# SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices

Yongjian Fu[1], Shuning Wang[1], Linghui Zhong[1], Lili Chen[2], Ju Ren[2,3*], Yaoxue Zhang[2,3]
[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]Department of Computer Science and Technology, BNRist, Tsinghua University
[3]Zhongguancun Laboratory, Beijing, China
Email:{fuyongjian,shuning.wang,zlh2021}@csu.edu.cn,{lilichen,renju,zhangyx}@tsinghua.edu.cn

## ABSTRACT

Silent Speech Interface (SSI) has been proposed as a means of reconstructing audible speech from silent articulatory gestures for covert voice communication in public and voice assistance for the aphasic. Prior arts of SSI, either relying on wearable devices or cameras, may lead to extended contact requirements or privacy leakage risks. The recent advances in acoustic sensing have brought new opportunities for sensing gestures, but their original intention is to infer speech content for classification instead of audible speech reconstruction, resulting in the loss of some important speech information (e.g., speech rate, intonation, and emotion). In this paper, we propose , the first system that supports accurate audible speech reconstruction by analyzing the disturbance of tiny articulatory gestures on the reflected ultrasound signal. The design of introduces a new model that provides the unique mapping relationship between ultrasound and speech signals, so that the audible speech can be successfully reconstructed from the silent speech. However, establishing the mapping relationship depends on plenty of training data. Instead of the time-consuming collection of massive amounts of data for training, we construct an inverse task that constitutes a dual form with the original task to generate virtual gestures from widely available audio (e.g., phone calls) for facilitating model training. Furthermore, we introduce a fine-tuning mechanism using unlabeled data for user adaptation. We implement using a portable smartphone and evaluate it in various environments. The evaluation results show that can reconstruct speech with a (Character Error Rate) CER as low as 7.62%, and decrease the CER from 82.77% to 9.42% on new users with only 1 hour of ultrasound signals provided, which outperforms state-of-the-art acoustic-based approaches while preserving rich speech information.

## CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**.

*Corresponding author.

*SenSys'22, Nov 6-9, 2022, Boston, USA*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/3560905.3568530

## KEYWORDS

Acoustic sensing, Silent Speech, Transformer, cGAN.

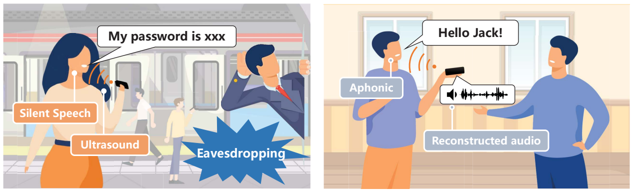
**ACM Reference Format:**
Yongjian Fu[1], Shuning Wang[1], Linghui Zhong[1], Lili Chen[2], Ju Ren[2,3*], Yaoxue Zhang[2,3]. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys'22).* ACM, New York, NY, USA, 15 pages. https://doi.org/3560905.3568530

## 1 INTRODUCTION

Voice communication is an important means of enabling people to communicate with humans or machines. However, voice communication often faces troubles in different scenarios, such as eavesdropping concerns in public places and soundless requirements in silent places. Silent Speech Interface (SSI) is an advanced technology that enables covert voice communication in public by reconstructing speech from silent articulatory gestures. It not only opens a covert way for voice-based interactions, but also can help the people who acquired voice disorders (e.g., laryngectomy) to regain the ability of voice communication.

In order to capture the articulatory gestures, classical SSI solutions commonly use various wearable sensors (e.g., EMG, EEG) [4, 19, 30] to sense the movements of vocal organs (e.g., lips, tongue). Nevertheless, these wearable sensors require body contact or even device implantation, which hinders the user's daily activities and may cause anaphylactic reactions (e.g., skin irritations). To achieve user transparency, contactless methods have been widely studied for various SSI applications. A representative category is to leverage vision information to extract the features of the articulatory gestures and generate corresponding speech [6, 16, 46, 47]. However, cameras introduce privacy concerns and cannot maintain high performance in low-light conditions. To address the limitations of vision-based solutions, recent advances have explored how to use wireless signals (e.g., WiFi [64] and acoustic [20, 59, 70, 72] ) for articulatory gesture sensing. Nevertheless, current wireless-based solutions are unsuitable for voice communication, since the original intention of infer speech content from reflected signals for command classification instead of reconstruction of audible speech, resulting in the loss of some important speech information such as speech rate, intonation and emotion. Therefore, how to acquire a manner that can enable silent voice communication using wireless signals remains an open challenge.

In this paper, we propose SVoice for recovering audible speech in silence using ultrasound sensing based on commercial mobile devices (e.g., smartphones). Figure 1 illustrates the practical application scenarios of SVoice, including covert voice communication



(a) Covert voice communication in public.(b) Assisting aphasics to speak with voice.

**Figure 1: Motivating examples of SVoice. The (a) shows how SVoice can help people to communicate covertly with the voice in silent conditions to prevent eavesdropping. The (b) shows SVoice as a new interactive interface to restore voice for people who have acquired voice disorders.**

in public and assisting aphasics to speak with voice. SVoice implements a regression model, which maps articulatory gestures directly to audible speech via ultrasound for restoring rich speech information.

To realize this high-level idea, we need to address the following challenges. To begin with, exploiting ultrasound as the vinculum between articulatory gestures and speech is nontrivial for two reasons. Firstly, it is challenging to extract fine-grained changes from reflected signals caused by the fast (-80cm/s~80cm/s) and subtle (<5cm) movements, then reconstruct high-dimensional speech from low-dimensional ultrasound. Secondly, since speech is the result of a high degree of overlap and continuous movements [39], the mapping relationship between articulatory gestures and speech is notoriously complex. The complex mapping relationship makes it difficult for DNN models trained on the dataset with a limited number of sentences to maintain generalization on unseen sentences. Therefore, it requires plenty of training data containing diverse sentences compared to classification tasks, which drives unaffordable data collection overhead. Moreover, SVoice should guarantee adaptability to different users. However, since speech reconstruction is notoriously speaker-dependent due to the differences in vocal timbre, getting a generic model by training is arduous. Although fine-tuning offers a viable option, sufficient pairwise training data is required to transform timbre. More importantly, people who have acquired voice disorders are inadequate to provide an audible speech.

To tackle these challenges, we first fully exploit the advantages of ultrasound and speech, i.e., high sampling rate and rich contextual information, respectively. We design an ultrasonic waveform to capture the high-resolution magnitude and phase gradients and apply a two-stream convolution structure and multi-attention heads to extract the multi-channel context of ultrasound signals for reconstructing the high-dimensional speech. To reduce the data collection overhead of establishing the mapping relationship, we exploit the rich sentence diversity in wild audio (e.g., phone calls), and construct an inverse task that constitutes a dual form with the original task inspired by the back-translation. The inverse task aims to generate virtual articulatory gestures from wild audio based on conditional GAN (cGAN) and cross-modal similarity measure network, enabling the widely available data from the target domain to facilitate model training. Moreover, we propose a fine-tuning scheme and design a tailored loss function that utilizes unlabeled ultrasound signals and cross-domain temporal calibration to fine-tune the DNN model for new users. Finally, we implement SVoice on a portable smartphone and verify its superior performance compared to existing baselines in various environments.

**Contributions:** To summarize, our main contributions are as follows:

- We propose SVoice, an end-to-end Silent Speech Interface that reconstructs audible speech using ultrasound signals reflected by articulatory gestures in silence and preserves rich speech information.
- We design a cross-modal data augmentation mechanism that can generate virtual articulatory gestures from wild audio. To the best of our knowledge, this is the first time to propose a dual form in wireless sensing tasks, enabling widely available data from the target domain to facilitate model training, which can be readily extended to other sensing tasks.
- We present a user adaptation technique that utilizes a small amount of unlabeled ultrasound data to fine-tune the model for different users.
- We collect a new dataset and conduct extensive experiments to demonstrate the performance of our system. The experimental results show the high efficiency and robustness of SVoice, achieving a CER of 7.62%, and decreasing the CER from 82.77% to 9.42% on new users with only 1 hour of ultrasound signals provided.

## 2 SPEECH SENSING VIA ULTRASOUND

This section presents the signal design and processing methods, as well as feasibility analysis, for reconstructing human speech using ultrasound in SVoice.

### 2.1 Acoustic Signal Design and Processing

*2.1.1 Acoustic Signal Design.* Human speech involves many organs. For example, the different vibration patterns of the vocal folds produce the pitch of the sound which makes up the basic unit of articulation(i.e. phoneme), and the combined movements of multiple organs (e.g. lips, tongue, jaw) produce different articulatory gestures [8]. During the normal speech, there is a certain correspondence between articulatory gestures and phonemes [13]. More specifically, uttering a phoneme corresponds to the coordinated movements of multiple articulators, including the production of airflow in the lungs, the vibration of the vocal cords, the protrusion of the lips, and the closure and extension and retraction of the tongue.

In the silent speech, the vibrations of the vocal cords are limited, but facial muscle movements such as the lips, jaw, and tongue are preserved. Therefore, it would be possible to recover the speech signals if we can fully capture and interpret the articulatory gestures. However, articulatory gestures are very subtle and rapid, typically lasting 100-700ms [57] and moving less than 5cm [61], making it challenging to capture the fine-grained gesture motion by using a single microphone [59]. In order to characterize the fine-grained articulatory gestures, the signal design of SVoice needs to satisfy the following two characteristics: i) High sampling rate that can track the rapid changes of the articulatory gestures; ii) High resolution to track the subtle changes in the facial muscles.

