the raw speech and transferred speech. The results demonstrate that the frequency range containing human speech (1 $\sim$ 3 kHz) is amplified, leading to improved speech clarity.

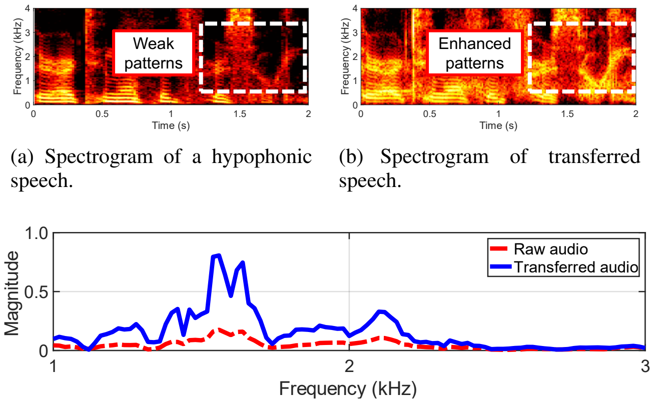
2) Style prediction: The goal of ClearAI is to assign a unique audio style for individual speech sample to achieve the best enhancement performance. Since we only have the hypophonic speech as the input x, we use a deep neural network (DNN) model to directly predict the target audio style without a reference audio. The audio style encoder E, which is composed of multi-layer perceptron (MLP) [13], will encode the input audio with ambient length as an 1×1024 linear style vector. To ensure the output vector matches the original dimensions, the predictive model F_p employs a symmetrical architecture with 4 down-sampling and 4 up-sampling layers. Moreover, as shown in Fig. 2, F_p is a residual network applying skip connections between down-sampling and up-sampling layers with the same dimension. We use the augmented healthy speech x_c and hypophonic speech x_h to optimize the style prediction model:

$$\underset{\gamma}{\text{minimize}} \;\; \mathcal{L}(\mathcal{F}_p(E(x_h)), E(x_c)), \tag{2}$$

where γ is the parameters in $\mathcal{F}_p$ and $\mathcal{L}$ is mean squared error (MSE) loss. In the inference phase of ClearAI, when we derive the hypophonic speech input, $\mathcal{F}_p$ predicts an optimized audio style $\mathcal{F}_p(x)$, based on which we can directly apply the audio style transfer model as: $x' = D(x, \mathcal{F}_p(E(x)))$. The outcome is a preliminary enhanced speech x'.

C. Speech restoration

Speech restoration models are designed for recovering high-quality speech from damaged or noisy speech. Inspired by VoiceFixer [9], the speech restoration module in ClearAI

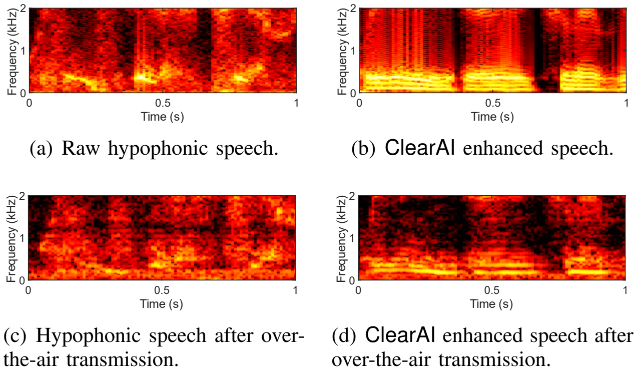


(a) Spectrogram of a hypophonic speech.

(b) Spectrogram of transferred speech.

(c) FFT results of the raw audio and transferred audio.

Fig. 4: Audio style transfer improves the speech intensity and equalizes frequency distribution for hypophonic speech.



(a) Raw hypophonic speech.

(b) ClearAI enhanced speech.

(c) Hypophonic speech after over-the-air transmission.

(d) ClearAI enhanced speech after over-the-air transmission.

