Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

NAM-to-Speech Conversion with Multitask-Enhanced Autoregressive Models

Neil Shah[1,2], *Shirish Karande*[2], *Vineet Gandhi*[1]

[1]CVIT, Kohli Centre for Intelligent Systems, IIIT Hyderabad, India
[2]TCS Research, Pune, India

neilkumar.shah@tcs.com, shirish.karande@tcs.com, vgandhi@iiit.ac.in

Abstract

We propose an alignment-free, end-to-end Non-Audible Murmur (NAM)-to-Speech conversion model. Existing methods rely on large NAM-text pairs per speaker to generate high-quality alignments for training non-autoregressive models. However, alignment quality deteriorates when trained on multi-speaker data, limiting their ability to generalize and effectively utilize the available training data. To address this, we introduce a streamlined autoregressive approach that eliminates the need for explicit alignment learning. By leveraging multi-speaker samples, synthetic training pairs, and multitask character recognition training, our method reduces the word error rate (WER) by 59.19% compared to the state-of-the-art (SOTA) on two public datasets. We demonstrate the model's zero-shot capability and validate the effectiveness of multitask training through ablation studies. Speech samples are available at `https://noalignnam.github.io/autoregressiveNAM/`

Index Terms: Non-Audible Murmur, NAM-to-speech, Autoregressive models, Seq2Seq networks

1. Introduction

A Silent Speech Interface (SSI) is a communication system synthesizing speech by capturing non-audible physiological signals associated with speech production. It offers versatile benefits, including facilitating communication in noisy or silent environments [1], assisting individuals with speech impairments [2], and providing user-friendly, customizable solutions that promote independence and immediate interaction [3]. Recent research in this field has made significant advancements [1, 4, 5, 6]. This progress is driven by a range of SSI techniques, such as the effective fine-tuning of self-supervised models on articulatory videos recorded using MRI [7] and the development of multi-speaker speech recognizers using ultrasound [6] to decode text content. These techniques provide deeper insights into the role of each vocal organ in sound production and enable the synthesis of natural, intelligible speech.

While these interfaces have advanced articulatory speech synthesis, many face challenges for everyday use due to their non-real-time nature [7], invasive characteristics [8], and issues such as equipment noise and poor lighting conditions [4]. Nakajima et al. [9] introduced a non-invasive SSI that used a specialized microphone to capture NAM (acoustically unintelligible and incomprehensible signals to nearby listeners) from tissues behind the ear. Although NAM-to-Speech conversion offers an intuitive approach to SSI, its success has been limited due to the absence of *"aligned"* ground-truth speech as training targets for the input NAMs. This challenge arises from data collection, as an individual cannot record ground-truth speech and NAM vibrations simultaneously [1, 9]. To address this, researchers typically follow a two-step process [1, 10, 11]: (a) exploring techniques to "*simulate aligned ground-truth speech*" as training targets to generate paired NAM-speech data, and (b) training a non-autoregressive (NAR) sequence-to-sequence (Seq2Seq) model using the paired NAM-speech data. Most efforts have focused on refining the alignment of NAM-speech training data, while the NAR Seq2Seq architecture has remained unchanged. Despite this progress, several challenges continue to hinder further development in this domain.

- The existing SOTA method [1] builds high-quality speaker-specific models by relying on large per-speaker NAM-text samples to learn precise alignments. However, alignment quality significantly degrades when trained on multiple speakers due to speaker variability. This degradation adversely affects the NAR Seq2Seq backbone, which depends on perfectly aligned data. Even minor misalignments in the training pairs can substantially compromise the intelligibility of the synthesized speech.
- An alternative approach is to simulate ground-truth speech from lip movements [11] directly, eliminating the need for explicit alignment learning, as NAMs are inherently aligned with lip movements. However, this method faces challenges in generalization and maintaining intelligibility across diverse speakers [11].

Generating aligned ground-truth training pairs is essential not only for existing NAM-to-Speech conversion methods [1, 10, 11] but also for other SSIs, such as Ultrasound Tongue Imaging (UTI) and rtMRI [12, 7]. However, traditional phoneme-speech alignment methods are often error-prone, introducing inconsistencies that hinder model performance. Our work challenges this dependency by eliminating reliance on such alignments, paving the way for more robust and scalable training strategies that enhance the effectiveness of SSI techniques.