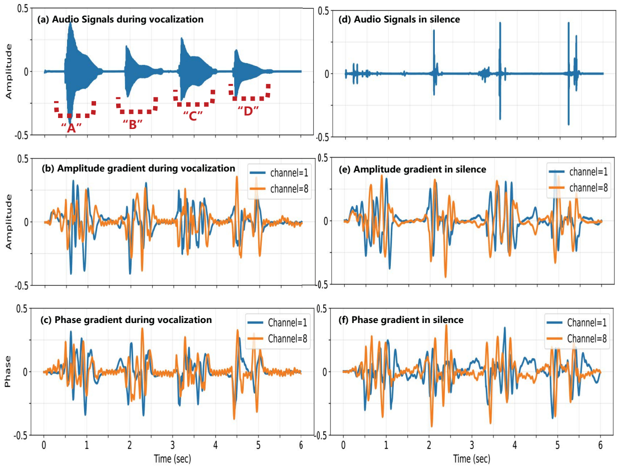


**Figure 2: A subject is asked to utter "A, B, C, D" while remaining vocalized and silent, respectively. (a), (b), (c) are the speech and ultrasound signals collected during vocalization, while (d), (e), (f) are collected during silence. The fluctuations in the gradients of phase and amplitude are correlated with human speech and similar during vocalized and silent speech.**

To satisfy these requirements, we set the transmit signal as a continuous wave (CW) of $Acos(2\pi ft)$, where $A$ is the amplitude and $f$ is the frequency of the signal. By analyzing the amplitude and phase of the reflected waves, we are able to obtain the changes in the articulatory gestures. Although modulated CW signals, such as Frequency Modulated Continuous Wave (FMCW) [9] and Orthogonal Frequency Division Multiplexing (OFDM) [43] signals are widely used in wireless sensing tasks as their resistance to multipath and frequency-selective fading. However, since these signals are processed in frames, the loss of inter-frame information results in reduced resolution. As for the Doppler shift, since the hardware artifacts and harmonics, mutual interference exists between ultrasound signals and audio signals [57], resulting in Doppler differences between vocalized and silent speech, which is detrimental to the training model. Therefore, we utilize the phase and amplitude of the CW signal for fine-grained gesture sensing. The CW signal is transmitted by commercial acoustic equipment with a sampling rate of 48 kHz, and each sampling point can capture one feature point, i.e., the resolution can reach 0.71 cm. In addition, multiple single-frequency subcarriers are chosen to resist multipath effects, where the frequency band $\Delta f$ is set to 700 Hz to prevent interference. Considering that most commercial devices support the inaudible frequency band of 17-22 kHz, the number of subcarriers $N$ is set to 8. Then, the final transmit signal is $T(t) = \sum_{i=1}^{N} Acos(2\pi ft)$, where $i$ is the i-th subcarrier and $f_i$ represents the frequency of each subcarrier.

*2.1.2 Signal Processing.* The ultrasonic waves are separated from the reflected signals by a high-pass filter, and passed through a band-pass filter to obtain the signal $R_i(t) = A_i' cos(2\pi f_i t + \phi_i)$, where $A_i'$ is the amplitude of the i-th subcarrier and $\phi_i$ is the phase shift of the i-th subcarrier. Then, the amplitude and phase of the signal are obtained using a coherent demodulator. Specifically, we multiply the filtered signal by $cos(2\pi ft)$:

$$\begin{aligned} R_i \times cos(2_i t) &= A_i cos(2\pi ft + \phi) \times cos(2\pi f_i t) \\ &= \frac{A_i}{2} cos(4\pi ft + \phi) \times \frac{A_i}{2} cos(\phi). \end{aligned} \tag{1}$$

Similarly, we multiply the filtered signal by $-sin(2\pi ft)$:

$$\begin{aligned} R_i \times sin(2_i t) &= A_i cos(2\pi ft + \phi) \times -sin(2\pi f_i t) \\ &= -\frac{A_i}{2} sin(4\pi ft + \phi) \times \frac{A_i}{2} sin(\phi). \end{aligned} \tag{2}$$

We filter out the additionally introduced high frequency $2f$ through a low-pass filter and then obtain $I = \frac{A_i}{2} cos(\phi)$, $Q = \frac{A_i}{2} sin(\phi)$. For reducing the computational cost of training the DNN model, the $I$, $Q$ signals are smoothed by the mean filter. Finally, in order to eliminate static interference, we obtain the signal gradients using the differential approach by subtracting the previous sample point from the latter one, i.e. $R_i(t)' = R_i(t) - R_i(t-1)$.

## 2.2 Feasibility Analysis

In order to capture the correlation between ultrasound and speech, the user's speech is vocalized instead of silent when collecting data for the DNN model training. Thus, the model can use the perfect alignment of ultrasound and clean speech in the temporal domain to establish correspondence. To validate the relationship between the received ultrasound signals and the speech content, we conduct this proof-of-concept. In this experiment, we ask a subject to say "A, B, C, D" once each while remaining vocalized and silent, respectively. The microphone on the bottom of the smartphone is fixed 4 cm in front of the subject. Figure 2 shows the fluctuations of the amplitude and phase gradients in channel 1 (17 kHz) and channel 8 (21.9 kHz) caused by articulatory gestures. We observe a clear correspondence between ultrasound and speech. In addition, the fluctuation patterns of phase and amplitude gradients are similar with the same speech content between vocalized and silent conditions, but distinguishable with different speech content. Based on this observation, we demonstrate the feasibility of reconstructing the human voice using ultrasound in silence.

# 3 SYSTEM OVERVIEW

SVoice is designed to fulfill the following objectives:1) reconstruct precise and natural speech from ultrasound in silence with affordable model training cost; 2) guarantee the adaptability to different users. Figure 3 depicts the system architecture of SVoice.

In the training stage, we collect the vocalized speech from source user, which contains synchronously recorded ultrasound and audio. After signal processing, the ultrasound signals are converted to multi-channel amplitude and phase gradients and form training data pairs with audio. SVoice utilizes the precise temporal alignment of ultrasound and clean audio to train the SiVoNet network to establish correspondence. Furthermore, a well-designed virtual gesture generation method is applied to substantially increase the number of training samples to facilitate model training. In the inference stage, a small amount of silent speech (without audio) from the target user is employed to fine-tune the model for optimal prediction outcomes. Note that the collected vocalized speech shown in Figure 3 is only used during the training stage and the fine-tuned SiVoNet model can be applied directly for inference. To achieve the

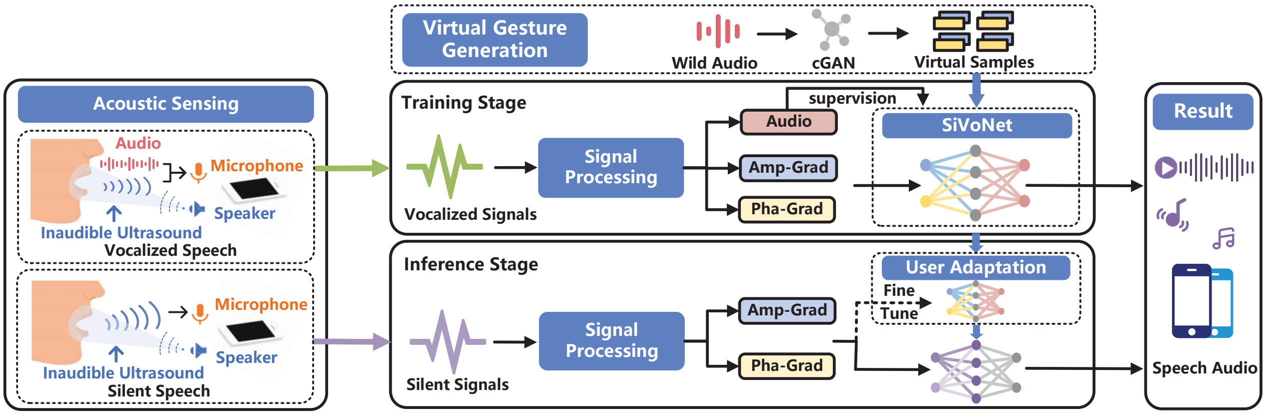


Figure 3: The system architecture of SVoice mainly consists of acoustic sensing and signal processing modules, SiVoNet for synthesizing speech from ultrasound, cGAN for generating virtual gestures from speech, and a fine-tuning mechanism using label-free data.

above goals, SVoice integrates three core components:

**SiVoNet** (Section 4) We design a tailored speech reconstruction model, named SiVoNet, which can extract and fuse the phase and amplitude features of ultrasound and contextual information, and then quickly generate the corresponding clear and intelligible speech.

**Virtual Gesture Generation** (Section 5) To establish the complex mapping between speech and articulatory gestures, we design a virtual gesture generation strategy based on cGAN to reduce involved data collection costs. Inspired by back translation in machine translation, our design utilizes easily collected audio (e.g., voice records or voice chats) to reversely synthesize ultrasound signals to increase the number of training samples for articulatory gestures and real speech data pairs, thereby advancing SiVoNet to learn complicated mapping relationships and enhancing the generalization ability of SiVoNet to unseen sentences.

**User Adaptation** (Section 6) The intuition is that the determining factor for the change of ultrasound signal is the speech content corresponding to the articulatory gesture, and ultrasound contains fewer personal characteristics (e.g., timbre) than audio. Consequently, to accommodate various users, SVoice develop a fine-tuning scheme to match the unlabeled ultrasound signal from the target user and the audio with the same speech content from the source user as training data pairs, avoiding the timbre problem. We construct a new loss function to reduce timing distortion between cross-user data pairs, which enables us to fine-tune the SiVoNet model for new users.

