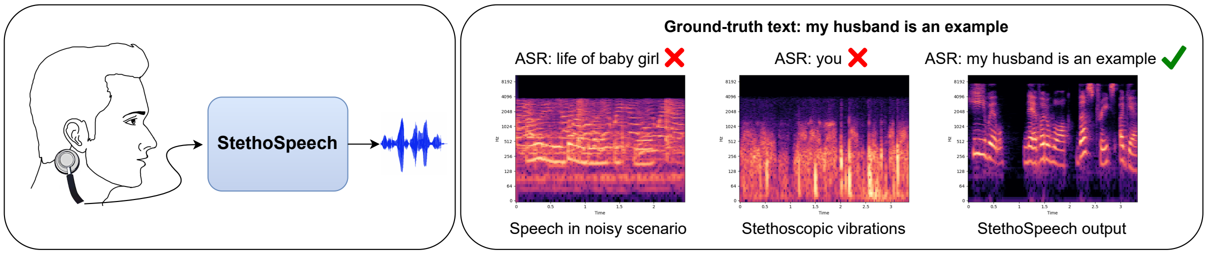
StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached to the Skin

NEIL SHAH, International Institute of Information Technology, Hyderabad, India and TCS Research, India
NEHA SAHIPJOHN, International Institute of Information Technology, Hyderabad, India
VISHAL TAMBRAHALLI, International Institute of Information Technology, Hyderabad, India
RAMANATHAN SUBRAMANIAN, University of Canberra, Australia
VINEET GANDHI, International Institute of Information Technology, Hyderabad, India



(a) (b)

Fig. 1. **Overview:** (a) StethoSpeech converts flesh-conducted vibrations into intelligible speech, and is effective even in noisy conditions (*e.g.*, loud background music). (b) (left to right) Mel-spectrograms of recorded speech under noise, Stethoscopic vibrations, and StethoSpeech-generated speech. Automatic Speech Recognition (ASR) outputs are shown on top. ASR completely fails to comprehend noisy audio and stethoscopic vibrations, but correctly predicts the speech converted via StethoSpeech.

We introduce ***StethoSpeech***, a silent speech interface that transforms flesh-conducted vibrations behind the ear into speech. This innovation is designed to improve social interactions for those with voice disorders, and furthermore enable discreet public communication. Unlike prior efforts, StethoSpeech does not require (a) paired-speech data for recorded vibrations and (b) a specialized device for recording vibrations, as it can work with an off-the-shelf clinical stethoscope. The novelty of our framework lies in the overall design, simulation of the ground-truth speech, and a *sequence-to-sequence translation network*, which works in the latent space. We present comprehensive experiments on the existing CSTR NAM TIMIT Plus corpus and our proposed *StethoText*: a large-scale synchronized database of non-audible murmur and text for speech research. Our results show that StethoSpeech provides natural-sounding and intelligible speech, significantly outperforming existing methods on several quantitative and qualitative metrics. Additionally, we showcase its capacity to extend its application to speakers

Authors' Contact Information: Neil Shah, International Institute of Information Technology, Hyderabad, Hyderabad, Telangana, India and TCS Research, Pune, Maharashtra, India, neilkumar.shah@research.iiit.ac.in, neilkumar.shah@tcs.com; Neha Sahipjohn, International Institute of Information Technology, Hyderabad, Hyderabad, Telangana, India, neha.s@research.iiit.ac.in; Vishal Tambrahalli, International Institute of Information Technology, Hyderabad, Hyderabad, Telangana, India, vishal.tambrahalli@research.iiit.ac.in; Ramanathan Subramanian, University of Canberra, Bruce, Canberra, Australia, ramanathan.subramanian@ieee.org; Vineet Gandhi, International Institute of Information Technology, Hyderabad, Hyderabad, Telangana, India, vgandhi@iiit.ac.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2474-9567/2024/9-ART123
https://doi.org/10.1145/3678515

not encountered during training and its effectiveness in challenging, noisy environments. Speech samples are available at https://stethospeech.github.io/StethoSpeech/.

CCS Concepts: • **Human-centered computing** → **Accessibility**; *Human computer interaction*; • **Computing methodologies** → Machine learning.

Additional Key Words and Phrases: silent speech, self-supervised learning, HuBERT, StethoSpeech, NAM-to-speech conversion, zero-pair setting, artificial learning

ACM Reference Format:
Neil Shah, Neha Sahipjohn, Vishal Tambrahalli, Ramanathan Subramanian, and Vineet Gandhi. 2024. StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached to the Skin. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 123 (September 2024), 21 pages. https://doi.org/10.1145/3678515

1 Introduction

Speech is the most effective and facile means of social interaction and communication. The air exhaled from the lungs cause the vocal folds to vibrate and produce speech due to the articulation of the tongue, cheeks, and lips. Unfortunately, adversarial medical conditions can impede the usual process of speech production. For instance, *Aphonia*, a neurogenic voice disorder [24] resulting in a person's inability to speak or produce audible sounds, occurs due to complete or partial airway obstruction. Several factors like vocal cord paralysis, vocal cord polyps, or psychological factors can cause Aphonia. Although scientific progress in medical science has found potential cures using medicines and surgeries, current solutions come with challenges and side effects [5, 7]. Hence, exploring alternatives through advancements in speech and signal processing is crucial, which has spurred research in Silent Speech Interfaces (SSI).

SSI is a form of spoken communication where an acoustic signal is not produced, *i.e.*, the subject silently articulates without producing sound. SSI techniques comprehend speech content by analyzing silent deformations or vibrations resulting across the vocal folds. SSI provides numerous benefits, including user-friendliness, customization potential, being more comprehensible to the general population, and facilitating independence and immediate communication. Lip reading [48] is one of the most straightforward SSI techniques. Other alternatives include Ultrasound Tongue Imaging (UTI) [53], real-time MRI (rtMRI) [31], Electromagnetic Articulography (EMA) [14], Permanent Magnet Articulography (PMA) [12], Electrophysiology [21], Electrolarynx (EL) [10, 22] and Electropalatography (EPG) [23]. However, many of these techniques are not real-time and pose challenges for everyday use due to their invasive nature. For instance, UTI captures tongue movement using ultrasound, rtMRI records the mid-sagittal plane of the upper airway using MRI, and EMA measures the movement of coils attached to articulators like the lips and tongue. Other limitations include intense vibrator noise, poor lighting scenarios, and extended time requirements for achieving speech proficiency.

Two decades ago, a non-invasive SSI technique was proposed by Nakajima et al. [29], employing a microphone capturing the flesh-conducted vibrations from behind the ear. The authors fabricated a specialized device by implanting a miniature condenser microphone into a standard medical-use stethoscope called a Non-Audible Murmur (NAM) microphone. They demonstrated the feasibility of speech recognition from NAM vibrations using a small Japanese language corpus. Almost a decade later, Yang et al. [58] released the CSTR NAM TIMIT Plus corpus, a 40-minute paired corpus of NAM and whispered speech in English. This corpus led to a few efforts translating NAM vibrations into conventional speech [27, 44].

The following shortcomings constrain most of the present initiatives. Firstly, State-Of-The-Art (SOTA) approaches utilize custom fabricated devices, which hinders their wide-scale applicability. Secondly, most existing efforts assume that auxiliary data in the form of a paired NAM-whisper or NAM-speech corpus are available. In numerous situations, such a requirement may be excessive (*e.g.*, for a patient who cannot utter everyday speech or has difficulty whispering). Thirdly, most recent methods [27, 44] map Mel-based spectral features from NAM vibrations directly to speech features. Such models tend to predict speaker and ambient noise characteristics in the ground-truth Mels. This further affects the quality of the output speech, in addition to existing challenges in NAM vibrations like the absence of a fundamental frequency [29] and heavy attenuations in the high-frequency components [50]. Finally, due to the lack of multi-speaker NAM datasets, the current research is yet to explore real-time adaptability and generalizability of existing approaches to novel speakers, whose data was not used for training (known as the *zero-shot* setting).

To address these issues, we propose *StethoSpeech* (illustrated in Figure 1), a novel machine learning-based framework for NAM-to-speech conversion in real-time and zero-shot scenarios. StethoSpeech differs from SOTA as follows:

- StethoSpeech takes input from an off-the-shelf and affordable wireless medical stethoscope used by general medical practitioners. This opens up possibilities for wide-scale adoption of the NAM-to-speech conversion technology.
- Our framework does not require auxiliary information in the form of paired speech or whisper data. We only need recorded NAM vibrations corresponding to a given (ground-truth) text and propose a novel way to simulate ground-truth speech by aligning textual data with corresponding NAM vibrations. This hugely improves accessibility and customization, as it opens up avenues to train subject-specific models, even for subjects unable to produce normal speech or whispering sounds.
- Prior studies on NAM-to-speech conversion limit their experiments to a single subject [58]. In contrast, we perform multi-subject experiments and demonstrate how StethoSpeech generalizes to novel speakers not encountered during training.
- Our framework exploits content-rich Self-Supervised Learning (SSL) representations, ignoring background and speaker-specific characteristics. We obtain SSL embeddings for NAM vibrations and translate them to SSL embeddings of the simulated speech data. We show that such a Sequence-to-Sequence (Seq2Seq) mapping makes it possible to learn efficient embeddings that preserve content, improve intelligibility, and restore prosody using as little as 35 minutes of paired text-NAM data.
- The obtained SSL speech embeddings are then converted to speech using a frozen speech decoder. The output speech can be rendered via pre-observed voices during initial speech-to-speech training. This speaker-conditioned framework allows for flexibility in voice selection. The speech decoder can also be augmented for novel voices with $15 - 20$ minutes of recorded speech for novel subjects.
- Our system design is minimalistic, where the NAM vibrations from the wireless stethoscope are transmitted onto the mobile phone via Bluetooth, and speech is output through the phone speaker. StethoSpeech achieves real-time performance for inference, and translating a 10s NAM vibrations takes less than 0.3s.