In this work, we simplify the existing two-step process for NAM-to-Speech conversion by eliminating the need to generate "time-aligned" ground-truth speech. Instead, we leverage the ability of autoregressive (AR) Seq2Seq models to predict the next token(s) consecutively. Specifically, our contributions include:

- Proposing a streamlined transformer-based NAM-to-discrete speech network with an auxiliary character prediction task, utilizing an AR Seq2Seq model to eliminate the need for aligned NAM-speech training pairs, ensuring robust NAM-to-Speech conversion.
- Introducing the multi-speaker NAM-to-Speech conversion model that leverages additional training pairs from multiple speakers while also utilizing augmented samples.
- Achieving a 59.19% average reduction in WER across two

10.21437/Interspeech.2025-1537

public datasets compared to the SOTA [1] in a multi-speaker setting while also enhancing generalization in zero-shot scenarios.

2. Related work

Nakajima et al. [9] pioneered a specialized microphone for capturing NAM vibrations. Later efforts focused on improving NAM microphone design [13, 14, 15]. A typical NAM-to-Speech conversion system requires aligned ground-truth speech corresponding to input NAMs to train a deep learning framework. Prior methods [16, 17, 18] explicitly use studio-recorded speech as ground-truth and apply dynamic time warping to align training pairs for Seq2Seq learning.

In our previous work [10], we were the first to explore the idea of "*simulating*" aligned ground-truth speech, moving away from reliance on studio-recorded speech for training. Specifically, we investigated the use of whisper data for this simulation; however, the approach faced generalization challenges across different speakers. It depended on paired whisper data, which may not be available, especially for patients with speech difficulties. To avoid relying on whisper data, we later introduced the StethoSpeech [1] approach, which learns speaker-specific phoneme-level duration alignments between NAMs and text and then synthesizes aligned ground-truth using a pre-trained text-to-speech (TTS) module. However, with limited NAM data per speaker and in a multi-speaker scenario, alignments tend to be noisy, significantly degrading the intelligibility of converted speech [11]. To eliminate reliance on NAMs and text for ground-truth simulation, our recent work [11] explored lip-to-speech synthesis using diffusion methods to generate aligned speech directly from lip movements in a zero-shot manner. However, the approach encountered generalization and intelligibility issues despite fine-tuning the diffusion model using video modalities from speakers available in the dataset. As a result, the StethoSpeech [1] method remains the current best solution for simulating speaker-specific aligned ground-truth speech. All these methods used NAR Seq2Seq models as the backbone to learn the correspondence between the NAM and speech modalities in the latent space. We also further emphasized data augmentation to improve NAR Seq2Seq models' intelligibility [10]. However, these models fail to leverage augmentation when trained on speakers from different regions.

This work eliminates reliance on high-quality alignments for ground-truth speech simulation by proposing a unified Seq2Seq framework that enables alignment-free conversion by leveraging multi-speaker and synthetically augmented training pairs. While similar techniques exist for speech-to-speech translation models [19, 20], their effectiveness for non-speech signals remains unvalidated, and our work explores this aspect by demonstrating the potential of alignment-free strategies in SSI tasks. We propose directly mapping raw, misaligned NAM vibrations to discrete speech units using AR translation to achieve this. These speech units are derived from simulated ground-truth speech using a HuBERT encoder [21], chosen for its superior ability to preserve detailed content information while reducing speaker-specific and background noise [22, 23, 24].

3. Method

Our proposed system comprises three main modules: simulating ground-truth speech from available text, augmenting data to generate additional NAM-speech training pairs, and utilizing a transformer-based NAM-to-unit translation (N2UT) model with a NAM encoder and a discrete speech decoder, as shown in Figure 1. During training, we use auxiliary tasks conditioned on intermediate representations from the NAM encoder and the discrete speech decoder to predict characters. A vocoder is trained separately to convert the predicted discrete units into speech.