## 4 SPEECH RECONSTRUCTIONS

SiVoNet is designed to reconstruct audio from ultrasound signals, as shown in Figure 4, which consists of three main modules: (1) *Feature Embedding Module*, which is to extract feature embeddings from phase and amplitude fluctuations in ultrasound signals; (2) *Speech Parameters Predictor*, which is to predict speech parameters using the contextual information in feature embeddings; (3) *Vocoder Module*, which is to rebuild clear and audible audio from speech parameters.

### 4.1 Feature Embedding

The phase and amplitude gradients of the ultrasound signal with time series length $T^u$ are sent into the feature embedding module, denoted as $U^p \in \mathbb{R}^{T^u \times C^u}$ and $U^a \in \mathbb{R}^{T^u \times C^u}$ respectively, where $C^u = 8$ is determined by the number of subcarriers. The "blue" part in Figure 4 illustrates the basic structure of the module. Considering the different noise and fluctuation patterns of phase and amplitude, SiVoNet first applies a dual-stream structure with 1D convolution Resblocks [26] to extract the features of $U^p$ and $U^a$ separately along the time dimension. In each stream, downsampling convolutional kernels are employed to reduce duplicate information in the temporal dimension and to align with the temporal resolution of speech features. Notably, the dual streams are trained independently without sharing weights. Then, SiVoNet concatenates the outputs of the dual streams along the channel dimension and feeds them into multi-layer 1D convolution to fuse the two types of features and transform them into the same feature space.

Note that all the above convolutional layers employ dilated convolution to increase the perceptual field of the convolutional kernels in the time domain and to better aggregate the information in the temporal dimension.

### 4.2 Speech Parameters Predictor

The critical challenge of SiVoNet is to reconstruct the audio waveform of a high-dimensional stream from the sensing data of a low-dimensional stream due to information loss. To address this problem, SiVoNet avoids directly predicting temporal waveforms, and instead utilizes the speech features of audio, Mel-frequency cepstral coefficients (MFCC) [49], as intermediate representations. Consequently, the objective of the speech parameters predictor is to map feature embedding on each time slot to the corresponding MFCC. Since MFCC retains only the amplitude envelope information of the audio signal in the time-frequency domain and discards the phase information, this greatly reduces the difficulty of the task. Nevertheless, the sensing feature on a single time slot is still insufficient to accurately predict MFCC due to the lack of sufficient information.

We thus design a bi-directional Transformer encoder [12], where a multi-headed attention mechanism can extract the contextual information of the feature embedding from multiple dimensions over a longer time horizon.

Treat the output $U^o = (U^o_1, \ldots, U^o_{T^m})$ of the feature embedding module as the input, where $U^o_i \in \mathbb{R}^d, i = 1, \ldots, T^m$ and $d$ represents the hidden layer dimension. For the input time series $U^o$, each attention head computes the scaled dot-product attention over each subspace in parallel, and outputs a sequence $Z = (Z_1, \ldots, Z_{T^m})$ of the same time length. Thus, the output sequence $Z$ is linearly weighted by each element in the input sequence $U^o$. The naive multi-head attention mechanism ignores the positional relationship of each element in the input sequence. Although there is no strict timing restriction on vocalizations, some vocalizations are generally regulated by grammatical expression rules, some vocalizations still have a relative sequence relationship (such as fixed phrases). SiVoNet uses learnable relative position embeddings instead of absolute position embeddings for capturing time-invariant relative position relationships in $U^o$. Each element $Z_i \in \mathbb{R}^{d/h}, i = 1, \ldots, T^m$ in the output Z of each attention head can be expressed as:

$$Z_i = \sum_{j}^{T^m} \alpha_{i,j}(U^o_j W^V + p^V_{i,j}), \tag{3}$$

where

$$\alpha_{i,j} = \text{softmax}(\frac{(U^o_i W^Q)(U^o_j W^K + p^K_{i,j})^{\mathrm{T}}}{\sqrt{d/h}}). \tag{4}$$

$W^Q, W^K, W^V$ represent the trainable query matrix, key matrix, and value matrix, respectively. The matrices are unique in each attention head. $\alpha_{i,j}$ represents the weight coefficient which is calculated by computing $W^Q$ and $W^K$. $p_{(i,j)}$ represents the relative position distance between $U^o_i$ and $U^o_j$ within a clipping distance $k$, which is only related to the relative distance of $i$ and $j$ and the learnable parameters $\omega$. $p^K_{(i,j)}$ and $p^V_{(i,j)}$ can be written as:

$$p^K_{i,j} = \omega^K_{clip(j-i,k)}, \tag{5}$$

where

$$\text{clip}(x, k) = \max(-k, \min(k, x)). \tag{6}$$

The predicted MFCC speech parameters $M \in \mathbb{R}^{T^m \times d^m}$ can be obtained by concatenating the outputs of multiple attention heads and going through a linear projection layer. $d^m = 39$ is the dimension of MFCC.

## 4.3 Vocoder

The vocoder module aims to recover the waveform of the speech in the time domain from the speech parameters, which determines the quality of the synthesized speech. The vocoder is generally based on the classical source-filter model of the human vocal mechanism that divides the speech generation process into two main independent modules of excitation source and vocal tract response. Traditional vocoder based on digital signal processing is fast, but the synthesized speech quality is poor due to simple speech modeling. The vocoder based on the full neural network synthesizes speech with high quality but cannot run in real-time. In order to balance the speed and quality of speech synthesis, we use LPCNet [63] as the vocoder. LPCNet combines the advantages of the nonlinear fitting capability of neural networks and the simplicity of LPC filters. Specifically, LPCNet reduces the difficulty of acoustic modeling by simplifying the modeling of the vocal tract using LPC filters and fitting the excitation sources using only neural networks. The generation process of the sampling point $R_i$ of the speech at $t$ is as follows:

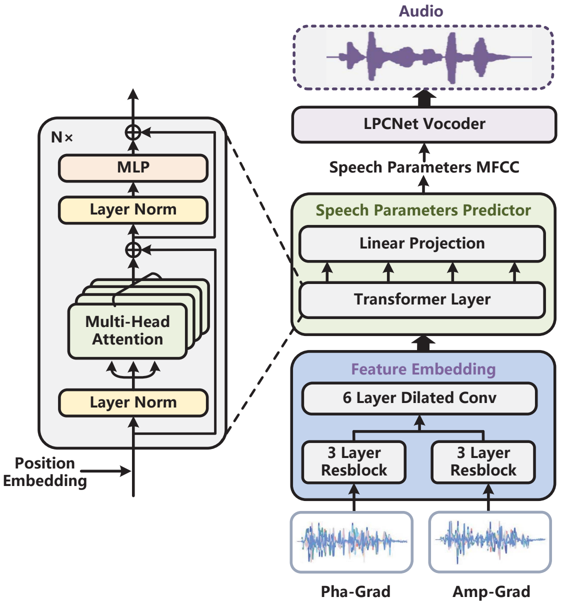


**Figure 4: The structure of SiVoNet to perform exhaustive extraction of speech information from ultrasound.**

$$R_t = \sum_{k=1}^{M} \beta_k s_{t-k} + e_t, \tag{7}$$

where $\beta_k$ represents the $k - th$ LPC parameter of the current frame, which is calculated by the LPC filter, $s_{(t-k)}$ represents the sampling point at time $t - k$, and $e_t$ represents the residual at time $t$. LPCNet regards $e_t$ as an excitation source for neural network fitting. As shown in Figure 5, LPCNet is divided into three modules, where the LPC filter module computes $\beta$ from speech parameters. Notably, we choose MFCC as the input of LPCNet to recover the time-domain waveform instead of Bark-frequency cepstral coefficients (BFCC), which are too narrow in the low-frequency part and lead to inaccurate estimation. The frame rate network comprises two convolutional and fully connected layers, providing a conditional vector input to the sample rate network. And the sample rate network is the core module that uses two GRU layers to predict $e_t$ in an autoregressive manner.

# 5 VIRTUAL GESTURE GENERATION

In this section, we present our virtual gesture generation scheme based on cGAN.

## 5.1 Intuitive Insight

Establishing complex mapping relationships between speech and ultrasound signals relies on large amounts of pairwise training data,

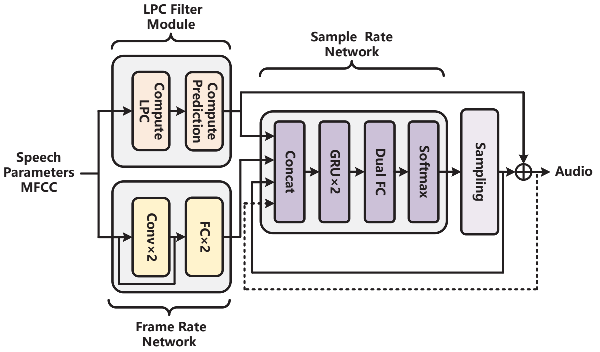


**Figure 5: The structure of LPCNet Vocoder, which synthesizes speech from MFCC.**

which requires unaffordable data collection overhead. Therefore, inspired by back-translation from machine translation [5, 34, 35, 52], we design a novel virtual gesture generation scheme that generates ultrasound signals from wild audio.

In machine translation tasks, the key idea of back-translation is that abundant pseudo-data pairs can be synthesized by training a back translation model from the target language to the source language, under the condition that the target corpus is more readily available [51]. Similarly, based on the fact that it is easy to collect audio data from daily life (e.g., calling or voice chatting), we can train an inverse model for generating ultrasound signals from audio, which encourages additional data pairs to be obtained from the new audio alone. Our insight is that generating virtual gesture data from audio can increase the diversity of sentence sets and serve as a denoising training method to disambiguate forward models, thereby improving generalization performance.

Note that unlike the previous work [69] in sign language translation, which achieved "back translation" by changing the order of the words in the original sentence, we are working on completely new sentences. We believe that the proposed insight is equally applicable to other sensing tasks, significantly reducing the data collection overhead.