To evaluate our approach, we conduct experiments on the in-house compiled *StethoText* corpus from twelve speakers, plus the existing CSTR NAM TIMIT Plus dataset [58]. We summarize below our research contributions:

- We propose the novel StethoSpeech framework for generating intelligible speech using an off-the-shelf stethoscope attached to the skin behind the ear. Our novelty lies in (a) simulating paired speech from NAM and text via noisy alignment, and (b) the methodology for NAM-to-speech translation via SSL embeddings.
- We release the StethoText corpus, comprising over fifteen hours of NAM vibrations and corresponding text. This dataset includes NAM samples collected under noise (*e.g.*, slightly noisy everyday office environment, and high-noise scenarios such as deafening background music).
- We present comprehensive experiments on the CSTR NAM TIMIT Plus and the proposed StethoText datasets. Our method outperforms SOTA methods without employing paired speech, providing more intelligible, natural-sounding speech with consistent prosody. We show that StethoSpeech is usable even

under extreme noise (Figure 1(b)), where normal speech is unintelligible. Furthermore, the framework demonstrates robustness to user mobility, such as walking.

2 Related Work

2.1 Speech Conversion

Our study is related to a wider body of research on speech conversion. It shares a common objective with voice cloning, where speech from one person is translated to a different voice [1, 34]. An alternative problem involves modifying accents within a multilingual context [43]. Another closely related application is whisper-to-normal speech conversion [20, 33, 38, 52]. These papers employ algorithms such as cycle Generative Adversarial Networks (GANs) [33], Artificial Neural Networks [20], and self-supervised representation learning for whisper talk [38] for their tasks. Most of these approaches assume paired whisper-to-speech data is available, and the whisper is noiselessly recorded in a studio setup. Recent work by Jun Rekimoto [38] shows that speech data can be mechanically converted to whispered voice using a linear predictive coding-based audio conversion tool [60]. Another salient direction in speech conversion research is silent lip reading, where translation happens between the visual and text modalities [9, 35, 41].

Our work of translating stethoscopic vibrations into everyday speech is related to speech conversion efforts. However, our task involves several new challenges, with the primary challenge being data scarcity. Unlike lip reading or whisper-to-speech conversion, abundant paired data is unavailable for this task. Collecting paired vibration-speech data is impossible in several scenarios (*e.g.*, medical patients). Secondly, unlike other methods [34], a pre-trained self-supervised backbone is unavailable for vibration data. The third challenge is posed by the nature of vibrational data like the absence of fundamental frequencies [50], difficulty in accurately mapping their corresponding speech acoustic characteristics, and incorrect detection of vowel and consonant sounds in the vibrations [29]. Further complications arise from factors like low-pass characteristics of soft tissues, which attenuates high-frequency components [50], and the lack of a closed-form expression for radiation impedance [36, Chapter 4, pp. 130].

2.2 Speech Signal Representations

State-of-the-art methods in the applications above heavily rely on self-supervised learning, wherein information extracted from the input audio serves as a label for learning representations in downstream tasks. Self-supervised models fall into categories like generative, contrastive, or predictive. Generative models, such as [8, 54], reconstruct input or predict future inputs, facing challenges with speech signals containing more information than text. Contrastive models, such as [3, 42], distinguish positive from negative samples but encounter issues with invariances in representations when exclusively sampling from the same utterance. Recent predictive models, like Discrete BERT [2], HuBERT [17], and WavLM [6], use additional loss functions to prevent information loss due to input quantization and outperform other approaches, as demonstrated in [6, 9, 34, 38]

Due to the availability of paired text, we also utilize linguistic information in our StethoSpeech framework. Specifically, we simulate ground-truth speech derived from a Text-to-Speech (TTS) system. More recent TTS involves a Transformer-based acoustic model [39, 40] that directly generates self-supervised representations from text, while a GAN-based decoder [25] converts these representations into speech. These models are mostly Non-AutoRegressive (NAR), with constant computation (O(1)) during both training and inference. As a result, real-time SSI techniques [38] prefer variants of *FastSpeech 2*-based models to attain quick, accurate and controllable TTS.

2.3 Vibrational or motion-based speech translation

Previous research has examined converting speech with novel device fabrication. *V-Speech* system proposed by Maruri et al. [28], leverage a piezoelectric disk located above the nasal pads of smart glasses. However,

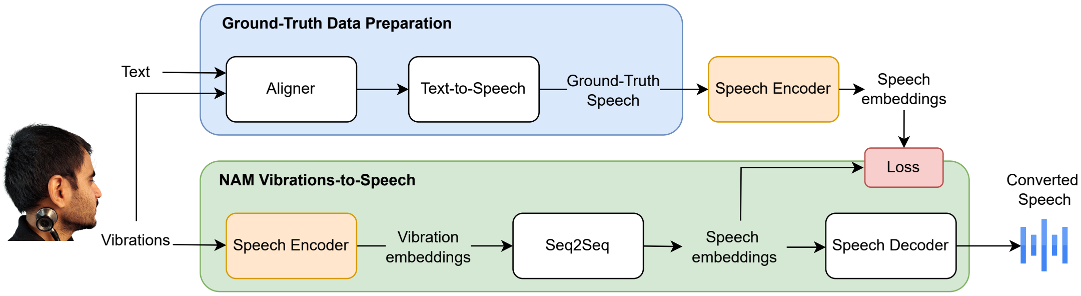


Fig. 2. StethoSpeech is a speech conversion mechanism from NAM vibrations. It comprises a data preparation step to generate ground-truth speech corresponding to NAM vibrations, a shared speech encoder (pre-trained and frozen) to extract self-supervised embeddings, a sequence-to-sequence network to map self-supervised embeddings of vibrations to that of speech, and a speech decoder to synthesize speech from the self-supervised speech embeddings.

they require clean speech for training and the performance remains unseen on novel speakers. *AlterEgo* [21] represents a wearable interface composed of plastic and gold-plated silver electrodes, designed to detect signal signatures across seven facial regions, However, its evaluation is restricted to a narrow vocabulary of spelling ten digits. *RFTattoo* [57] utilizes batteryless and flexible RFID tattoos on the face to track lip-stretch and tongue position. Zeng et al. [59] use radar-based wireless sensing to capture articulatory features, to communicate and recognize speech up to a specific range. However, occlusions and background movements can negatively impact performance of this approach, which also falters on words unseen during training.

Our work closely relates with the seminal work by Nakajima et al. [29], which demonstrated the feasibility of generating speech from NAM. The NAM recording device is fabricated by implanting a miniature condenser microphone into a standard medical-use stethoscope. They chose the microphone configuration to ensure it effectively captured NAM while avoiding external noise. Subsequently, they improve their devices via better design and usability [30]. Significant efforts were made by Hirahara et al. [16], Shimizu et al. [46] in a similar direction, as they studied frequency characteristics and sensitivities of five different NAM microphone designs. Our work does not concentrate on device fabrication but instead explores the direct utilization of readily available clinical stethoscopes.

Several methods exist to translate the signal captured via the specialized NAM microphones into intelligible speech. Hidden Markov Models (HMMs) for NAM recognition (*i.e.*, NAM-to-text conversion) were employed by Nakajima et al. [29]. Heracleous et al. [15] improved HMM models through speaker adaptation. An initial attempt to obtain direct speech from NAM vibrations was made in Toda and Shikano [51], which found that the converted speech lacks prosody and intelligibility due to difficulties in estimating pitch contours from NAM vibrations. The authors instead suggested converting whispers from NAMs to address this issue and improve intelligibility and naturalness. Since then, NAM-to-whisper speech conversion has been studied and replaced by neural variants in recent years [27, 44]. GANs are employed in [44], while [27] uses multiple auto-encoders aligned in the latent space. Overall, paired NAM and whisper data requirements are essential to prior approaches. [27] demonstrates the experiments on a limited 40 minutes of data captured from a specialized device [58]. In contrast, our work eliminates the necessity for paired-whisper or paired-speech data and achieves speech translation utilizing a standard clinical stethoscope.