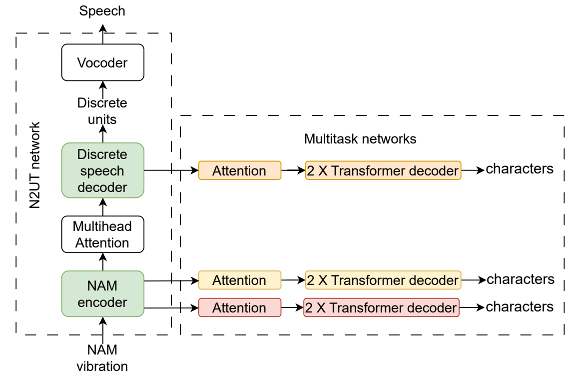

Figure 1: *Illustration of the end-to-end NAM-to-Speech conversion framework. The framework comprises: (1) a transformer-based NAM-to-unit translation (N2UT) model with a NAM encoder and discrete speech decoder, (2) multitask networks conditioned on both the encoder and decoder, and (3) a vocoder trained to convert the predicted units into speech.*

3.1. Ground-truth speech simulation

The current SOTA method [1] trains an acoustic model using NAM-text pairs to predict phoneme durations, which a TTS model then utilizes for aligned speech synthesis. In contrast, our proposed framework eliminates the need for aligned data and directly utilizes text with a FastSpeech2 [25] TTS module to simulate ground-truth speech.

3.2. Data augmentation

Acquiring large-scale NAM vibrations is expensive, which limits the effectiveness of Seq2Seq networks. Data augmentation with synthetic data has been shown to enhance the quality of synthesized speech [10, 26]. Following the approach in [10], we first extract HuBERT representations from NAM vibrations and train a vocoder to synthesize speech in the NAM voice. Subsequently, we extract HuBERT representations from the LJSpeech dataset [27] and the IndicTTS dataset [28]. These representations are passed through the trained vocoder to generate speech in the NAM style. This method effectively expands the NAM corpus, augmenting it to include approximately 60 hours of additional unaligned NAM-speech training pairs.

3.3. NAM-to-unit translation (N2UT) model

NAM encoder: We build the NAM encoder and a discrete speech decoder by adapting the transformer architecture designed by [29] and using fairseq repository[1]. While existing approaches [1, 10, 11] rely on HuBERT embeddings to process input NAM vibrations, we take a different approach by using mel-filterbank features. As predominantly low-frequency, NAM signals benefit from mel-filterbank features' ability to capture low-

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs

level acoustic information that HuBERT representations may miss. Additionally, mel-filterbank features allow the encoder to learn NAM-specific patterns without being constrained by the pre-encoded linguistic features inherent in HuBERT. For the input NAM vibrations, we extract 80-dimensional mel-filterbank features at every 10 ms, followed by cepstral mean and variance normalization. These features undergo a series of preprocessing steps, implemented as a module appended before the transformer layers in the NAM encoder. To preprocess the input features, we apply two $1D$ convolutional layers with a kernel size of 5 and 1024 channels, resulting in a downsampling factor of 4 on the input NAM vibrations. The convolutional layers reduce the temporal resolution while preserving essential acoustic features, with each layer followed by a gated linear unit activation to retain critical NAM information and reduce computational complexity [30]. The NAM encoder comprises 12 transformer layers, each with an embedding size of 512, a feedforward network (FFN) embedding size of 2048, and 8 attention heads.

Discrete speech decoder: To extract discrete ground-truth speech units as training targets, we use the HuBERT [21] Base model [2], which learns self-supervised representations emphasizing only the speech content. A HuBERT model, pre-trained on an unlabeled speech corpus in the target language (e.g., English), encodes the speech into continuous representations at every 20 ms frame. These learned representations undergo k-means clustering to generate 100 cluster centroids, representing the target ground-truth speech utterances as sequences of cluster indices. To accelerate training and inference, as recommended by [31], we collapse consecutive identical HuBERT units in the target sequence into a single unit, resulting in a sequence of unique discrete units. The discrete speech decoder comprises 6 transformer layers, with an embedding size of 512, an FFN embedding size of 2048, and 8 attention heads. A linear layer is appended to the output to project the decoder's representations into the predicted speech units. We train the N2UT model using cross-entropy loss with label smoothing. We train the models for $1000k$ steps using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$, and label smoothing of 0.2.

Multitask training: Supervision from text transcripts is incorporated into the N2UT model through co-training with an auxiliary module, as shown in Figure 1. We observe that predicting characters corresponding to input NAM is crucial for effectively addressing the NAM-to-Speech conversion task. Losses from the auxiliary networks are integrated only during training to help the primary discrete speech decoder learn intelligible speech units. This approach also encourages the model to disentangle linguistic content from other information in the mel-filterbank features, promoting feature diversity and reducing the risk of overfitting. Based on preliminary experiments, we append an attention module and a decoder to the eighth and tenth layers of the NAM encoder. Additionally, an attention module and decoder are attached to the third intermediate layer of the discrete speech decoder. This is because conditioning on the last few layers of the discrete speech decoder, which is trained to predict speech units, may result in suboptimal text generation. Each auxiliary loss is given a weight of 8.0. The model includes multi-head attention modules with 4 heads, while the decoders consist of 2 transformer layers, with 512-dimensional embeddings and an FFN embedding size of 2048.