## 5.2 Model Design

The ability of cGAN [42] to transform across domains has been commonly demonstrated in areas such as image generation [3, 21, 42], speech enhancement [7, 18, 41], and melody synthesis [66]. Our proposed cGAN framework for cross-domain audio-to-ultrasound translation is illustrated in Figure 6. The framework consists of two key components: a generator $G$ and a discriminator $D$. Let $(U, M)$ denote a pair of ultrasound vectors and its audio vector. By conditioning the model on additional audio information $M$, $G$ is trained to capture the true distribution of the corresponding ultrasound data $U^r$, while $D$ tries to distinguish the true data $U^r$ from the pseudo data $U^g$ synthesized by $G$. The objective function $V(G, D)$ is as follows:

$$\min_{G} \max_{D} V(D, G) = \mathcal{L}_D(G, D) + \mathcal{L}_G(G, D). \tag{8}$$

Using the trained $G$, we can generate dozens of corresponding ultrasound data for an unseen audio to enable virtual gesture generation.

### 5.2.1 *Similarity Assessment Discriminator.*

A key challenge for discriminators is how to design for effective identification of real and generated data. The traditional approach is to project the prediction into 1 dimension, which represents the "real or fake" probability of the input data, and then calculate the loss accordingly. However, due to the 1-dimensional output containing limited information, it is hard to force the discriminator to learn the correlation between the two modalities, audio and ultrasound, to reasonably distinguish "real or fake". Hence, inspired by the work on speech enhancement involving multiple modalities similarly [57, 71], we use the Siamese network [33] and Triplet loss [24] to train the cross-domain discrimination model and evaluate the similarity between the two modalities.

Defining $(U^r, M)$ and $(U^g, M)$ as two types of input pairs (real pair and fake pair) for $D$, where $U$ is a vector of ultrasound phase and amplitude gradients in series and $M$ represents the MFCC parameters. We first design two sub-networks: the ultrasound sub-network and the audio sub-network, which extract embedded similarity features for $U$ and $M$ respectively. These two sub-networks have a similar network structure, with a Resblock and a convolutional layer to abstract features, a BLSTM layer to obtain temporal context information, followed by an FC layer to project the final similarity feature vector. Note that in order to fairly estimate the similarity vectors of two different inputs of the same modality, the ultrasound sub-network deploys a Siamese network structure that shares architecture and weights for the inputs $U^r$ and $U^g$.

To enable $D$ to better "understand" the correlation between ultrasound and audio, facilitating correct identification, we use the Triplet loss function to update the model parameters. Specifically, for $D$, the input to loss is a triple $(M, U^r, U^g)$, where $M$ is considered as the anchor, $U^r$ as the positive sample, and $U^g$ as the negative sample. Aiming to minimize the distance of real pairs $(M, U^r)$ and maximize the distance of false pairs $(M, U^g)$ in the feature space, the loss function for $D$ is designed as follows:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{D}}(G, D) &= \mathbb{E}_{M,U^r,U^g \sim p_{\text{data}}(M,U^r,U^g)} \\ &[\|f_a(M) - f_u(U^r)\|_2^2 - \|f_a(M) - f_u(U^g)\|_2^2 + \alpha]_+,\end{aligned} \tag{9}$$

where $f_a$ and $f_u$ represent the audio and ultrasound sub-networks respectively, and $\alpha$ is a margin distance that is enforced between real and fake pairs. Based on the insight that the closer the distance between two points in the same feature space, the more similar they are implied to be (i.e., the higher the correlation), $D$ is trained to learn how to capture the correlation between audio and ultrasound to distinguish between real and fake ultrasound.

### 5.2.2 *Audio-to-Ultrasound Generator.*

For the generator $G$, our goal is to learn the mapping from the auxiliary audio to the underlying ultrasound variation patterns. The input of $G$ is the MFCC speech features $M \in \mathbb{R}^{T^m \times d^m}$ calculated from the real audio and a random noise vector $N \in \mathbb{R}^{T^m \times d^m}$. Note that we set the dimensions of $M$ and $N$ to be the same. Two modules, feature extractor and ultrasound predictor, are developed in G based on our intuition that pairwise tasks can use a similar model structure. In order to extract features, we likewise apply a Resblock-based dual-stream structure on the input sequence and then feed the concatenated output into a 6-layer 1D dilated convolution. In particular, contrary to SiVoNet (Section 4), we utilize transposed convolution [15] in the Resblock

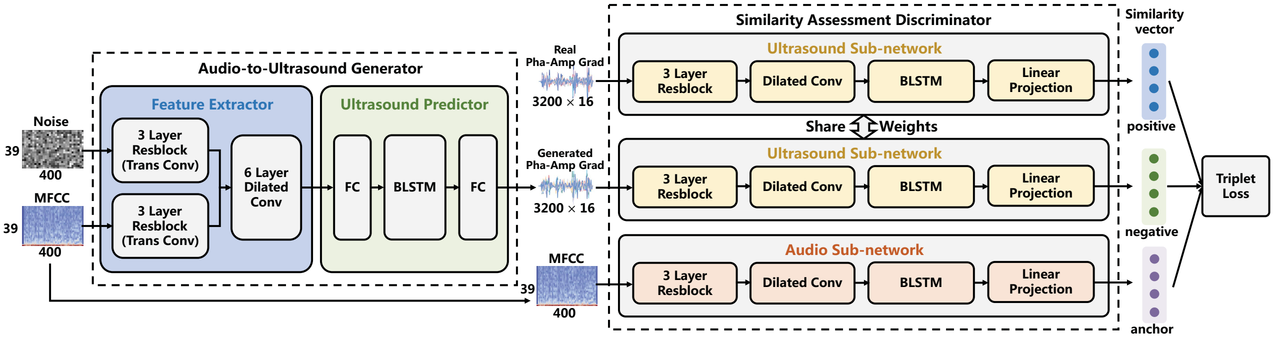


**Figure 6: The architecture of cGAN-based virtual gesture generation network, which generates ultrasound from audios.**

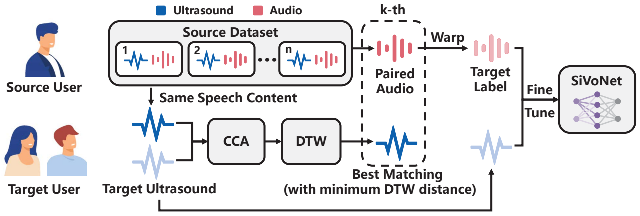


**Figure 7: The workflow of target user adaptation design, which utilizes audio of the source user data to achieve label-free model fine-tuning.**

structure rather than general convolution to upsample audio parameters (MFCC) to reconstruct ultrasound parameters (amplitude and phase) with much smaller temporal resolution. Finally, we use a Bi-directional LSTM layer [23], sandwiched between two linear projections, to jointly predict the differential change in amplitude and phase of the ultrasound, i.e. the first 8 dimensions of the output represent the phase and the remaining 8 dimensions correspond to the amplitude.

$G$ also uses the Triplet loss but adversarially swaps the positions of the positive and negative samples, i.e. the input is $(M, U^g, U^r)$. Since the generator $G$ is tasked to not only fool the discriminator $D$, but also to get as close to the ground truth as possible, we mix the triplet loss with a traditional loss (L1 distance) to encourage alignment, which has been considered beneficial in previous studies [29]. The final loss function for $G$ is shown in Equation 8:

$$\begin{aligned}&\mathcal{L}_G(G,D) = \mathbb{E}_{M,U^g,U^r\sim p_{\text{data}}(M,U^g,U^r),N\sim p_N}\\ &[\mu(\|f_a(M) - f_u(G(M,N))\|_2^2 - \|f_a(M) - f_u(U^r)\|_2^2 + \alpha)_+\\ &+\lambda\|G(M,N) - U^r\|_1],\end{aligned} \tag{10}$$

where $\mu$ and $\lambda$ represent the coefficients of the Triplet loss term and the L1 loss term respectively, and $G : (M, N) \rightarrow U^g$.

## 6 USER ADAPTATION

In this section, we illustrate our approach to improving the user adaptability of SVoice. Since the differences in the timbre of each individual, a substantial quantity of training data from new users are necessary for fine-tuning the model, which is unsuitable for deployment in the practical world. Moreover, aphasics are incapable of providing audible speech as the label. To address this issue, our critical insight is that audios are rich in individual characteristic information (e.g., timbre), whereas ultrasound signals reflected by articulatory gestures focus on the semantic information inherent in speech, which is relatively constant between individuals. This encourages us to fine-tune the model by collecting unlabeled ultrasound signals from target users and enabling labeled data from the existing source training dataset. Figure 7 illustrates the workflow.

Denote the source dataset as $\mathcal{D}^s = \{(U_1^s, A_1^s), \ldots, (U_n^s, A_n^s)\}$, where $U^s$ represents the ultrasound signal and $A^s$ represents the audio signal aligned with the $U^s$ timing. For $U^t$ collected from target users, we cannot use traditional fine-tuning schemes due to the lack of time-aligned audio label $A^t$. Thus, we turn our attention to $\mathcal{D}^s$. From $\mathcal{D}^s$, we select $A^s$ that has the same speech content with $A^t$ as the label of $U^t$, since the timbre of the reconstructed audio by SVoice relies on labels rather than ultrasound signals. However, although the content of the audio is the same, $A^s$ and $A^t$ are temporally aligned in time series due to distortions caused by differences in individual speaking rates. The neural network predicts $\hat{A}^t$ from $U^t$, and the loss existing between $\hat{A}^t$ and $A^s$ can be expressed as:

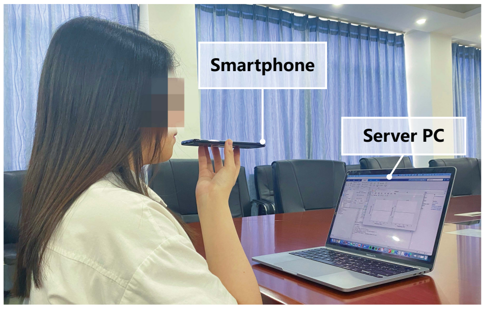
$$\begin{aligned}\mathcal{L} &= \|A^s - \hat{A}^t\|_2\\ &= \|A^t - \hat{A}^t\|_2 + \mathcal{L}_{\text{time}}\,,\end{aligned} \tag{11}$$

where $\mathcal{L}_{\text{time}}$ is the loss of timing mismatch between $A^s$ and $A^t$. The existence of $\mathcal{L}_{\text{time}}$ makes the model unable to establish the correct mapping relationship between $U^t$ and $A^t$.