3 Method

Prior NAM-to-speech approaches [27, 51] assume that access to paired speech or whisper audio corresponding to the NAM vibrations is available. These methods train their conversion models by initially leveraging studio-recorded speech and then applying Dynamic Time Warping (DTW) [27] to align speech with NAM vibrations in the Mel feature space. We note two concerns with the above approaches: (1) explicitly recording ground-truth speech may not be practical or feasible in certain situations, especially when dealing with patients who cannot produce speech. (2) the DTW alignment warps artifacts in the converted speech signal, reducing its quality and intelligibility. Instead, we propose the novel StethoSpeech method, which does not require paired speech data.

Figure 2 overviews the StethoSpeech framework, which consists of four modules: a ground-truth preparation module, a speech encoder, a Seq2Seq network, and a speech decoder. The ground-truth data preparation step represents our point of difference from prior work. It simulates ground-truth speech corresponding to NAM vibrations from the corresponding text in lieu of whisper or speech data. The shared speech encoder converts NAM vibrations plus ground-truth speech into corresponding SSL embeddings. The Seq2Seq network then inputs embeddings derived from vibrations and maps them to corresponding speech embeddings. Finally, the speech decoder utilizes these converted embeddings to reconstruct the output speech. We below explain each module separately.

3.1 Ground-truth Data Preparation

We propose a novel approach to simulate ground-truth speech solely by utilizing NAM vibrations and their corresponding textual content. Given the availability of textual data, we can employ a *text-to-speech* synthesis model to simulate the corresponding ground-truth speech. However, a significant challenge arises due to the potential temporal misalignment of the NAM vibrations and the simulated speech. Unaligned data makes it challenging to train a Seq2Seq network in a NAR manner, often requiring perfectly synchronized sequence pairs. Autoregressive models do not have this restriction; however, they face real-time challenges by their inference speeds and issues like word skipping/repetitions [39].

Some NAR TTS models [39, 43] allow for explicitly controlling each rendered phoneme's duration. Hence, we can simulate ground-truth speech by obtaining phoneme-level alignment between the NAM vibrations and ground-truth text. We could use a forced alignment tool to extract the duration of every phoneme in the NAM vibrations; however, this will internally require the acoustic model to detect phonemes and align them with the text-phonemicized tokens. Off-the-shelf acoustic models are unusable on NAM vibrations due to the differing characteristics of the vibrational data, which are predominantly inaudible, lack intelligibility, and do not convey any discernible prosodic patterns. Therefore, we train an acoustic model using NAM vibrations and corresponding textual representations to obtain phoneme-level alignment. We rely on the official Montreal Forced Aligner (MFA) repository [**?**] to train an acoustic model. The model first extracts 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features from the vibrations. These features undergo 40 iterations of monophone Gaussian Mixture Model (GMM) training and 35 iterations of triphone GMM training to capture contextual phoneme information. They learn the acoustic feature transformation using maximum likelihood linear regression to obtain phoneme-level durations. Subsequently, we force-feed these durations to a FastSpeech2 [39] TTS model with the corresponding text, enabling the synthesis of ground-truth speech aligned with the input NAM vibrations.

3.2 Speech encoder

SOTA NAM-to-speech conversion techniques [27] rely on Mel-cepstral features to encode the raw audio. However, such features encode the entirety of the input audio, including ambient noise characteristics. So, when training NAM-to-speech conversion models, the network is also prone to reconstructing the ambient noise information. This makes training complex and adversely impacts the intelligibility and quality of the converted speech.

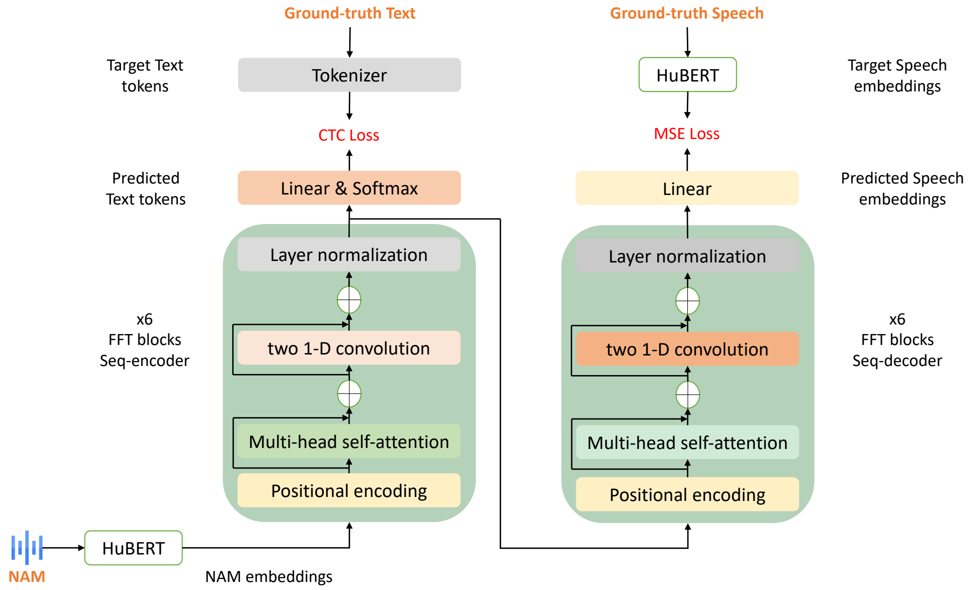


Fig. 3. **Seq2Seq architecture overview:** SSL embeddings of NAM vibrations traverse six transformer encoder blocks, followed by prediction of text tokens using a linear head and a Softmax layer optimized with CTC loss against ground-truth text tokens. Subsequently, the encoded NAM embeddings are transformed into target speech embeddings through six transformer decoder blocks and a linear layer utilizing MSE loss.

Recent advances in self-supervised learning achieve high-fidelity, compressed speech representations. Our work primarily focuses on content-rich, disentangled representations that ignore the speaker and ambient information. To this end, we employ BASE HuBERT [17], a self-supervised neural network trained on 960 hours of unlabeled speech using a BERT-like[8] masked-prediction loss. This method partitions the audio into small chunks, computes their features, and clusters them via the k-means algorithm. The network uses the cluster IDs as pseudo labels and then performs classification to predict pseudo labels for randomly masked segments. The cluster ids are iteratively computed to obtain more suitable self-supervised units.

3.3 Seq2Seq network

Following recent advancements in unit-based Seq2Seq translation networks [4, 18, 26, 34], we initially attempted speech translation at the discrete unit levels output by the speech encoder. However, the decoding resulted in poor-quality synthesized speech, indicating a significant loss of information when NAM vibrations are encoded and translated as discrete units. Deviating from most existing unit-based Seq2Seq architectures, we propose leveraging *latent space embeddings* to this end. We prefer this approach as encoding NAM vibrations via embeddings improved the synthesized speech's intelligibility. The shared speech encoder is employed to extract 768-dimensional latent SSL embeddings from the last hidden layer of the pre-trained HuBERT for both NAM vibrations and simulated ground-truth speech. Our novel transformer-based Seq2Seq network then learns the mapping between these embeddings. To further enhance real-time efficiency, we opt for a NAR encoder-decoder network. This architecture predicts output embeddings in a single iteration, providing a significant latency advantage over autoregressive models, which generate output embeddings one at a time based on past embeddings.

The 6-layer Seq-encoder and Seq-decoder networks contain feed-forward transformer blocks with two multi-head self-attention [55] blocks and two 1-dimensional convolutions inspired by Fastspeech2 [39]. The Seq-encoder processes embeddings from the NAM vibrations into a sequence of fixed-dimensional vectors, while the Seq-decoder predicts ground-truth speech embeddings. Our approach does not rely on variance adaptor [39] or length regulator [40], as input and target embeddings are aligned through our proposed ground-truth data preparation step. We set the batch size to 16 and the maximum number of steps to $30,000$. We employ the Adam optimizer with an initial learning rate of 4.4 x 10^{-2} and an annealing rate of 0.3, with annealing steps at $[3000, 4000, 5000]$. The HuBERT model encodes speech and vibrations into embeddings at a frame rate of 50Hz. The model optimizes the Mean Squared Error (MSE) loss, quantifying the difference between the decoded and ground-truth speech embeddings. We define the objective as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{i=1}^{T} ||S_{\text{ssl\_i}} - \hat{S}_{\text{ssl\_i}}||^2, \tag{1}$$

where S_{ssl} are the ground-truth speech embeddings, $\hat{S}_{\text{ssl}}$ are the decoded speech embeddings, and i indexes the time-steps ranging from $1 \dots T$.

We augment the model with an additional fully connected linear head to predict Connectionist Temporal Classification (CTC) [13] tokens after the Seq-encoder layer. For tokenizing text sequences, we utilized the Wav2Vec2 tokenizer [3] and followed the text preprocessing steps outlined in [56]. For given input NAM embeddings N_{ssl}, let Enc_{ssl} be the output of the Seq-encoder. Suppose C denotes the character label corresponding to the ground-truth text; the goal is to minimize the negative log-likelihood by using $P_{CTC}(C|Enc_{\text{ssl}})$ defined as :

$$\mathcal{L}_{CTC} := -\log P_{CTC}(C|Enc_{\text{ssl}}). \tag{2}$$

Taking the weighted sum of the MSE and CTC loss functions, the final objective function is summarized as:

$$\mathcal{L}_{\text{Tot}} = \alpha_{CTC} * \mathcal{L}_{\text{CTC}} + \alpha_{MSE} * \mathcal{L}_{\text{MSE}}, \tag{3}$$

where $\alpha_{\text{CTC}} \in \mathbb{R}$ and $\alpha_{\text{MSE}} \in \mathbb{R}$ are the hyper-parameters that balance the influence of the two loss terms. We empirically found optimal values for α_{CTC} and α_{MSE} to be 0.001 and 1, respectively.