Fig. 5: Spectrograms of hypophonic speech and ClearAI enhanced speech in digital and physical transmission scenarios.

consists of two stages: mel-spectrogram reconstruction and waveform synthesis. In order to train the mel-spectrogram reconstruction model, we set the augmented hypophonic speech in Section II-A as the input, and the corresponding healthy speech as the target output. Also, to improve the robustness of speech restoration, we apply regular signal distortions, including additional noise, low resolution, and amplitude clipping on the hypophonic speech x_h. Next, x_h is converted to a 2D mel-spectrogram matrix X_h to reduce the model complexity. The mel-spectrogram reconstruction model, featuring a similar architecture to Res-UNet [14], comprises 6 encoder and 6 decoder blocks. Each block contains a convolutional (or transposed convolutional) layer paired with a corresponding down-sampling (or up-sampling) layer. The training phase of the mel-spectrogram reconstruction model can be formulated as:

$$\underset{\theta}{\text{minimize}}\ \mathcal{L}(\mathcal{F}_r(X_h), X_c), \tag{3}$$

where $\mathcal{F}_r$ is the model function with parameter θ, X_h and X_c are respectively the mel-spectrogram of the distorted hypophonic speech and healthy speech, and $\mathcal{L}$ is mean absolute error (MAE) loss function [15].

Next, a TFGAN [16] based vocoder is trained to convert mel-spectrogram to an audible waveform. The network consists of convolutional blocks and up-sampling layers to reduce the frequency domain dimension and extend the length of output signal. Finally, a (128, T) mel-spectrogram will be converted to a (1, 441 $\times$ T) waveform, where T is 10 milliseconds. The final inference of speech restoration model is $x_{out} = \mathcal{G}(\mathcal{F}_r(X'))$, where X' is the mel-spectrogram of the audio style transfer output x' and $\mathcal{G}$ is the pre-trained vocoder function.

III. Evaluation

A. Dataset introduction

Currently, there is a lack of dedicated speech datasets tailored for hypophonic speech collection. As a result, we select relevant speech samples from existing datasets and employ data augmentation techniques to facilitate large-scale training. We collect hypophonic speech samples in Perceptual Voice Qualities Database (PVQD) [17] and a curated dataset of hypophonic speech (herein referred to as the Cushnie-Sparrow Database [18]. PVQD contains 296 speech samples and each sample is from a individual speaker from 14 to 93 years old. 207 of them have different levels of voice complaints. Each sample contains 5 complete sentences and 2 vowels. The Cushnie-Sparrow dataset is a dataset collected from 56 PD patients from 54 to 88 years old (42 males and 14 females). Each speaker produced a set of 11 sentences from the Sentence Intelligibility Test [19], ranging from 5 to 15 words in length. For both datasets, the records were captured through a headset condenser microphone placed 6 cm from the speaker's mouth to ensure the best audio clarity.

B. Experiment setup

In the data augmentation phase, we choose VCTK Corpus dataset [20] as source speech to generate hypophonic speech. After training the style prediction and speech restoration models, we select hypophonic speech samples from PVQD and Cushnie-Sparrow datasets as the input and obtain the enhanced speech. We compare the speech enhancement performance of ClearAI with 5 baseline methods: ① Equalizer: amplifying the most sensitive frequencies (1.6 $\sim$ 4 kHz) according to equal-loudness contour [21], ② SEGAN [8], a speech enhancement generative network, ③ SpeechBrain [22], a general speech toolkit, ④ AudioSR [23], an audio super-resolution based on diffusion model, ⑤ VoiceFixer [9], a speech restoration model for audio distortion recovery. In our evaluation, the loudness of all speech samples are normalized as 70 dB SPL to eliminate the impact of volume.

C. Audio quality comparison

Although many metrics can be used for speech quality assessment, e.g., Perceptual evaluation of speech quality (PESQ) [24], they are all based on the comparison between clean speech and distorted speech. These metrics are not suitable for evaluating the quality of hypophonic speech as the raw speech has lower quality. Therefore, we choose Signal-to-Noise Ratio (SNR) to evaluate the audio quality of hypophonic

TABLE I: SNR comparison of unprocessed hypophonic speech and enhanced speech from different enhancement methods.

Methods	Digital (dB)	Over-the-air (dB)
Raw speech	12.41	7.70
Equalizer	13.25	8.55
SpeechBrain	14.11	8.43
SEGAN	10.35	7.14
AudioSR	13.85	8.44
VoiceFixer	14.52	8.80
ClearAI	**15.97**	**9.67**

speech and existing enhancement methods. In Table I, we list the SNR values of hypophonic speech and enhanced speech from ClearAI and other existing methods in digital format. The equalizer can improve the speech intensity but the vocal breathiness is still strong. For speech enhancement models targeting noise cancellation, SpeechBrain can improve the audio quality by mitigating the background noise. On the other hand, SEGAN degrades the speech quality as it incorrectly eliminates weak speech signals. AudioSR method reconstructs the high-frequency band but leaves the low-frequency signals unchanged. VoiceFixer shows better performance since it can effectively filter out the noise and compensate weak audio patterns. Among all enhancement methods, ClearAI shows the best SNR results. Fig. 5(a) and Fig. 5(b) show the comparison of raw speech and enhanced speech by ClearAI. Because ClearAI is trained on hypophonic samples and corresponding healthy speech, it shows better capacity while reconstructing hypophonic speech with low intensity and strong breathy voice.