To address this problem, we use Dynamic Time Warping (DTW) [49] to construct a loss function to eliminate the effects of the timing distortion loss of $A^s$ and $U^t$. Specifically, we use a pre-trained model to predict $\hat{A}^t$ with poor quality from $U^t$, and use DTW to align $A^s$ to $\hat{A}^t$ in time dimension to obtain $\bar{A}^s$. $\bar{A}^s$ and $\hat{A}^t$ are aligned in the time dimension, eliminating $\mathcal{L}_{\text{time}}$. Therefore, we have the new loss function:

$$\mathcal{L} = \|\bar{A}^s - \hat{A}^t\|_2. \tag{12}$$

However, DTW is used to achieve $A^s \rightarrow \bar{A}^s$, but due to the non-differentiability of DTW, the loss caused by DTW itself in the process of $A^s \rightarrow \bar{A}^s$ will not participate in the back-propagation of the model. Consequently, obtaining $\bar{A}^s$ still loses some accuracy due to timing matching, and we denote this loss as $\delta_A$. To reduce the accuracy drop when using DTW for $A^s$, we propose a matching



!t

**Figure 8: Experimental setup. The target sits in a normal posture and holds the smartphone naturally to record ultrasound signals during the target's speech with/without audio.**

scheme instead of randomly selecting $A^s$ from $\mathcal{D}^s$. Since the size of $\delta_A$ depends on the degree of distortion of $A^s$ and $A^t$, and $A^t$ cannot be obtained directly, we instead use the degree of timing matching loss $\delta_U$ between $U^s$ and $U^t$ as a measure of the size of $\delta_A$. Note that we can use $\delta_U$ to indirectly represent $\delta_A$ due to the complete time synchronization between the ultrasound signal and the audio signal during a recording. To better capture the difference between $U^s$ and $U^t$, we utilize Canonical Correlation Analysis (CCA) [25] to find more correlated components in the magnitude and phase of multiple subcarriers in $U^s$ and $U^t$. Then, use DTW to obtain the corresponding timing matching loss $\delta_U$. Finally, we find the $A^s$ corresponding to the minimum $\delta_U$ in the source dataset $\mathcal{D}^s$ as the label of $U^t$, thereby minimizing the $\delta_A$ loss.

## 7 IMPLEMENTATION

**Prototype:** We implement a prototype of  for comprehensive evaluation. Figure 8 shows the experimental setup. The target is asked to sit in a normal posture, naturally holding the smartphone to record the ultrasound signals during the target's speech with/without audio. Without loss of generality, we exploit a Samsung Galaxy S8 to validate the performance of  on the commercial smartphone platform. We configure and control the smartphone using the LibAS [62] development toolkit run on a PC (i.e., MacBook Pro laptop) to synchronously collect both ultrasound and audio from the bottom microphone.

**Data Collection:** In our experiments, we select 250 unique sentences covering all Chinese phonemes from the open-source dataset AISHELL-3 [53] as our speech corpus. We recruit 10 volunteers, including 5 males and 5 females, ranging in age from 19 to 25 years old, and all volunteers are native speakers of mandarin. We explicitly inform volunteers about the purpose of the experiments. To increase the diversity of the environment, all volunteers repeat each sentence at least 5 times in 3 quiet environments (i.e. laboratory, meeting room, office). Therefore, we obtain a total of 37500 data pairs with 4s length of each pair (takes about 41.7 hours). We adopt corpus to randomly split the collected dataset into the training and test sets that include 200 and 50 sentences, respectively.

**Virtual Gesture Generation:** To generate additional virtual articulatory gestures, we train the cGAN-based model using the training set. Furthermore, each volunteer provides an additional 500 audios that are not included in our corpus. cGAN is controlled to repeatedly generate gestures from each audio 15 times, after which the generated dataset is combined with the original training set to form a new training set. Note that the sentences used in the generated data do not appear in the training or test set. As a result, the final training set has 105000 data pairs with 700 unique sentences, whereas the final test set has 7500 data pairs with 50 unique sentences. Finally, is trained and evaluated using the final training/test set.

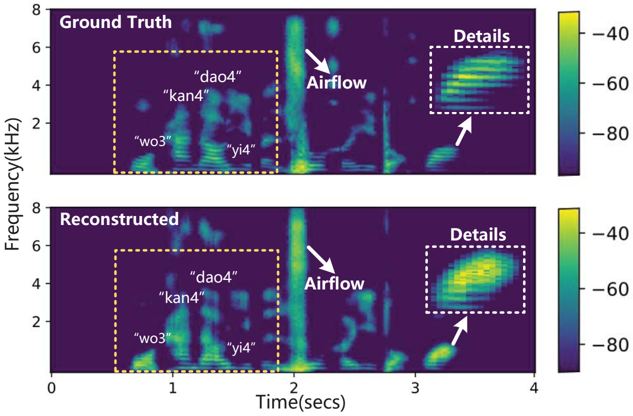


**Figure 9: The T-F spectrogram of ground truth and the reconstructed speech. The speech contents are "wo3 kan4 dao4 yi4 ben3 shu1 he2 yi4 duo3 hua1." (I saw a book and a flower.).**

**Metrics:** We characterize performance and conduct a comprehensive comparison with the state of arts from two perspectives: speech recognition accuracy and speech reconstruction quality. These can be quantified by the following four metrics:

*Character Error Rate (CER):* The minimum difference in characters between system output and baseline, which can be calculated by [67]:

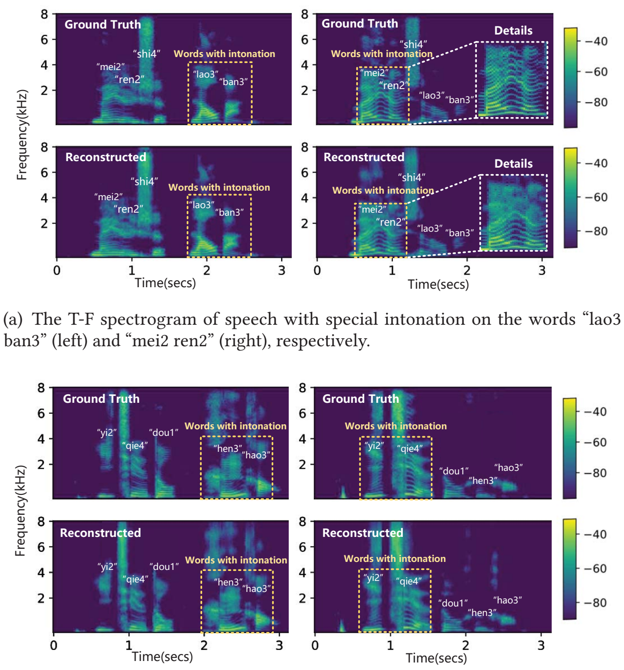
$$CER = \frac{I_c + S_c + D_c}{N_c}, \tag{13}$$

where $N_c$ represents the total number of reference characters. $I_c$, $S_c$, and $D_c$ respectively represent the minimum number of characters inserted, replaced, and deleted to convert the output of Automatic Speech Recognition (ASR) to a reference. Using the Microsoft Azure Speech-to-Text API [2], we calculate CER by comparing the reconstructed speech to the reference speech recorded with the same microphone. A lower CER indicates a higher quality of the reconstructed speech.

*STOI [58]:* Short-time objective intelligibility is a state-of-the-art speech intelligibility estimator that uses a linear correlation of temporal envelopes between reconstructed speech and ground truth, with values ranging from 0 (poor) to 1 (excellent).

*ESTOI [31]:* Extended short-time objective intelligibility ranges from 0 (poor) to 1 (excellent).

*PESQ [50]:* Perceptual evaluation of speech quality is an objective and fully referenced approach for assessing speech naturalness that creates models with mean opinion scores ranging from 1 (poor) to 5 (excellent).



(a) The T-F spectrogram of speech with special intonation on the words “lao3 ban3” (left) and “mei2 ren2” (right), respectively.

(b) The T-F spectrogram of speech with special intonation on the words “hen3 hao3” (left) and “yi2 qie4” (right), respectively.

**Figure 10: Spectrograms of speech with special intonation on various parts of the content. The speech contents in (a) and (b) are “mei2 ren2 shi4 lao3 ban3.”(No one is the boss.), “yi2 qie4 dou1 hen3 hao3.”(Everything is fine.), respectively.**

# 8 EVALUATION

In this section, we conduct comprehensive experiments in the real world to evaluate the effectiveness and robustness of .

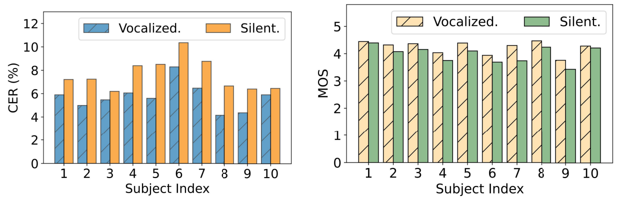
## 8.1 Overall Performance

This section focuses on verifying the effectiveness of the proposed series of techniques: i) speech reconstruction, ii) virtual gesture generation, and iii) adaptability to different users.