Vocoder: For unit-to-speech conversion, we employ a modified version of the HiFi-GAN vocoder, as outlined in [22, 32]. The vocoder is trained on the LJSpeech dataset, with a batch size of 32, a learning rate of 2×10^{-6}, 100 embeddings, an embedding dimension of 128, and an input dimension of 512. The HiFi-GAN vocoder generates speech from units predicted by the N2UT network.

[2]https://github.com/pytorch/fairseq/tree/master/examples/hubert

Table 1: *Error rates in speaker-specific training.*

Dataset	Speaker	Ours		StethoSpeech [1]	
		CER ↓	WER ↓	CER ↓	WER ↓
StethoText	$S01$	43.12	89.16	**9.76**	**20.78**
	$S02$	41.45	89.16	**12.19**	**26.85**
MultiNAM	$S01$	49.32	96.89	**27.17**	**56.65**
	$S02$	51.98	105.84	**34.08**	**69.45**

4. Experiments

We evaluate the effectiveness of our framework using NAM-text pairs from two public datasets: the StethoText [1] corpus, which comprises NAM vibration and text pairs, and the MultiNAM [11] corpus, which includes paired NAM, video, whisper, and text data. The StethoText corpus includes 4.38 hours of data from a male speaker ($S01$) and 4.18 hours from a female speaker ($S02$). Additionally, it provides 7 hours of data from ten other speakers ($S03$ to $S12$), primarily intended for evaluating method performance in a zero-shot setting. The MultiNAM corpus provides 5.66 hours of data from a different male speaker ($S01$) and 2.30 hours from a different female speaker ($S02$). In both datasets, the subjects were instructed to murmur while holding the stethoscopic microphone below the back of their ears on the tissue region. In Section 4.1, we evaluate the converted speech using StethoSpeech [1] and our proposed method in a speaker-specific setting. In Section 4.2, we assess the impact of training these methods using all available data, including a synthetically generated dataset. In Section 4.3, we conduct ablation studies to gauge the impact of multitask training. Additionally, in Section 4.4, we explore the performance of these methods in a zero-shot scenario. To assess the NAM-to-Speech conversion performance, we compute character error rate (CER) and WER scores as an objective measure of speech intelligibility, using OpenAI's whisper-medium speech recognizer [33].

4.1. Speaker-specific training

Table 1 compares StethoSpeech [1] and our proposed method on the complete NAM-to-Speech conversion task in a speaker-specific setting across both datasets. In this setting, each method is trained exclusively on data from a single speaker and evaluated on a 5% test set from the same speaker. StethoSpeech outperforms our proposed method in intelligibility but relies heavily on a large quantity of aligned NAM-ground truth pairs available from a single speaker. This is evident from its performance: with 4.38 hours of aligned data from speaker $S01$ in the StethoText corpus, it achieves the lowest WER of 20.78%. However, with just 2.30 hours of aligned data from speaker $S02$ in the MultiNAM corpus, the WER increases dramatically by 234.21%. These results highlight that StethoSpeech requires substantial aligned data per speaker to achieve highly intelligible speech conversion.

Table 2: *Error rates using all data and synthetic NAM data.*

Dataset	Test speaker	Ours		StethoSpeech [1]	
		CER ↓	WER ↓	CER ↓	WER ↓
StethoText	$S01$	**7.87**	**16.97**	14.53	30.51
	$S02$	**11.21**	**18.37**	13.11	30.175
MultiNAM	$S01$	**9.94**	**17.85**	40.51	75.89
	$S02$	**19.11**	**40.10**	70.54	135.30
	all	**8.65**	**18.44**	23.89	45.19

Table 3: *Performance comparison of multitask network configurations for NAM-to-Speech conversion.*

Multitask loss	CER ↓	WER ↓
None	22.32	39.48
3^{rd} layer speech decoder	19.08	33.27
8^{th} layer NAM encoder	10.91	20.64
10^{th} layer NAM encoder	10.95	20.32
all	**8.65**	**18.44**

4.2. Training with combined and synthetic NAM data

StethoSpeech performs well with many NAM-text pairs from a single speaker to generate high-quality alignments. However, in extreme cases, such as patients with chronic obstructive pulmonary disease [34], where muscles are too weak to murmur, it becomes infeasible for individuals to record large amounts of NAM vibrations. This underscores the need for effectively leveraging data from other speakers, including augmented data.