In an over-the-air transmission scenario, it is challenging for microphones to adequately capture hypophonic speech. To evaluate ClearAI performance in over-the-air transmission, we use a loudspeaker to play all speech samples with the same volume setup, and then we place a smartphone (iPhone 13) 1.2 meters away from the loudspeaker to record the audio with 48 kHz sampling rate. The hypophonic speech example is shown in Fig. 5(c). The higher frequency spectral components of the signal will attenuate during over-the-air propagation, further weakening the captured audio intensity. This attenuation pattern is worse at greater distances. While methods such as AudioSR can reconstruct high-frequency components to boost digital audio power, it is less effective during over-the-air transmission since high frequency signals are vulnerable during over-the-air transmission. In comparison, as shown in Fig. 5(d), ClearAI yields the clearest speech patterns in the replay recording, because it enhances the audio intensity while eliminating the vocal breathiness in the speech.

D. Speech recognizability

ASR models can recognize unprocessed hypophonic speech in quiet environments as they use denoising methods to eliminate breathy noise. However, the hypophonic speech is not robust in noisy environments. To address the human computer interaction for hypophonic patients, we apply ClearAI in the amplification process to enhance the speech. We select the OpenAI's open-sourced ASR model, Whisper, and statistic its word error rate (WER) for evaluation. We use ClearAI to process the hypophonic speech and compare the ASR accuracy with the unprocessed raw speech signal. Additionally, we inject multi-talker noise into the speech samples to simulate noisy environments with other speakers. In the evaluation, all the speech samples are normalized to 75 dBFS, and we adjust the noise power to manipulate the input audio SNR.

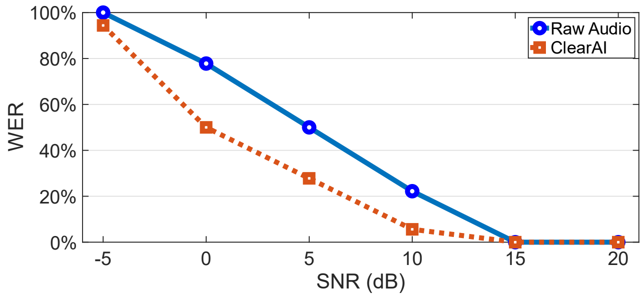


Fig. 6: WER comparison under different noise levels.

Fig. 6 presents the WER comparison results under different noise levels. When the SNR is high, for example, greater than 15 dB, the ASR model achieves a 0% WER even for unprocessed hypophonic speech samples. As we increase the noise level to an SNR of 10 dB, the ASR model fails to recognize 22% of the words in the speech, as the noise overwhelms the weakest speech patterns. In comparison, the speech samples enhanced with ClearAI maintain a very low WER ($\sim 4\%$) since ClearAI eliminates the breathy noise and selectively enhances the human speech patterns in the audio samples. Notably, even under extremely strong noise levels (SNR = 0 dB), where the noise has similar power to human speech, the audio samples enhanced by ClearAI can still be recognized with less than 50% WER, achieving 30% of improvement compared with the raw hypophonic speech samples. The results show that ClearAI can effectively improve the recognizability of noisy hypophonic speech. ClearAI can also be conveniently deployed on mobile devices to enhance speech signals.

E. Human study

We also conduct a human perception study to compare the intelligibility of the enhanced speech samples from the human listeners' perspective. 16 volunteers (10 males and 6 females, 24 to 33 years old, all with normal hearing ability) participate in this evaluation. We select 4 speech samples from the Cushnie-Sparrow dataset and 4 samples from PVQD dataset with hypophonia symptom for the test. First, we let the volunteers listen to the raw speech audio as a reference. Next, they listen to the enhanced speech samples from ClearAI and other methods. After that, they will rank the speech clarity and intelligibility for all enhanced samples from 1 (lowest) to 6 (highest) based on their personal judgement.