*8.1.1 Speech reconstruction ability.* To give an intuitive awareness of the speech reconstruction performance of , we show the T-F spectrograms of the speech synthesized by and the ground truth in Figure 9. The similarity in the overall shape of the T-F spectrum indicates that the information for reconstructing the human voice is well preserved in the articulatory gestures. The reason is that SVoice capture multi-channel contextual information using the transformer to reconstruct the high-dimensional spectrum. We annotate the position of the word at the corresponding formant and show the details of the low-frequency part. Formants are unique frequency components of the human voice. The results show that can reconstruct the low-frequency part and distinguishable formant, which indicates that our vocoder can recover natural human voice from speech parameters. In addition, we find that SVoice can even capture the airflow that usually appears in some plosives, which further confirms the feasibility of the proposed SVoice in feature extraction.

**Table 1: Objective speech quality, intelligibility, and CER for Vision-based method and SVoice.**

| Method | STOI | ESTOI | PESQ | CER |
|---|---|---|---|---|
| Lip2Wav(Vision-based) | 0.73 | 0.54 | **1.77** | 14.08% |
| SoundLip [70] | / | / | / | 12.56% |
| **SVoice(ours)** | **0.77** | **0.72** | 1.53 | **5.72%** |



(a) CER for 10 subjects when vocalized/silent.

(b) MOS for 10 subjects when vocalized/silent.

**Figure 11: The CER and MOS of each subject during vocalization/silence.**

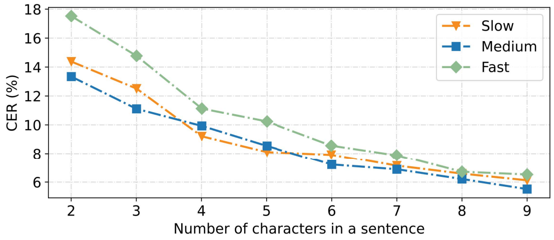


**Figure 12: The CER on sentences of different lengths at different speech rates.**

To further evaluate the ability of  to preserve user-relevant speech characteristics (i.e., intonation, speech rate), one user is asked to repeat the same sentence in various intonations. Intonation is known to represent the configuration of the energy and the pattern of pitch variation at the sentence level [44]. As shown in Figure 10, The similarity of the energy configuration and the pattern of pitch variations between the reconstructed speech and the ground truth indicates SVoice's ability to preserve intonation information. Similarly, the temporal alignment of the reconstructed speech and the ground truth in multiple repetitions demonstrates the preservation of speech rate information.

We also compare  with two state-of-the-art techniques, vision-based Lip2Wav [48] and acoustic-based SoundLip [70]. We reproduce the experimental results of Soundlip using our dataset. Note that since we are unable to obtain the corresponding vision dataset, we directly use the results claimed in the paper. From the comparisons shown in Table 1, we can see that our system achieves better intelligibility and CER for speech reconstruction than Lip2Wav. While the speech quality of is slightly lower than that of Lip2Wav. The reason is that we leverage MFCC as an intermediate feature for speech reconstruction, instead of using the directly predicted spectrum. Compared to spectrum, MFCC leaves out spectrum details,

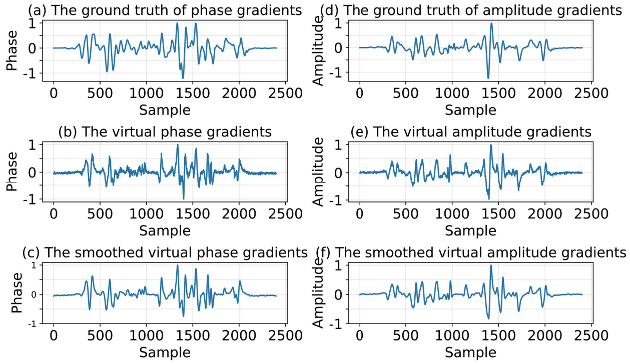
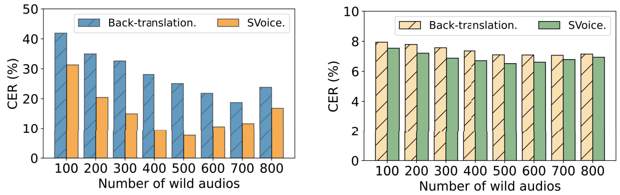


**Figure 13: The virtual gesture generated by cGAN where speech content is "fan3 zheng4 wo3 ting3 han2 xin1 de5" (Anyway, I am quite chilled).**

thus degrading the speech quality. Nevertheless, MFCC can prompt the model to focus on speech content, which benefits the accuracy and intelligibility of reconstructed speech. Also, in most cases, even if the speech quality is a bit poor, it actually does not hinder people very much from getting the key information. Notably, although we use a lightweight neural vocoder to trade off latency and quality, it is possible to reconstruct high-quality speech from MFCC with heavy neural vocoders [45]. In addition, outperforms SoundLip on CER since the latter focuses more on the task of voice command classification rather than building complex mappings.

Speaker-dependant characteristics (e.g., speech rate) and speech content (e.g., sentence length) may also affect the performance of . As shown in Figure 12, we roughly divide the speech speed into three levels: fast, medium, and slow, corresponding to faster than 260, between 160 and 260, and below 160 characters per minute. We evaluate the effect of different numbers of characters of speech content at each speech rate, where the speech content is randomly selected from our corpus. The experimental results show that the CER of  decreases as the number of content characters increases, which is expected since SiVoNet relies on contextual information. maintains a CER of less than 10% at slow or moderate speech speeds when the number of characters in a sentence is more than 4, which is sufficient for everyday conversation in Mandarin. Sentences with less than 4 characters are usually common phrases, which can be considered separately for building multitasking networks [70]. We note that although maintaining a fast speech rate (above 260/min) causes more phoneme overlap between characters resulting in a decrease in CER, SiVoNet can use contextual information to recover the correct information when the number of characters is more than 5. The above results show that  can maintain high robustness at various speech rates, especially in medium and long sentences.

Note that the above experiments are conducted in the case of target vocalization for the systematic evaluation. However, our actual goal is to reconstruct the speech of silent people, so we also analyze the effect of vocalization/silence on the ultrasound signal for speech construction. We re-collected ultrasound information for each target in silence as the silent test set. Due to the lack of time-aligned reference speech to objectively compare the quality of reconstructed speech in silence, we adopt a widely used subjective evaluation technique, i.e., Mean Opinion Score (MOS) [55], to compare the quality of speech. We recruit 20 listeners in the age range of 19 to 50. The listeners are required to evaluate the reconstructed speech quality of the test set collected during vocalization/silence, then score the speech quality on a scale of 1 (poor) to 5 (excellent), where 1 (poor) means that the audio is unintelligible and unclear, and 5 (excellent) requires natural and smooth content. Besides, the audio examples on a scale of 1 to 5 are provided to listeners as a reference for scoring. We ensure that the listener does not know the sentence's content in advance. As shown in Figure 11, the speech reconstruction quality of using the silence test set is slightly lower than that of vocalization. The average CER and MOS scores of all subjects during vocalization and silence are 5.72%, 4.23 and 7.62%, and 3.98, respectively. This is expected due to the auditory feedback effect [4], where humans unconsciously suppress the movement of vocal organs (e.g., the tongue) during silent speech. Although the performance of SVoice degrades in silence, its average CER is still lower than that of the start-of-the-art methods. In addition, by collecting additional silent data, the voiced/silent data discrepancies can be eliminated using the techniques described in section 6.



(a) Effect of the number of wild audio on unseen sentences.

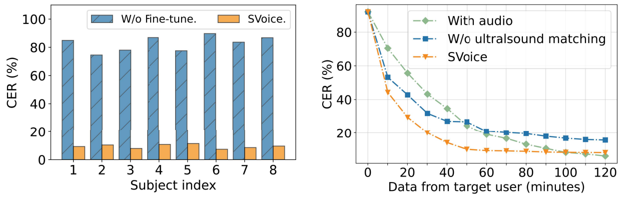
(b) Effect of the number of wild audio on seen sentences.

**Figure 14: Performance of SVoice's virtual gesture generation scheme.**

**Table 2: Objective speech quality, intelligibility and CER for Ablation Study.**

| Method | STOI | ESTOI | PESQ | CER |
|---|---|---|---|---|
| Single Frequncy | 0.61 | 0.48 | 1.33 | 45.98% |
| W/o Gradient | 0.74 | 0.65 | 1.25 | 14.57% |
| W/o Transformer | 0.72 | 0.64 | 1.51 | 13.56% |
| W/o RPE | 0.71 | 0.63 | 1.45 | 10.36% |
| W/o Two-stream | 0.74 | 0.66 | 1.27 | 9.27% |
| W/o Vocoder | 0.59 | 0.41 | 1.14 | 8.71% |
| Doppler | 0.77 | 0.67 | 1.41 | 6.41% |
| **SVoice** | **0.77** | **0.72** | **1.53** | **5.72%** |

*8.1.2 Effect of virtual gesture generation.* This experiment is designed to investigate the effectiveness of the virtual gesture generation scheme. As shown in Figure 13, the directly generated ultrasound signal is doped with a certain amount of noise compared to the ground truth. We apply a Savitzky-Golay filter [36] for high-quality speech synthesis to smooth the generated signal. The smoothed ultrasound signals act as new training data along with the original dataset to train the DNN model. To verify the superiority of the virtual samples generated by cGAN in enriching



(a) CER of diverse targets is basically stable.