3.4 Speech decoder

To generate speech, we need a decoder that takes the speech embeddings predicted by a Seq2Seq network as input and synthesizes speech as the output. We rely on HiFiGAN-v2 [25, 34] for speech synthesis from these self-supervised embeddings. However, in its basic form, we have identified that HiFiGAN-v2 [34] accepts discrete units as input, while our Seq2Seq model predicts embeddings. To address this, we employ a two-step process. First, we train a k-means model on English voices from the datasets [19, 47] using the Fairseq repository [32]. During inference, we deploy the trained k-means model to predict cluster units. These cluster units then serve as discrete inputs to the HiFiGAN speech decoder, allowing us to utilize both models' strengths for speech generation effectively.

The generator in the decoder employs transposed convolutions for upsampling the ground-truth speech's SSL units and a residual block for receptive field expansion, generating the synthesized signal. The discriminator distinguishes between the synthesized and original signals, using multi-period and multi-scale networks to capture temporal patterns, details, and global structure. The main objective is to minimize the dissimilarity between the two signals, enhancing speech fidelity. We train a multi-speaker speech decoder with the voices available from

Table 1. Description of the StethoText dataset. Gender Labels (M: Male, F: Female) and the duration shown are in minutes.

Speaker	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
Gender	M	F	M	M	M	M	M	F	M	M	F	M
Files	3325	3961	536	450	572	535	318	203	536	600	475	315
Duration	263.11	251.28	36.74	38.96	60.72	49.13	27.31	16.29	39.58	52.65	49.48	25.36

[19, 47], allowing the user to generate content in the voice of their interest. For the model configuration, we set the batch size to 16, the learning rate to 2×10^{-4}, the number of embeddings to 100, the embedding dimension to 128, and the model input dimension to 256.

3.5 Model size and computational requirements

The stethoscope pairs wirelessly with a smartphone via Bluetooth, leveraging a dedicated application, which instantaneously transmits the recordings to our AWS server (c5.xlarge instance with 4 vCPUs, 8.0 GiB of memory and priced at 0.17 USD per hour). The processing happens on the cloud, and we assess computational performance by measuring the time taken to convert speech for a 10s segment of recorded vibration. Our HuBERT model (94.69M trainable parameters) performs feature extraction in 0.368s, and the conversion of these features to speech embeddings through a Seq2Seq model (33.78M trainable parameters) consumes 0.173 seconds. The Hifi-GAN model (13.75M trainable parameters) generates output speech in 1.203 seconds. Overall, with CPU-only inference, our pipeline takes approximately 1.754 seconds. The end-to-end processing time sharply reduces to 0.305 seconds when performing inference on a g4ad.xlarge instance with a single AMD Radeon Pro v520 GPU with 16 GiB memory, the lowest cost GPU in AWS priced at 0.37 USD per hour.

4 Experiments

4.1 Datasets

4.1.1 CSTR NAM TIMIT Plus corpus. The CSTR NAM TIMIT Plus corpus [58] is a public dataset comprising NAM vibrations plus their corresponding text and whisper audio. It comprises 421 sentences uttered by a female speaker from the *Herald text* in a studio setup. The dataset spans a duration of 40 minutes and contains 1601 unique words. The audio sampling frequency is 16000 Hz. In alignment with previous works [27, 44], we randomly assign 13% of the data for the test set and the remaining portion for training. Our StethoSpeech framework ignores the whisper audio while only being used for baseline comparisons.

4.1.2 StethoText corpus. We detail various attributes of the StethoText corpus below.

Subjects: We focused on four key aspects to evaluate the speech conversion ability of the StethoSpeech framework. The ***speaker-specific*** study assesses the viability of our framework to produce highly intelligible speech by training on extensive samples from a single speaker. To this end, we compiled the StethoText corpus with approximately 8 hours of data from two speakers– one male (s1) and one female (s2). Again, we employ a random selection method for performance evaluation, with 10% of the speaker data used for testing and the remaining portion for training. For a second study, aiming to showcase the efficacy of StethoSpeech under ***zero-shot learning***, we opted to gather NAM vibration data from ten speakers (s3 to s12), including two females and eight males. These speakers narrated 4540 English sentences, amounting to 7 hours of data. Here, we explicitly reserved speakers s11 (female) and s12 (male) for testing while training with other speakers to assess StethoSpeech in a zero-shot scenario. A third study showcases the robustness of StethoSpeech under ***background noise*** involved recording an additional 10 minutes of NAM vibrations from speaker s2 while playing a high-bass

Table 2. Tokenized distribution of unique words in the evaluated corpora.

Tokens	1	2	3	4	5	6	7
% of words in CSTR NAM TIMIT Plus	39.04	26.80	2.00	0.19	0.12	0.00	0.00
% of words in StethoText	26.72	25.80	4.69	1.38	0.37	0.03	0.01

musical piece in the background. The fourth study investigates the robustness of the StethoSpeech framework to ***mobility*** by recording 10 minutes of NAM vibrations from subject s1 while walking. All speakers were paid a token fee for their time and effort. Our StethoText corpus comprises 15.78 hours of recorded NAM vibrations plus corresponding text. Table 1 presents StethoText details regarding gender distribution, the number of recorded files per speaker, and the recorded duration in minutes.

NAM vibrations acquisition: We used an *AyuSynk*[1], a Bluetooth-enabled digital stethoscope, primarily due to its affordability at 185 USD. The stethoscope has denoising filters, which are particularly useful for recording heartbeats and lung movements. We recorded NAM vibrations directly without denoising to preserve the original signal. The participants were instructed to place the stethoscope-head below the back of the ears, on the tissue region, loosely adhering to guidelines from the prior efforts [29, 30]. Although we did not formally experiment with alternative positions, our findings indicate that minor positional variations did not significantly impact the recording quality.

Speakers aged between 26 and 30 individually narrated English sentences sourced from the TIMIT [11] dataset and sentences published in [47]. Speakers were instructed to narrate the sentences in low quiet voice (at most, producing a soft whisper, not meant to be heard by others) while holding the stethoscope behind their ear. Each speaker was given non-overlapping text content to narrate and each of them contributes unique sentences to the dataset. None of the participants displayed abnormalities in hearing or speech development, nor did they manifest any pathologies. Each participant delivered sentences at their natural speaking pace in a slightly noisy environment, mimicking a regular office setting. All participants were naive to the purpose of their study. An off-the-shelf wireless clinical stethoscope captures NAM vibrations. We noticed that involuntary human movements, such as gestures and facial expressions during the recording process, did not hamper the quality of NAM vibrations and the overall speech conversion process. No noise suppression nor signal enhancement techniques were employed during the recording process, ensuring the fidelity of the captured NAM vibrations. We did not record any paired speech and instead relied on our proposed data generation module to simulate ground-truth speech.

Dataset diversity: Via StethoSpeech, we aim to create a dataset that includes NAM vibrations and aligned text, closely simulating real-world speech conditions while incorporating diverse and challenging words. The existing CSTR NAM TIMIT Plus corpus consists of sentences from the original TIMIT dataset [11]. While these sentences are phonetically balanced, they do not capture the broad vocabulary used in everyday speech. The twelve subjects in the StethoText corpus contributed to recording a total of $17,747$ unique words, compared to the $1,601$ unique words narrated by a single speaker in CSTR NAM TIMIT Plus. To assess the overall diversity of our corpus, we subjected all sentences from both corpora to a BERT-based-uncased tokenizer [8]. We analyzed the number of tokens per word from both corpora, which serves as a proxy for measuring the diversity in the dataset. For instance, the word "Is" is tokenized into a single token, while a word like "Tenaliraman" is tokenized into four tokens ("ten","ali","rama","n"). Table 2 shows the comparison. The StethoText corpus exhibits a higher prevalence of diverse and challenging words. Approximately 6.48% of words in StethoText require three or more

[1] https://www.ayusynk.ai

tokens, compared to 2.31% in the CSTR NAM TIMIT Plus corpus. This observation reveals that the StethoText corpus incorporates diverse words.

Ethical Disclosure: In alignment with institutional ethical regulations, the authors secured ethical approval for the study, plus informed consent from participants for the public release of their vibration data. A copy of the consent form is uploaded in the supplementary material. Prior to participation, subjects were thoroughly briefed regarding the task procedures, and their rights. Participant confidentiality is ensured by removing identifiable information during data storage and analysis.