Fig. 7 displays the average scores for different speech enhancement methods. Among all methods, SEGAN has the lowest score because it incorrectly filtered out some speech features, rendering the enhanced speech incomplete. VoiceFixer primarily targets regular audio distortions. Therefore,

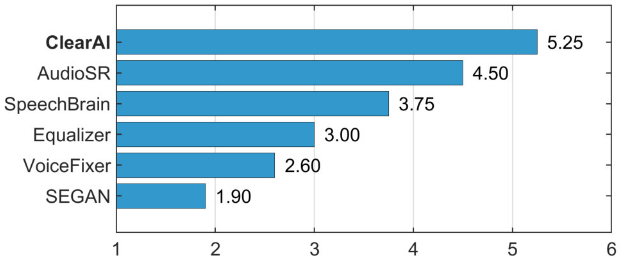


Fig. 7: Human test result of different speech enhancement methods on weak hypophonic speech samples (x-axis denotes the speech clarity and intelligibility score from 1 to 6).

it presents an unnatural timbre when restoring hypophonic speech. The equalizer method can enhance the energy of frequencies that human hearing is sensitive to, while SpeechBrain's speech enhancement tool mitigates breathing noise, thereby improving speech clarity. AudioSR reconstructs the high-frequency components of speech, improving the speech clarity and stereophony. Compared to all these methods, ClearAI consistently achieves the highest average scores, corresponding to the highest audio quality and intelligibility. The training mechanism on parallel healthy speech and hypophonic speech not only restores the speech content but it also corrects the hypophonic timbre, bringing it closer to the speech from healthy individuals.

IV. Conclusion

In this paper, we present ClearAI, an AI-driven speech enhancement model designed for hypophonic speech. Facilitated by data augmentation, ClearAI strengthens speech features by leveraging audio style transfer. It then utilizes a two-stage speech restoration model to reconstruct the clear speech audio. Our evaluation results show that ClearAI's enhanced speech achieves the highest SNR compared with other methods in both digital and physical environments, and it can reduce ASR error rate when recognizing hypophonic speech in noisy environments. Moreover, the results from human study demonstrate that ClearAI delivers the clearest and most intelligible speech among all the enhanced hypophonic speech samples.

Acknowledgment

We deeply appreciate the anonymous reviewers for their valuable comments. This research is supported by National Science Foundation Grant CNS-2310207.

References

[1] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *The Journal of Speech and Hearing Disorders*, vol. 43, no. 1, p. 47–57, Feb. 1978.

[2] S. G. Adams and A. D. Dykstra, *Hypokinetic dysarthria*. New York: Thieme Publishing Group, 2009.

[3] A. T. Neel, "Effects of loud and amplified speech on sentence and word intelligibility in parkinson disease," 2009.

[4] K. Tjaden, J. E. Sussman, and G. E. Wilding, "Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis," *Journal of speech, language, and hearing research: JSLHR*, vol. 57, no. 3, p. 779–792, Jun. 2014.

[5] Y. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, p. 417–429, 2011.

[6] T. Knowles, S. G. Adams, A. Page, D. Cushnie-Sparrow, and M. Jog, "A comparison of speech amplification and personal communication devices for hypophonia," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 8, pp. 2695–2712, 2020.

[7] Y. Wang, X. Wu, D. Wang, L. Meng, and H. Meng, "Unit-dsr: Dysarthric speech reconstruction system using speech unit normalization," *arXiv preprint arXiv:2401.14664*, 2024.

[8] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[9] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," *arXiv preprint arXiv:2204.05841*, 2022.

[10] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5954–5958.

[11] E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 586–590.

[12] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *arXiv preprint arXiv:2207.08759*, 2022.

[13] A. A. Heidari, H. Faris, S. Mirjalili, I. Aljarah, and M. Mafarja, "Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks," *Nature-Inspired Optimizers: Theories, Literature Reviews and Applications*, pp. 23–46, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[16] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, "Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis," *arXiv preprint arXiv:2011.12206*, 2020.

[17] P. R. Walden, "Perceptual voice qualities database (pvqd): Database characteristics," *Journal of Voice*, vol. 36, no. 6, pp. 875.e15–875.e23, 2022.

[18] D. A. Cushnie-Sparrow, "Modelling loudness: Acoustic and perceptual correlates in the context of hypophonia in parkinson's disease," Ph.D. dissertation, The University of Western Ontario (Canada), 2021.

[19] K. M. Yorkston, D. R. Beukelman, and R. Tice, "Sentence intelligibility test," *Communication Disorders Software. Distributed by the Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital, Lincoln, NE.*, 1996.

[20] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[21] H. Guo, Y. Wang, N. Ivanov, L. Xiao, and Q. Yan, "Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 1353–1366.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[23] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "Audiosr: Versatile audio super-resolution at scale," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1076–1080.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.