Table 2 compares the performance of these methods on converted speech trained using all available NAM-text pairs from both datasets, along with synthetic data, resulting in 83.52 hours of training pairs. For our proposed method, since aligned ground-truth speech is not required, we directly pass the text to the FastSpeech2 TTS model to generate the ground-truth speech, as explained in Section 3.1. For StethoSpeech, the acoustic model was trained on the NAM-text data to obtain aligned durations, which, along with the text, were processed through the FastSpeech2 TTS model to generate aligned ground-truth speech. Results are reported on a 5% test split for two primary speakers from each dataset. Our proposed method achieves the best performance, showing significant improvement across all speakers compared to the speaker-specific setting in Section 4.1, by leveraging synthetic training targets and all available data. Interestingly, augmenting data in StethoSpeech reduces its performance, likely due to the pre-trained HuBERT model generating out-of-distribution samples, as it was not trained on geographically relevant speaker data. In contrast, our method leverages out-of-domain augmented samples to generate highly intelligible speech consistently. Additionally, the last row in Table 2 shows the performance of these methods when tested on a 5% test split from all speakers, with our method reducing the WER by over 59.19% and the CER by over 63.79% compared to StethoSpeech.

4.3. Ablation studies on multitask training

Table 3 compares the performance of our model trained with various multitask network combinations to a baseline that conditions intermediate representations from all auxiliary decoders trained on the same data. Training without auxiliary decoders results in poor intelligibility, as the model struggles to generate coherent syllables despite occasionally producing plausible sounds. Conditioning the discrete speech decoder at the third layer slightly enhances intelligibility and helps the model capture complementary information for speech units. Conditioning on the NAM encoder at the eighth and tenth layers provides significant performance improvements, demonstrating that leveraging the NAM encoder to extract linguistic information from low-level acoustics enhances the discrete speech decoder's ability to learn coherent speech units. Combining all tasks yields the best results, underscoring the effectiveness of an alignment-free, end-to-end NAM-to-speech conversion approach.

Table 4: *Error rates on StethoText corpus in a zero-shot setting.*

Test speaker	Ours		StethoSpeech [1]	
	CER ↓	WER ↓	CER ↓	WER ↓
$S11$	**13.34**	**27.72**	22.64	32.37
$S12$	**14.45**	**30.65**	22.54	34.19
all male	**18.98**	**34.81**	24.90	36.80
all female	**16.78**	**33.70**	27.43	39.29

4.4. Zero-shot evaluation

Table 4 presents the evaluation of converted speech for speakers whose data was not used during training. Since StethoSpeech does not work well with augmentation, it is trained solely on the available data from the dataset. In contrast, our proposed method is trained on all available NAM-text pairs, excluding the samples from the test speaker. We conducted two studies: first, we reserved speakers $S11$ and $S12$ from the StethoText corpus for testing. Second, we evaluate the methods in a more challenging intra-gender setting, where the model is trained on data from all female speakers and tested on male speakers, and vice versa. The reported error rates for the converted speech using our proposed method ranged from 27.72% to 33.70%. We attribute the strong generalizability of our method to its ability to avoid generating aligned training pairs and the effectiveness of autoregressive modeling in leveraging large amounts of data.

5. Conclusion

We present an alignment-free, end-to-end framework for NAM-to-Speech conversion. Existing models heavily depend on large per-speaker NAM-text pairs to learn alignments for their Seq2Seq architectures but fail to leverage data augmentation effectively. Moreover, their performance degrades when trained on multi-speaker data—a critical challenge in SSI research, where collecting large datasets per speaker is impractical. Our AR-based Seq2Seq model with multitask training addresses these limitations by utilizing non-aligned NAM-speech pairs and synthetically augmented data. Ablation results demonstrate that multitask training acts as a regularizer and helps the model learn complementary information, leading to notable performance gains in both multi-speaker and zero-shot settings across multiple public datasets. Beyond NAM-to-Speech conversion, similar training strategies can improve other SSI methods like UTI and rtMRI by reducing reliance on error-prone phoneme-speech alignments. This shift can make SSI technologies more affordable and accessible, broadening their practical impact. In future work, we plan to explore techniques for preserving paralinguistic aspects in NAM-to-Speech conversion.