4.2 Experimental setup

We first present a quantitative evaluation of the simulated ground-truth speech and then delve into the performance of StethoSpeech on both the CSTR NAM TIMIT Plus and our proposed StethoText corpora. On each dataset, we perform three types of evaluation: (a) Qualitative assessment via Mel-spectrograms; (b) Quantitative evaluation through Word Error Rates (WER), Character Error Rates (CER), and Mel-Cepstral Distortion (MCD); (c) Subjective user study to evaluate the intelligibility, quality, and naturalness of the output speech. We also perform quantitative and qualitative experiments to test StethoSpeech's generalizability to novel speakers in the zero-shot setting and evaluate the framework's robustness under severe background noise.

Evaluation Metrics: We use Whisper-ASR [37] to transcribe the synthesized speech and calculate the error rates. WER measures the percentage of words that are incorrectly transcribed, while CER measures the percentage of incorrectly transcribed characters. MCD captures the Mel-cepstral distortions between the generated and the ground-truth utterances [49]. For the user study, we recruited 20 participants following informed consent, ensuring gender balance. All subjects, fluent in English and aged 21-35, rated utterances on a 5-point Mean Opinion Score (MOS) scale.

Comparisons: On the CSTR NAM TIMIT Plus corpus, we compare StethoSpeech against SOTA methods MSpec-Net [27] and DiscoGAN [27, 45]. Additionally, we propose a new baseline method for comparison, which utilizes paired whisper audio accessible within the CSTR NAM TIMIT Plus corpus to simulate ground-truth speech. We obtain quantized HuBERT representations of the whisper data and pass it on to the speech decoder to simulate corresponding ground-truth speech. The idea works similarly to conversion from a whispering style to everyday speech. On obtaining ground-truth speech from a whisper, we follow the same steps as StethoSpeech to train our Seq2Seq method for NAM-to-speech conversion. Since it utilizes paired data, we term this baseline "Paired-Baseline".

5 Results

5.1 Simulated ground-truth speech

As the simulation of the ground-truth speech by aligning text with NAM vibrations represents the major point-of-difference of the StethoSpeech framework from prior work [27, 51], we firstly evaluate simulated speech quality on the CSTR NAM TIMIT Plus and StethoText datasets.

CSTR NAM TIMIT Plus: As this corpus includes both text and paired whisper audio along with NAM vibrations, we employ our proposed alignment method (see Section 3.1) on text, plus ground-truth speech simulation from whisper for the paired-Baseline approach. The simulated speech via the paired-baseline yields a 24.73% WER, while StethoSpeech-based speech generation results in a 4.53% WER, conveying that NAM-text alignment is more effective than speech generation from whisper. The higher WER with simulated speech in the paired-baseline

Table 3. Simulated ground-truth speech error rates for speakers in the StethoText corpus using the StethoSpeech approach.

Speaker	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
CER	4.90	4.66	25.66	31.07	21.72	24.35	11.35	15.12	13.74	10.95	11.39	6.91
WER	9.50	9.30	42.50	47.09	32.67	34.51	17.33	27.08	23.18	17.27	18.17	14.76

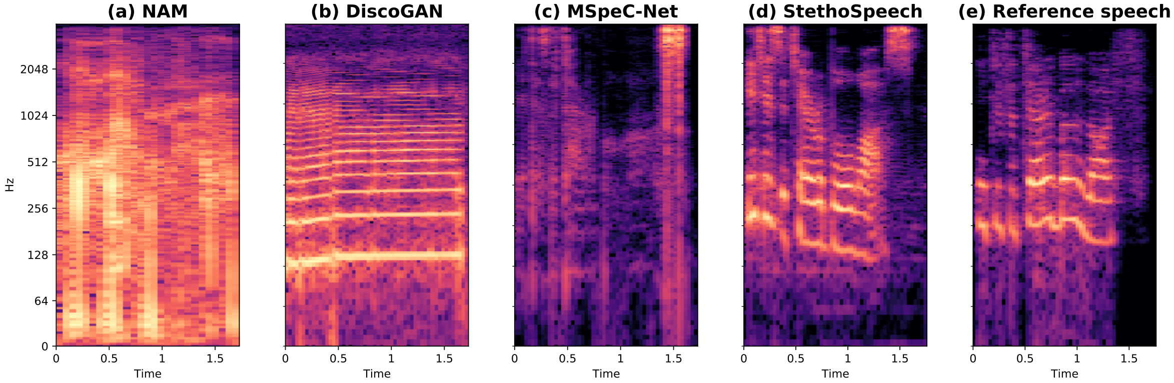


Fig. 4. Mel spectrograms of (a) NAM vibrations and converted speech using (b) DiscoGAN, (c) MSpec-Net, (d) StethoSpeech. (e) presents the spectrogram corresponding to the ground-truth utterance. Sentence is borrowed from the CSTR NAM TIMIT Plus corpus (sentenceID: 401, Transcription: 'It is a terrible loss.')

can be attributed to error propagation from the speech encoder (HuBERT) not pre-trained on whisper audio, followed by a speech decoder.

StethoText: Since our corpus lacks paired whisper audio, we rely on our NAM-text alignment method to simulate ground-truth speech. Table 3 presents error rates for the twelve speakers in the StethoText corpus. We can observe that having more samples for a specific speaker (s1 and s2) aids the alignment module to simulate highly intelligible ground-truth speech. We also observed higher error rates for faster speakers.

5.2 Output Speech Quality Evaluation on CSTR NAM TIMIT Plus

5.2.1 Qualitative Mel assesment. Figure 4 compares Mel-spectrograms of NAM vibrations with speech outputs generated via StethoSpeech and competing state-of-the-art methods. The Mel-spectrogram of the reference speech with the same textual content is shown in Figure 4 (e). These outputs enable insights into two key questions: (1) Why do NAM vibrations lack intelligibility? (2) How similar is the Mel-spectrogram of the StethoSpeech-generated output to the reference, versus SOTA methods? Figure 4 (a) shows that NAM vibrations typically encompass frequencies up to 1-2 kHz while excluding high-frequency components. Additionally, the amplitude of NAM vibrations decreases at higher frequencies, possibly due to the cumulative effects of sound propagation from its source to the skin, along with inherent energy attenuation as frequency increases. These characteristics result in reduced intelligibility and the absence of prosody in NAM vibrations.

Figures 4 (b), (c), and (d) depict Mel-spectrograms of the generated speech samples via the DiscoGAN, MSpec-Net, and our proposed StethoSpeech methods. The DiscoGAN fails to generate understandable speech, as evident from its spectrogram visualization. Although the MSpec-Net output contains some understandable content, it

Table 4. Comparison of metrics on NAM vibrations and Converted Speech using existing methods and our proposed Paired-Baseline and StethoSpeech approach on the **CSTR NAM TIMIT Plus** dataset.

Method	**CER ↓**	**WER ↓**	**MCD ↓**
NAM vibration	48.37	66.99	-
DiscoGAN [27]	76.06	98.87	9.37
MSpec-Net [27]	149.53	148.59	9.14
Paired-Baseline (ours)	27.38	43.79	**4.36**
StethoSpeech (ours)	**16.32**	**26.80**	4.58

struggles to generate natural-sounding speech. Furthermore, MSpec-Net does not adequately retain the lower-frequency components present in the original NAM vibrations. Interestingly, the similarity in Mel-spectra between the StethoSpeech output (Fig. 4 (d)) and reference speech (Fig. 4 (e)) signifies the efficacy of our approach. StethoSpeech enhances lower-frequency formants while also accurately predicting the higher-frequency formants. This dual capability of enhancing and predicting formants at both lower and higher frequencies significantly improves intelligibility. Consequently, StethoSpeech establishes a new benchmark in NAM-to-speech synthesis.

5.2.2 Quantitative evaluation. Quantitative results comparing StethoSpeech with existing methods on the CSTR NAM Timit Plus dataset are presented in Table 4. We report the metrics for the NAM vibrations and the speech generated using StethoSpeech and existing methods on the test split, as outlined in Section 4.1. We directly employ the open-source codes released by [27] for training the DiscoGAN and MSpec-Net methods and report the metrics on our test split. For a fair comparison, we used our simulated ground-truth speech to train and test these methods.

- *NAM vibrations vs. StethoSpeech-converted speech:* Applying ASR on the raw NAM vibrations gives CER = 48.37% and WER = 66.99%. It reflects that the output from the specialized NAM microphone device is partially comprehensible since about 50% of the characters are correctly recognized. Applying StethoSpeech on data recorded using a standard wireless stethoscope brings significant improvements to speech intelligibility (CER = 16.32%, WER = 26.80%).
- *SOTA vs. converted speech from StethoSpeech:* StethoSpeech demonstrates superior performance with the lowest error rates and the least presence of distortion among SOTA results, as measured by the respective metric. It achieves a significant reduction of 99.56% and 104.58% in MCD scores when compared to the MSpec-Net and DiscoGAN methods, respectively.
- *SOTA vs. Paired-Baseline:* Our paired-baseline model achieves a remarkable reduction of almost 50% in MCD compared to SOTA methods, which also rely on paired whisper-speech data for NAM-to-speech synthesis. This observation reveals the efficacy of our proposed Seq2seq architecture, comprising the encoder and decoder blocks.