(b) CER varies in different number of data for fine-tuning.

**Figure 15: Performance of SVoice's user adaptability.**

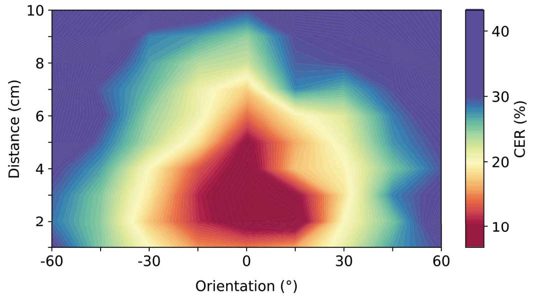


**Figure 16: CER on different distance and orientation.**

the corpus' diversity, we compare the performance of our method and the original back-translation on the various wild audios. Figure 14 displays the impact of the number of wild audios on the CER of . Obviously, the introduction of virtual samples generated by cGAN benefits speech reconstruction of both seen and unseen sentences. As the audio signals of unseen sentences increase, the CER of SVoice decreases continuously. Especially for unseen sentences, the new data generated by cGAN dramatically reduces the CER of  from 40.19% to 7.63% when the audio is up to 500. Our method outperforms back-translation with the same number of audio since cGAN can create more virtual samples from the distribution of the training set. With numerous virtual samples, however, both techniques experience a performance reduction. As the number of virtual samples rises, excessive noise is injected into the model, which is detrimental to training the model. Although the number of virtual samples created by back-translation is substantially lower than that of our technique, the poor sample quality prevents it from using a greater quantity of audio.

*8.1.3 Adaptability to different users.* We investigate adaptability to different users when they only provide ultrasound signals. Considering the variation in timbre between males and females, we only fine-tune the model using data from users of the same gender. In our experiments, one male/female user serves as the source user, while four male/female users serve as the target user. Figure 15 depicts the experimental results. demonstrates its effectiveness in user adaptation using only ultrasound data. We observe that using ultrasound matching results in a lower CER since ultrasound matching reduces the penalty associated with using temporal calibration. In addition, we compare our method to the original scheme of fine-tuning using ultrasound data with audio (i.e., changing the timbre). Compared to the original scheme, converges faster, requiring only one hour of data collection to reduce the CER to 9.42. However, as the amount of data used for fine-tuning increased, the original scheme eventually outperforms , as timing calibration errors are always present. Since our fine-tuning scheme aims to leverage unlabeled data for user adaption, we use only naive fine-tuning methods, resulting in the need for one-hour data from the new user. Advanced transfer learning [17] methods can reduce the dependence on data, which is one of our future works. Nonetheless, guarantees user adaptation using only ultrasound data, which is crucial for the usability of SSI in the aphasic.

## 8.2 Ablation Study

In this section, we conduct ablation experiments to quantitatively investigate performance in speech reconstruction. To simultaneously compare the accuracy and quality of speech reconstruction, we use the test dataset when subjects remain vocalized. We validated our approach by ablating specific components, the results are shown in Table 2.

**Single Frequency** means only a single frequency CW signal is emitted when sensing articulatory gestures. The results show that using multiple frequency subcarriers can effectively suppress multipath effects and frequency selective fading, thus improving the system's stability.

**W/O Gradient** represents that coherent demodulation's initial phase and amplitude are directly used without differential processing in the signal processing stage. The results show that differential processing can eliminate the influence of static interference and improve system performance.

**W/o Transformer** means that the transformer structure is replaced by a typical bidirectional LSTM module. The results show that using the transformer can better capture the context information in the time series and improve the sensing accuracy.

**W/O RPE** represents using the original absolute position encoding instead of the relative position encoding. The results show that the overall performance is slightly reduced, and the absolute position-coding can better represent the positional relationship of the speech temporal sequence.

**W/o Two-stream** means that we remove the dual-stream structure and stitch the differential phase and amplitude directly along the channel dimension as the input of the feature embedding module. The results indicate that the two-stream structure has a positive effect on extracting and fusing the phase and amplitude variations.

**W/o Vocoder** represents that the LPCNet Vocoder is replaced by the classic digital speech synthesis method Griffin Lim. It can be seen that compared with the digital speech synthesis method, LPCNet Vocoder greatly improves the speech quality while has a slight impact on the content accuracy.

**Doppler** represents the use of Doppler shift instead of differential phase and amplitude. The results show that using Doppler can achieve comparable performance with SVoice. However, due to the obvious difference in Doppler of Ultrasound signals in vocalized and silent, such as harmonics [57], the performance of synthesizing voice using Doppler features drops significantly in the case of remaining silent.

## 8.3 Robustness Analysis

In this section, we analyze the robustness of SVoice under different user-smartphone topologies, ambient noises, body motions, and

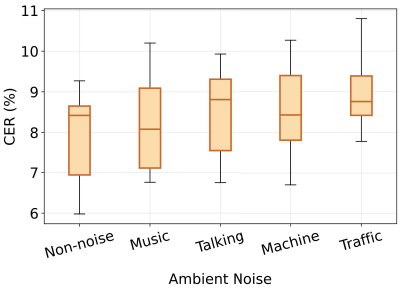


**Figure 17: CER on different ambient noise.**

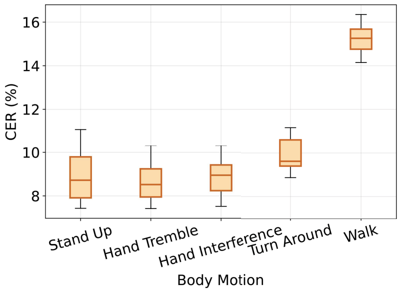


**Figure 18: CER on typical body motions.**

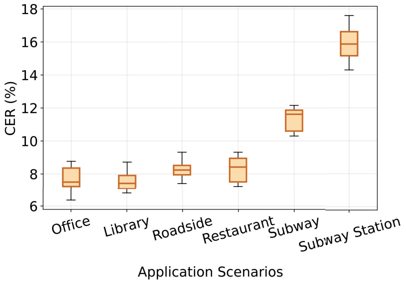


**Figure 19: CER on various scenarios.**

real application scenarios. In all experiments, subjects are asked to hold the smartphone in their hands and remain silent.

*8.3.1 Distance and Orientation.* We evaluate the performance of SVoice at different distances and orientations of the user's mouth relative to the bottom microphone of the smartphone. In this experiment, the smartphone is used at different distances (from 1cm to 10cm) and different orientations (from -60° to 60°). The experimental results are shown in Figure 16. We can see that SVoice can achieve the CER within 10% stably from 2 to 5 cm and from -15 to 15 degrees, which is consistent with the habit of human voice input. As the distance between the smartphone and the user's mouth increases, the CER of SVoice decreases continuously due to the decay of ultrasound signals in the air and the weak signal fluctuations caused by the lip movements. Nevertheless, SVoice is able to keep the CER below 20% in the range of less than 8 cm. Notably, a distance of less than 2 cm also degrades the performance, as the ultrasound signal cannot fully capture lip movements at such a close distance. Considering the fact that users are used to sending voice to a smartphone within the range of 2cm to 5cm in most cases, especially in public places, we believe that SVoice can be easily integrated into the voice applications of the smartphone and can maintain stable performance without changing users' habits.

*8.3.2 Ambient Noise.* To evaluate the performance of SVoice in different ambient noise conditions, we utilize an additional speaker placed 40 cm apart from the subject as a noise source. Our experiments involve four types of noise: music, speech, traffic, and machine noise. The noise level is controlled at around 65dB. The experimental results are depicted in Figure 17. It is obvious that SVoice is robust to different types of noise even though our training data is obtained in a controlled, quiet environment. This is because the frequency band of the noise signal in daily life is mostly lower than 10kHz, which can be successfully filtered by the coherent detector in the signal processing stage of SVoice.

*8.3.3 Body Motion.* Users may be engaged in various body motions while using the smartphone, so we evaluate the robustness of SVoice under several typical body motions, which we hypothesized to generate large signal interference. We ask four subjects to perform body motions, including 'Stand Up', 'Hand Tremble', 'Hand Interference', 'Turn Around' and 'Walk' while using SVoice in silence, where 'hand interference' means the motion interference from the other hand, and 'hand tremble' means that the user's hand shakes or moves unconsciously while holding the device. We test the CER of 4 subjects under different body motions. As shown in Figure 18, except for 'walk', SVoice achieves lower than 10% average CER for various motion artifascts of daily living. This is due to the fact that we maintain the diversity of body poses of the user's handheld device during the data collection stage. We believe the inferior accuracy for 'Walk' is caused by the complex dynamic disturbances in larger spatial degrees of freedom during walking. However, we believe there is a potential to mitigate these interferences by leveraging richer body motion datasets, and we leave it as our future work.

*8.3.4 Environmental Disturbance.* In addition to evaluating in an uncontrolled laboratory environment, we also look at performance in real-world scenarios full of uncontrollable factors. Here we consider six typical scenarios: a quiet office as a reference, a library with many people, a roadside with heavy traffic, a noisy restaurant, a crowded subway, and walking in a crowded subway station. From the experimental results displayed in Figure 19, we observe that the average CER in each scenario is 7.64%, 7.58%, 8.25%, 8.29%, 11.31%, and 15.97%, respectively. The result indicates that SVoice is robust to most open scenarios, where SVoice achieves a CER of less than 9%. We note that the performance of SVoice degrades slightly in some crowded scenarios due to interference from other individuals and worsens when coupled with the user's motion (e.g., walking). This challenge may be addressed by using neural networks or modeling to eliminate the effects of non-interesting user movements [11] and device motions [37], which is critical for wireless sensing tasks but not our main work in this paper. In a word, the results of this experiment confirm that SVoice can reconstruct accurate and audible speech from the user's silent speech in a variety of real-world scenarios, demonstrating the capability of SVoice to enable covert speech communication in public.