6. References

[1] N. Shah, N. Sahipjohn, V. Tambrahalli, R. Subramanian, and V. Gandhi, "Stethospeech: Speech generation through a clinical stethoscope attached to the skin," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 3, Sep. 2024. [Online]. Available: https://doi.org/10.1145/3678515

[2] M. Janke and L. Diener, "Emg-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[3] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, ser. IUI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 43–53. [Online]. Available: https://doi.org/10.1145/3172944.3172977

[4] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Proc. INTERSPEECH 2023*, 2023, pp. 127–131. [Online]. Available: https://www.isca-archive.org/interspeech_2023/otani23_interspeech.html

[5] H. Li, M. Wang, H. Gao, S. Zhao, G. Li, and Y. Wang, "Hybrid silent speech interface through fusion of electroencephalography and electromyography," in *Interspeech 2023*, 2023, pp. 1184–1188.

[6] R. Beeson and K. Richmond, "Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video," in *Interspeech 2023*, 2023, pp. 1149–1153.

[7] N. Shah, A. Kashyap, S. Karande, and V. Gandhi, "Mri2speech: Speech synthesis from articulatory movements recorded by real-time mri," in IEEE ICASSP 2025, Hyderabad, India. [Online]. Available: https://arxiv.org/abs/2412.18836

[8] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[9] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *2003 IEEE ICASSP.*, vol. 5. IEEE, 2003, pp. V–708.

[10] N. Shah, S. Karande, and V. Gandhi, "Towards improving nam-to-speech synthesis intelligibility using self-supervised speech models," in *Interspeech*, 2024, pp. 2470–2474.

[11] N. Shah, S. Karande, and G. Vineet, "Advancing nam-to-speech conversion with novel methods and the multinam dataset," in IEEE ICASSP 2025, Hyderabad, India. [Online]. Available: https://arxiv.org/abs/2412.18839

[12] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech synthesis from articulatory movements recorded by real-time mri," in *Interspeech 2023*, 2023, pp. 127–131.

[13] Y. Nakajima and K. Shikano, "Methods of fitting a nonaudible murmur microphone for daily use and development of urethane elastmer duplex structure type nonaudible murmur microphone," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3330–3330, 2006.

[14] S. Shimizu, "Acoustic characteristics of non-audible murmur and response characteristics of nam microphone," *IEICE Tech. Rep.*, vol. 2007, pp. 85–90, 2007.

[15] S. Shimizu, M. Otani, and T. Hirahara, "Frequency characteristics of several non-audible murmur (nam) microphones," *Acoustical science and technology*, vol. 30, no. 2, pp. 139–142, 2009.

[16] T. Toda and K. Shikano, "Nam-to-speech conversion with gaussian mixture models." in *Interspeech*, 2005, pp. 1957–1960.

[17] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion." in *Interspeech*, 2018, pp. 3157–3161.

[18] H. Malaviya, J. Shah, M. Patel, J. Munshi, and H. A. Patil, "Mspec-net: Multi-domain speech conversion network," in *IEEE ICASSP*. IEEE, 2020, pp. 7764–7768.

[19] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Interspeech 2017*, 2017, pp. 2625–2629.

[20] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Interspeech 2019*, 2019, pp. 1123–1127.

[21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[22] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech 2021*, 2021, pp. 3615–3619.

[23] N. Shah, S. Kosgi, V. Tambrahalli, N. S, A. Nelakanti, and V. Gandhi, "ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 79–91. [Online]. Available: https://aclanthology.org/2024.findings-eacl.6

[24] J. Rekimoto, "Wesper: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–12.

[25] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[26] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7180–7184.

[27] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[28] A. Baby, A. L. Thomas, N. L. Nishanthi, T. Consortium *et al.*, "Resources for indian languages," in *Proceedings of Text, Speech and Dialogue*, 2016.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[31] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu *et al.*, "Textless speech-to-speech translation on real data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 860–872.

[32] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[34] P. J. Barnes, S. D. Shapiro, and R. A. Pauwels, "Chronic obstructive pulmonary disease: molecular and cellularmechanisms," *European Respiratory Journal*, vol. 22, no. 4, pp. 672–688, 2003.