Although Paired-Baseline and StethoSpeech approaches achieve optimal results in terms of error rates and MCD, it is essential to acknowledge that MCD is not an appropriate metric for comparison, as it relies on the nature of ground-truth speech (background noise, prosody *etc.*). For instance, if the output is denoised clean speech, however, the ground truth is noisy, the MCD will show an increase. We also conducted an ablation study by excluding the CTC loss from Equation 3. We observed that the WER increased from 26.80% to 29.87%, and the CER increased from 16.32% to 19.57%. These results underscore the effectiveness of incorporating the CTC loss.

5.2.3 User study. Figure 5 depicts the outcomes of the subjective evaluation performed by 20 users evaluating the converted speech outputs on the CSTR NAM TIMIT Plus dataset. To assess output speech quality, each participant rates five samples: NAM vibrations, and converted speech via the MSpec-Net, DiscoGAN, Paired-Baseline and

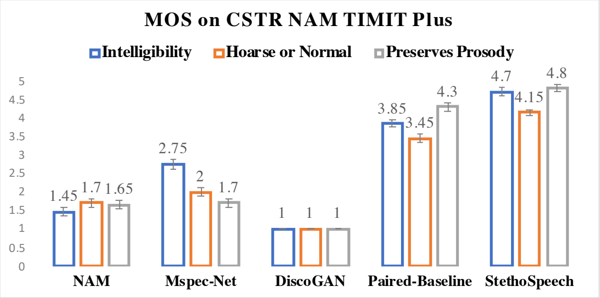


Fig. 5. Quality of NAM-to-speech conversion on **CSTR NAM TIMIT Plus** corpus: MOS evaluated on Intelligibility, Hoarse-normal voice rating and natural prosody rating for original NAM, and converted speech using MSpec-Net, Disco-GAN, Paired-Baseline, and StethoSpeech

Table 5. Recognition performance of NAM vibrations and converted speech using the StethoSpeech approach in different voices, evaluated for speakers s1 and s2 in the **StethoText** corpus.

Speaker	Method	Signal	CER ↓	WER ↓
s1 (male)		NAM vibrations	116.03	141.87
	StethoSpeech	Speech in Voice1 (male)	**11.89**	**20.78**
	StethoSpeech	Speech in Voice2 (female)	15.85	26.76
	StethoSpeech	Speech in Voice3 (female)	13.08	22.77
s2 (female)		NAM vibrations	100.57	112.93
	StethoSpeech	Speech in Voice1 (male)	16.82	27.12
	StethoSpeech	Speech in Voice2 (female)	15.95	27.14
	StethoSpeech	Speech in Voice3 (female)	**15.33**	**26.85**

StethoSpeech methods. To ensure a fair comparison with DiscoGAN and MSpec-Net outputs, we directly utilized the speech samples provided on their demo page [27], and inferred the same set of samples using our approach, as done in [38]. Here, the participants were tasked with evaluating speech samples based on three questions: (a) "Is this speech intelligible for a human to comprehend? (Intelligibility)." Notably, StethoSpeech received the highest score (mean: 4.70), significantly improving intelligibility of NAM vibrations (mean: 1.45). (b) "Is this voice hoarse or normal?" A distinct enhancement was observed in speech converted using StethoSpeech, resulting in a more natural and normal voice quality (mean = 4.15). (c) "Does the speech exhibit consistency in prosody, intonation, and articulation? (*i.e.*, whether it preserves prosody)." StethoSpeech (mean: 4.80) again excelled in replicating speech attributes akin to a healthy speaker. Overall, the user study emphasizes that StethoSpeech clearly outperforms other baselines with respect to desired aspects such as naturalness, intelligibility, and consistent prosody.

5.3 Evaluations on the StethoText corpus (speaker-specific models for speakers s1 and s2)

5.3.1 Quantitative evaluation. Table 5 presents error rates computed for NAM vibrations and their converted speech using the StethoSpeech approach for s1 and s2 speaker data. Via the Seq2seq architecture, we render the converted speech using StethoSpeech in three different voices, to highlight the voice selection ability of our framework. We observe that feeding the raw NAM vibrations into the ASR engine results in higher than 100% error rates (it exceeds 100% because of additional word predictions). From this result, we conclude that the NAM

Table 6. Evaluation of StethoSpeech framework on the **StethoText** corpus in a zero-shot setting.

Train speakers	**Test speakers**	**CER ↓**	**WER ↓**
s1 to s10	s11 (female)	22.64	32.37
s1 to s10	s12 (male)	22.54	34.19
all male	all female	24.90	36.80
all female	all male	27.43	39.29

vibrations in the StethoText corpus are completely incomprehensible, and feeding them to ASR model results in near-gibberish output. This is because we obtained NAM vibrations from a normal clinical stethoscope, while the CSTR NAM TIMIT Plus corpus (Table 4) samples were recorded using a specially fabricated device.

We notice that the word error rates on the converted speeches using StethoSpeech for both speakers, irrespective of any target voice, remain within 20%-27%. Just to set the context, these error rates remain in line with the SOTA performance on the problem of whisper-to-speech conversion [38] (WER: 26.68%). It is worth noting that whisper-to-speech conversion has much lower complexity compared to our studied task. Obtained results undeniably underscore the effectiveness of our suggested framework, and lay a solid foundation for additional research on the problem. Finally, when rendering speech in different voices, the best results are achieved when the same gender voice is used. The differences are insignificant in the case of the female speaker. We also experimented with DiscoGAN and MSpec-Net methods on speakers s1 and s2 from our StethoText corpus. In both cases, the resulting CER and WER were over 90% and 110%, respectively, rendering them incomparable to our approach.

5.3.2 User study. Figure 6 presents comparisons of MOS for simulated ground-truth, speech converted using the StethoSpeech approach, and original NAM for speakers s1 and s2. Participants were asked to evaluate 10 randomly selected samples from each speaker across three aspects as specified in Sec. 5.2.3. Results reveal that (a) the intelligibility of the converted speech remains closer to the simulated ground truth. (b) StethoSpeech's capability to synthesize high-quality natural speech, and (c) the converted speech using StethoSpeech is able to preserve tonal aspects similar to ground-truth speech.

5.4 Zero-Shot evaluation

Here, we evaluate StethoSpeech on NAM vibrations of subjects, whose data was not used for training (zero-shot setting). Table 6 presents the quantitative results. First, we train the framework using speakers s1 to s10 and assess the intelligibility of the synthesized speech for speakers s11 and s12. Furthermore, we also evaluate StethoSpeech's speech conversion ability in a more challenging intra-gender setting, where we train the framework on all male speakers' data and evaluate the converted speech for all female speakers, and vice versa. The reported CER's for both the scenarios ranged from 22.54%–27.43%, with WER's in the range of 32.37%–39.29%. We believe that adding more data per speaker may further reduce these error rates. We attribute the generalizability of our framework to the absence of dependency on the fundamental frequency and speaker characteristics in the NAM vibrations.

5.5 Recognition performance in noisy scenarios

In this experiment, we assess the robustness of StethoSpeech-generated speech against recorded NAM vibrations in the challenging environment of loud background music. We conduct qualitative and quantitative evaluations to address the following key questions: (a) Does loud background music interfere with bodily conducted NAM vibrations? (b) Can a speech recognition engine accurately identify speech content from StethoSpeech when

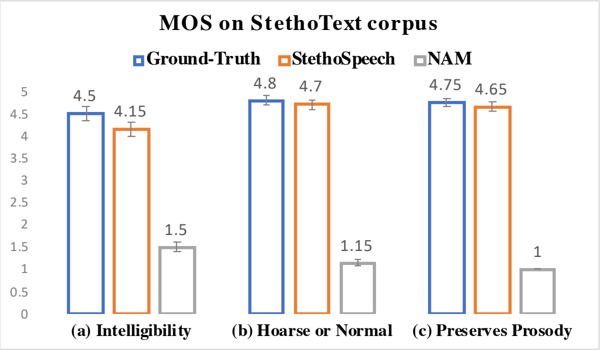


Fig. 6. Quality of Speech conversion from NAM vibrations using StethoSpeech approach on **StethoText** corpus: (a) MOS on Intelligibility, (b) Hoarse-Normal voice rating, and (c) natural prosody rating.

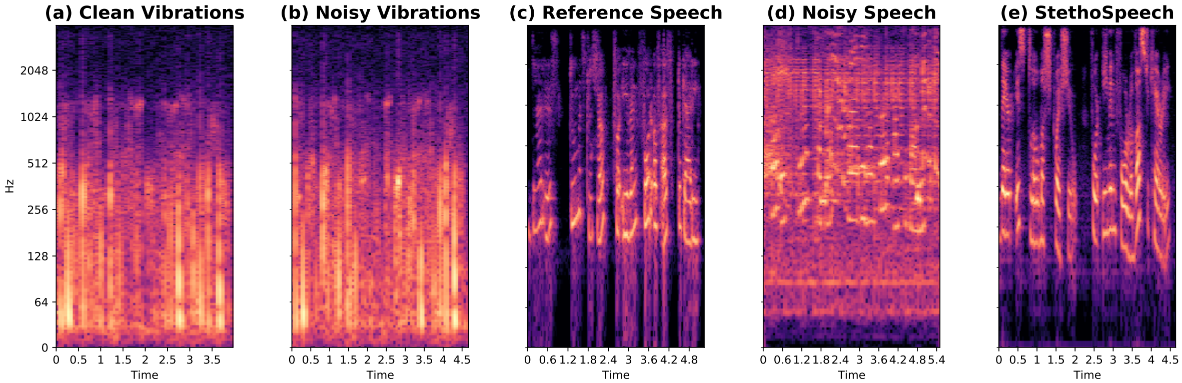


Fig. 7. Mel-spectrum comparison of clean vibrations, noisy vibrations, clean speech, noisy speech, and StethoSpeech output. Transcription: 'There is so much treasure for even four of us to carry'.

NAM vibrations are recorded in the presence of loud background music? (c) How does the spectral quality of StethoSpeech-converted speech compare to reference speech recorded in clean conditions?

5.5.1 Qualitative Mel assessment. In Figure 7 (a) and (b), we present a comparison of the Mel-spectra obtained with clean vibrations, and with loud background music (noisy vibrations). Notably, the formants at low frequencies in the noisy vibrations remain discernible, akin to those in the clean vibrations. This signifies that the NAM vibrations remain largely unaffected by the high-volume background noise. Figure 7 (c) and (d) display the speech spectrum recorded without background music (reference speech) and with loud background music (noisy speech). Evidently, the formants in the noisy speech become completely distorted due to the convolution of loud musical elements with the speech content. In Figure 7 (e), we present the converted speech using StethoSpeech with noisy vibrations as input. The Mel spectra of the converted and reference speech (Figure 7 (e),(c)) are remarkably similar, highlighting StethoSpeech's proficiency in enhancing and predicting intelligible speech content even when NAM vibrations are recorded in a noisy environment.

Table 7. Recognition performance of reference speech, vibrations, and speech converted using StethoSpeech in clean and noisy scenarios, evaluated for speaker s2 in the **StethoText** corpus.

Setting	Signal	CER ↓	WER ↓
Clean	Reference speech	0.51	1.71
	NAM vibrations	119.14	131.90
	Converted speech (StethoSpeech)	**14.69**	**25.57**
Noisy (loud music)	Reference speech	58.73	79.90
	NAM vibrations	88.17	101.47
	Converted speech (StethoSpeech)	**16.95**	**27.45**

5.5.2 Quantitative evaluation. For a thorough quantitative evaluation in noisy environments, we curate a dataset of 223 samples totaling 10 minutes of NAM vibrations. The recordings were compiled from the female subject (s2) in three phases. In the first phase, the reference clean speech was recorded. In the second phase, the same sentences were uttered in the presence of loud background music. In the third phase, the corresponding NAM vibrations were recorded with the loud background music played at the same volume. The ASR results of reference speech, NAM vibrations and StethoSpeech outputs under clean and noisy scenarios are presented in Table 7. In the presence of elevated noise levels, the ASR algorithm applied to the reference speech experiences considerable challenges, resulting in a notable 78% drop in the WER metric compared to the clean setting. In contrast, the converted speech using StethoSpeech gives similar results in both clean and noisy settings (with a minor drop of 2% WER). This remarkable recognition performance, even in noisy environments, highlights StethoSpeech's robustness for achieving real-time and widely-deployable applications.

5.6 StethoSpeech performance on NAM vibrations recorded during walking

To assess our method's capability in synthesizing speech from NAM vibrations recorded under user motion, we gathered 130 samples- equivalent to 10 minutes of data from speaker s1. We extract the sentences from a freely available online story book[2] and exclusively reserve these samples as a test set. Notably, our approach was trained on samples recorded while the speaker was seated, whereas test samples were recorded while the speaker was walking.

The WER computed on the converted speech was 26.70%, while the CER was 17.14%. These error rates align closely with those observed on speech generated from NAM samples compiled with a seated speaker (Table 5). This observation suggests that our method exhibits robustness to body movements. Additionally, we have displayed these samples on our demo page for illustration purposes.

6 Conclusions, Limitations and Future Work

In this work, we introduce StethoSpeech, a novel framework to convert NAM vibrations recorded via a clinical stethoscope from behind the ear into audible and intelligible speech. Through extensive experimentation, we demonstrate that NAM-to-speech conversion is possible using a standard and affordable stethoscope, and our work does away with the reliance on specially fabricated devices for SSI. The second significant advancement on the accessibility front is demonstrating NAM-to-speech conversion without needing paired whisper/speech data. We devise methods to generate synthetic ground-truth speech aligned with the NAM vibrations, which then can be used to train a Seq2Seq network in a non-autoregressive manner. Our proposed novel Seq2Seq network explicitly learns content-specific embeddings to preserve and enhance the intelligibility of NAM vibrations. We validated

[2] https://monkeypen.com/pages/free-childrens-books

the intelligibility, quality, and prosody estimation of the converted speech through extensive experiments on the existing CSTR NAM TIMIT Plus dataset and the newly collected StethoText corpus. Our proposed database amounting to 15.78 hours of NAM vibrations and their corresponding text sets a foundation for further research in both speaker-specific and zero-shot NAM-to-speech conversion. We also showcase that the StethoSpeech framework is successfully usable in loud background noise, such as in a musical concert and is robust to user movements.

We observe that subjects with slower speaking rates exhibit lower word and character error rates in the simulated ground-truth speech using our existing alignment module compared to faster-speaking subjects. Future work would look at refining the acoustic model training to improve the alignments between NAM vibrations and text at faster-speaking rates. While the potential application of our work as a non-invasive conversion tool for voice-impaired patients is intriguing, its actual efficacy in medical settings remains to be seen. In future work, we plan to collaborate closely with ENT specialists and speech therapy centers to investigate the real-time usability of our approach in enhancing the quality of life for patients. StethoSpeech focuses on converting NAMs recorded across the neck. An avenue for future work is designing translation systems that integrate visual modality features, like lip movements and facial expressions, which play a crucial role in shaping intended speech.

Reproducibility statement

All experiments were performed in Python using the pytorch library. All experiments were conducted on computational cluster nodes equipped with a NVIDIA GeForce RTX 2080 Ti GPU.

References

[1] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural Voice Cloning with a Few Samples. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/4559912e7a94a9c32b09d894f2bc3c82-Paper.pdf

[2] Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of Self-Supervised Pre-Training for ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7694–7698. https://doi.org/10.1109/ICASSP40776.2020.9054224

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1044, 12 pages. https://dl.acm.org/doi/abs/10.5555/3495724.3496768

[4] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. AudioLM: A Language Modeling Approach to Audio Generation. 31 (jun 2023), 2523–2533. https://doi.org/10.1109/TASLP.2023.3288409

[5] Itzhak Brook and Joseph F Goodman. 2020. Tracheoesophageal voice prosthesis use and maintenance in laryngectomees. *International Archives of Otorhinolaryngology* 24, 04 (2020), 535–538. https://www.scielo.br/j/iao/a/X6VSZFNCS4VSHwDYBH4WXqp/?lang=en#

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9814838

[7] Steven R. Cox, Julie A. Theurer, Sandi J. Spaulding, and Philip C. Doyle. 2015. The multidimensional impact of total laryngectomy on women. *Journal of Communication Disorders* 56 (2015), 59–75. https://www.sciencedirect.com/science/article/abs/pii/S002199241500043X

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[9] Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Haithem Boussaid, Ebtessam Almazrouei, and Merouane Debbah. 2023. Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Visual to Audio Representation Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13790–13801. https://openaccess.thecvf.com/content/ICCV2023/papers/Djilali_Lip2Vec_Efficient_and_Robust_Visual_Speech_Recognition_via_Latent-to-Latent_Visual_ICCV_2023_paper.pdf

[10] Carol Y. Espy-Wilson, Venkatesh R. Chari, Joel M. MacAuslan, Caroline B. Huang, and Michael J. Walsh. 1998. Enhancement of Electrolaryngeal Speech by Adaptive Filtering. *Journal of Speech, Language, and Hearing Research* 41, 6 (1998), 1253–1264. https://doi.org/10.1044/jslhr.4106.1253

[11] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n* 93 (1993), 27403. https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G/abstract

[12] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. 2017. Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 369–376. https://doi.org/10.1145/1143844.1143891

[14] Horn H, Göz G, Bacher M, Müllauer M, Kretschmer I, and Axmann-Krcmar D. 1997. Reliability of electromagnetic articulography recording during speaking sequences. *European journal of orthodontics* 19, 6 (1997), 647–655. https://doi.org/10.1093/ejo/19.6.647

[15] Panikos Heracleous, Tomomi Kaino, Hiroshi Saruwatari, and Kiyohiro Shikano. 2006. Unvoiced speech recognition using tissue-conductive acoustic sensor. *EURASIP Journal on Advances in Signal Processing* 2007 (2006), 1–11. https://asp-eurasipjournals.springeropen.com/articles/10.1155/2007/94068

[16] Tatsuya Hirahara, Shota Shimizu, and Makoto Otani. 2007. Acoustic characteristics of non-audible murmur. In *The Japan China Joint Conference of Acoustics*, Vol. 100. 4000. https://www.researchgate.net/profile/Tatsuya-Hirahara/publication/251737500_ACOUSTIC_CHARACTERISTICS_OF_NON-AUDIBLE_MURMUR/links/00b7d52a80765613f2000000/ACOUSTIC-CHARACTERISTICS-OF-NON-AUDIBLE-MURMUR.pdf

[17] 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 (oct 2021), 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[18] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. 2023. ReVISE: Self-Supervised Speech Resynthesis With Visual Input for Universal and Generalized Speech Regeneration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18795–18805. https://openaccess.thecvf.com/content/CVPR2023/html/Hsu_ReVISE_Self-Supervised_Speech_Resynthesis_With_Visual_Input_for_Universal_and_CVPR_2023_paper.html

[19] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[20] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad. 2014. Fundamental frequency generation for whisper-to-audible speech conversion. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2579–2583. https://doi.org/10.1109/ICASSP.2014.6854066

[21] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[22] Yoshinobu Kikuchi and Hideki Kasuya. 2004. Development and evaluation of pitch adjustable electrolarynx. In *Speech Prosody 2004, International Conference* (Nara, Japan). https://www.isca-archive.org/speechprosody_2004/kikuchi04_speechprosody.pdf

[23] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Alex Olwal, Jun Rekimoto, and Thad Starner. 2021. Mobile, Hands-free, Silent Speech Texting Using SilentSpeller. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 178, 5 pages. https://doi.org/10.1145/3411763.3451552

[24] Juerg Kollbrunner, Anne-Dorine Menet, and Eberhard Seifert. 2010. Psychogenic aphonia: no fixation even after a lengthy period of aphonia. *Swiss medical weekly* 140, 1-2 (2010), 12–17. https://boris.unibe.ch/1509/1/smw-12776.pdf

[25] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1428, 12 pages. https://dl.acm.org/doi/abs/10.5555/3495724.3497152

[26] Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022. Textless Speech Emotion Conversion using Discrete & Decomposed Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11200–11214. https://doi.org/10.18653/v1/2022.emnlp-main.769

[27] Harshit Malaviya, Jui Shah, Maitreya Patel, Jalansh Munshi, and Hemant A. Patil. 2020. Mspec-Net : Multi-Domain Speech Conversion Network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7764–7768. https://doi.org/10.1109/ICASSP40776.2020.9052966

[28] Héctor A. Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Beltman, Lama Nachman, and Hong Lu. 2020. V-Speech: Noise-Robust Speech Capturing Glasses Using Vibration Sensors. *GetMobile: Mobile Comp. and Comm.* 24, 2 (sep 2020), 18–24. https://doi.org/10.1145/3427384.3427392

[29] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 5. V–708. https://doi.org/10.1109/ICASSP.2003.1200069

[30] Yoshitaka Nakajima and Kiyohiro Shikano. 2006. 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* 120, 5 (2006), 3330–3330. https://www.researchgate.net/publication/272259200_Methods_of_fitting_a_nonaudible_murmur_microphone_for_daily_use_and_development_of_urethane_elastmer_duplex_structure_type_nonaudible_murmur_microphone

[31] Yuto Otani, Shun Sawada, Hidefumi Ohmura, and Kouichi Katsurada. 2023. Speech Synthesis from Articulatory Movements Recorded by Real-time MRI. In *Proc. INTERSPEECH 2023*. 127–131. https://doi.org/10.21437/Interspeech.2023-286

[32] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 48–53. https://doi.org/10.18653/v1/N19-4009

[33] Mihir Parmar, Savan Doshi, Nirmesh J. Shah, Maitreya Patel, and Hemant A. Patil. 2019. Effectiveness of Cross-Domain Architectures for Whisper-to-Normal Speech Conversion. In *2019 27th European Signal Processing Conference (EUSIPCO)*. 1–5. https://doi.org/10.23919/EUSIPCO.2019.8902961

[34] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*. 3615–3619. https://doi.org/10.21437/Interspeech.2021-475

[35] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Prajwal_Learning_Individual_Speaking_Styles_for_Accurate_Lip_to_Speech_Synthesis_CVPR_2020_paper.pdf

[36] Thomas F. Quatieri. 2002. *Discrete-time speech signal processing: principles and practice.* Pearson Education India. https://books.google.co.in/books/about/Discrete_time_Speech_Signal_Processing.html?id=5KYeAQAAIAAJ

[37] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

[38] Jun Rekimoto. 2023. 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 700, 12 pages. https://doi.org/10.1145/3544548.3580706

[39] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*. https://openreview.net/forum?id=piLPYqxtWuA

[40] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. *FastSpeech: fast, robust and controllable text to speech.* Curran Associates Inc., Red Hook, NY, USA. https://dl.acm.org/doi/abs/10.5555/3454287.3454572

[41] Neha Sahipjohn, Neil Shah, Vishal Tambrahalli, and Vineet Gandhi. 2023. RobustL2S: Speaker-Specific Lip-to-Speech Synthesis exploiting Self-Supervised Representations. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1492–1499. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10317357

[42] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*. 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

[43] Neil Shah, Saiteja Kosgi, Vishal Tambrahalli, Neha S, Anil Nelakanti, and Vineet Gandhi. 2024. ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations. In *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, St. Julian's, Malta, 79–91. https://aclanthology.org/2024.findings-eacl.6

[44] Neil Shah, Nirmesh Shah, and Hemant Patil. 2018. Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion. In *Proc. Interspeech 2018*. 3157–3161. https://doi.org/10.21437/Interspeech.2018-1565

[45] Nirmesh J Shah, Mihir Parmar, Neil Shah, and Hemant A Patil. 2018. Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion. In *Machine Learning in Speech and Language Processing (MLSLP) Workshop*. Google Office, 1–3. https://www.researchgate.net/publication/326668619_Novel_MMSE_DiscoGAN_for_Cross-Domain_Whisper-to-Speech_Conversion

[46] Shota Shimizu, Makoto Otani, and Tatsuya Hirahara. 2009. Frequency characteristics of several non-audible murmur (NAM) microphones. *Acoustical science and technology* 30, 2 (2009), 139–142. https://www.jstage.jst.go.jp/article/ast/30/2/30_2_139/_pdf

[47] Abhayjeet Singh, Amala Nagireddi, Anjali Jayakumar, Deekshitha G, Jesuraja Bandekar, Roopa R, Sandhya Badiger, Sathvik Udupa, Saurabh Kumar, Prasanta Kumar Ghosh, Hema A Murthy, et al. 2024. Lightweight, Multi-Speaker, Multi-Lingual Indic Text-to-Speech.

IEEE Open Journal of Signal Processing 5 (2024), 790–798. https://doi.org/10.1109/OJSP.2024.3379092

[48] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. LipLearner: Customizable Silent Speech Interactions on Mobile Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 696, 21 pages. https://doi.org/10.1145/3544548.3581465

[49] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. 2007. Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 8 (2007), 2222–2235. https://doi.org/10.1109/TASL.2007.907344

[50] Tomoki Toda, Keigo Nakamura, Hidehiko Sekimoto, and Kiyohiro Shikano. 2009. Voice conversion for various types of body transmitted speech. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 3601–3604. https://doi.org/10.1109/ICASSP.2009.4960405

[51] Tomoki Toda and Kiyohiro Shikano. 2005. NAM-to-speech conversion with Gaussian mixture models. In *Proc. Interspeech 2005*. 1957–1960. https://doi.org/10.21437/Interspeech.2005-611

[52] Viet-Anh Tran, Gérard Bailly, Hélène Lœvenbruck, and Tomoki Toda. 2010. Improvement to a NAM-captured whisper-to-speech system. *Speech Communication* 52, 4 (2010), 314–326. https://doi.org/10.1016/j.specom.2009.11.005 Silent Speech Interfaces.

[53] László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. 2018. Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces. In *Proc. Interspeech 2018*. 3172–3176. https://doi.org/10.21437/Interspeech.2018-1078

[54] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6309–6318. https://dl.acm.org/doi/10.5555/3295222.3295378

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

[56] Patrick von Platen. 2022. *Fine-tuning Wav2Vec2 for English ASR*. https://colab.research.google.com/github/patrickvonplaten/notebooks/blob/master/Fine_tuning_Wav2Vec2_for_English_ASR.ipynb Google Colab Notebook, Last accessed on 06 September 2023.

[57] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (sep 2020), 24 pages. https://doi.org/10.1145/3369812

[58] Chen-Yu Yang, Georgina Brown, Liang Lu, Junichi Yamagishi, and Simon King. 2012. Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation. In *2012 8th International Symposium on Chinese Spoken Language Processing*. 220–223. https://doi.org/10.1109/ISCSLP.2012.6423522

[59] Shang Zeng, Haoran Wan, Shuyu Shi, and Wei Wang. 2023. mSilent: Towards General Corpus Silent Speech Recognition Using COTS mmWave Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 39 (mar 2023), 28 pages. https://doi.org/10.1145/3580838

[60] zeta chicken. 2017. *toWhisper*. https://github.com/zeta-chicken/toWhisper 2017 github.