## 9 DISCUSSION

**Device Portability.** Different smartphones may have different layouts of speakers and microphones. For example, the bottom microphone and speaker of the Samsung S8 are on the same side, while the VIVO X20 is on each side. There are also differences in the frequency response of microphones and speakers among different smartphone models[1]. Applying the DNN model trained on the training data collected on the Samsung S8 to other devices may degrade the performance of SVoice. In order to be widely deployed on smartphones, one option is to consider collecting larger datasets containing a rich variety of phone models. Another approach is to fine-tune the model on a small number of samples collected on new devices, which has been proven to be an effective approach[70].

**System Efficiency.** We discuss the system efficiency of SVoice in terms of time and energy consumption. We implement SVoice in a client/server format, i.e., ultrasound signals acquisition and uploading are made on the Samsung Galaxy S8 smartphone, and signal processing and model inference are performed at the server. The average time delay of the signal processing is 168ms due to coherent detector and multi-frequency mechanism processing. Although the transformers can be run in parallel to improve model inference speed, the average latency is still 287ms. Currently, almost all computing tasks in SVoice are performed on the server. Extensive works have shown that using a mobile GPU/NPU can significantly reduce latency[57, 71]. We plan to design a more lightweight network to support inference on mobile devices fully. Since the function of the mobile device only needs to transmit and collect ultrasonic signals, the energy consumption level is relatively low (16.84 mAh) compared to the typical battery size of a smartphone (3000 mAh)[71]. Therefore, the current energy consumption of SVoice on mobile devices can fully meet the needs of daily life.

**Timbre and Language Adaptability.** Although SVoice provides a solution to guarantee the user adaptability of SSI to different users, it still requires new users to contribute some new data for model fine-tuning, which may not meet the expectations of general SSI. A potential solution is leveraging crowdsourcing to collect training datasets covering multiple users and train a general model using our proposed timing warping loss function. Considering that vocal timbre may a piece of potential privacy information, such as identity authentication, another possibility is to add a speaker embedding module[38] to extract speaker-independent features to train a general model that can change vocal timbre. In addition, another potential opportunity is inter-language generality. Although SVoice is only trained and evaluated on the Mandarin dataset, the correlations between speech and articulatory gestures in different languages are similar. Therefore, we believe that SVoice has inter-language potential if we collect datasets covering more languages.

**Limited Sensing Range.** The sensing range of SVoice is limited due to the fast attenuation of the acoustic signal in the air. In addition, wearing a mask has become the norm for most people outside under the influence of COVID-19. However, additional losses are incurred when ultrasound passes through the mask, resulting in a significant reduction in SVoice's performance (CER of 32.92%). To address this challenge, one of our future works is to unleash the potential of SVoice in reconstructing audible audio on other mobile devices, such as earphones [54], by utilizing ultrasound to sense the deformation of the ear canal during silent speech.

## 10 RELATED WORK

**Acoustic-based articulatory gesture sensing:** *SottoVioce*[32] converted ultrasonic images captured by ultrasonic imaging sensor into audible speech by using a two-level CNN network. It achieves high imaging accuracy but requires specialized equipment. Some works utilize ultrasound for articulatory gesture sensing based on commercial devices[10, 40, 57, 60, 68, 71]. *Zhang et al.*[68] authenticated live users by leveraging their unique articulatory gestures. However, similar authentication work focuses on the differences in users' articulatory gestures when speaking passphrases. *WaveVoice*[71] and *UltraSE*[57] fuse speech and ultrasonic features for speech enhancement, but they only use ultrasound as supplementary information. *SilentTalk*[59] is a lip-reading system based on the ultrasonic Doppler effect that enables 12 basic lip motions identification. *Endophasia*[72] distinguishes 20 silent commands by generating a two-dimensional motion profile of ultrasonic sensing signals. *EchoWhisper*[20] and *SoundLip*[70] provide word-level and sentence-level silent command recognition based on smartphones, respectively. There are three main differences between our work and recent works. First, SVoice aims to achieve a regression task that focuses on establishing mapping relations between articulatory gestures and speech, while recent works are devoted to classifying predefined words or sentences. Second, SVoice has the potential to recover speaker-dependant information from ultrasound, which is not considered in existing works. Third, we propose a data augmentation mechanism for generating ultrasound from wild audio, which can increase the scalability of the system on the unseen sentences, rather than being limited to the seen sentences.

**Silent speech reconstruction:** Vision-based approaches demonstrate superior performance in speech restoration[6, 10, 16, 27, 28, 46, 56]. *Petridis et al.* [46] present an end-to-end visual speech recognition system based on LSTM networks. *Ephrat et al.* [16] proposed a method to generate speech audio from silent video using a CNN-based model. However, vision-based solutions are constrained by lighting conditions and require cameras to monitor users' faces, which may cause privacy concerns. Wearable device-based approaches have also been proposed for speech recognition[4, 14, 19, 22, 30, 65]. *Gaddy et al.*[19] measured silent lip movements by EMG sensors to synthesize audio. *Gonzalez et al.* [22] converted articulatory gestures captured by permanent magnet articulography (PMA) into audible speech. *Anumanchpalli et al.* [4] designed a neural decoder that leverages electrocorticography (ECoG) signals to reconstruct audible speech. These works require users to wear specialized devices continuously, making them inconvenient and uncomfortable.

## 11 CONCLUSION

This paper presents SVoice, a silent speech interface that can reconstruct audible speech from silent articulatory gestures using an ultrasonic signal generated by a portable smartphone. It contributes three tailored techniques: an end-to-end model that establishes the independent mapping relationship between audible speech and signals disturbance information caused by articulatory gestures, a cross-modal data augmentation mechanism that can generate virtual articulatory gestures from widely available audio, a user adaptation technique that utilizes a small amount of unlabeled ultrasound data to fine-tune the model for different users. The extensive experiments demonstrate that SVoice has great potential to support secret communication in public and restore voice for aphasics.

## ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant No. 62122095, 62072472 and U19A2067, the National Key RD Program of China under Grant No. 2022YFF0604502, by a grant from the Guoqiang Institute, Tsinghua University and the Natural Science Foundation of Hunan Province, China under Grant No. 2020JJ2050.

## REFERENCES

[1] 2020. Best smartphones for audio. https://www.soundguys.com/best-smartphones-for-audio-16373

[2] 2022. Microsoft Azure Speech-to-Text API. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

[3] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. 2017. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2089–2093.

[4] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 7753 (2019), 493–498.

[5] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041* (2017).

[6] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* (2016).

[7] Deepak Baby and Sarah Verhulst. 2019. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 106–110.

[8] Catherine P Browman and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology* 6, 2 (1989), 201–251.

[9] Chao Cai, Zhe Chen, Jun Luo, Henglin Pu, Menglan Hu, and Rong Zheng. 2021. Boosting chirp signal based aerial acoustic communication under dynamic channel conditions. *IEEE Transactions on Mobile Computing* 21, 9 (2021), 3110–3121.

[10] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous acoustic sensing on commodity iot devices: A survey. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 432–454.

[11] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 392–405.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Randy L Diehl, Andrew J Lotto, Lori L Holt, et al. 2004. Speech perception. *Annual review of psychology* 55, 1 (2004), 149–179.

[14] Lorenz Diener, Matthias Janke, and Tanja Schultz. 2015. Direct conversion from facial myoelectric signals to speech using deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[15] Vincent Dumoulin and Francesco Visin. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).

[16] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 455–462.

[17] Chao Feng, Nan Wang, Yicheng Jiang, Xia Zheng, Kang Li, Zheng Wang, and Xiaojiang Chen. 2022. Wi-Learner: Towards One-shot Learning for Cross-Domain Wi-Fi based Gesture Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.

[18] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*. PMLR, 2031–2041.

[19] David Gaddy and Dan Klein. 2020. Digital voicing of silent speech. *arXiv preprint arXiv:2010.02960* (2020).

[20] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[21] Jon Gauthier. 2014. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester* 2014, 5 (2014), 2.

[22] Jose A Gonzalez, Lam A Cheah, James M Gilbert, Jie Bai, Stephen R Ell, Phil D Green, and Roger K Moore. 2016. A silent speech system based on permanent magnet articulography and direct synthesis. *Computer Speech & Language* 39 (2016), 67–87.

[23] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[24] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[25] Hotelling Harold. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Thomas Hueber and Gérard Bailly. 2016. Statistical conversion of silent articulation into audible speech using full-covariance HMM. *Computer Speech & Language* 36 (2016), 274–293.

[28] Thomas Hueber, Gérard Bailly, and Bruce Denby. 2012. Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. In *Interspeech 2012-13th Annual Conference of the International Speech Communication Association*. Tue–P3c.

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[30] Matthias Janke and Lorenz Diener. 2017. EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2375–2385.

[31] Jesper Jensen and Cees H Taal. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 11 (2016), 2009–2022.

[32] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[33] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille, 0.

[34] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043* (2017).

[35] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755* (2018).

[36] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojiang Chen, Zhanyong Tang, and Zheng Wang. 2021. Crossgr: accurate and low-cost cross-target gesture recognition using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–23.

[37] Jialin Liu, Dong Li, Lei Wang, Fusang Zhang, and Jie Xiong. 2022. Enabling Contact-free Acoustic Sensing under Device Motion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.

[38] Yi Liu, Liang He, Jia Liu, and Michael T Johnson. 2018. Speaker embedding extraction with phonetic information. *arXiv preprint arXiv:1804.04862* (2018).

[39] Andrew J Lotto, Gregory S Hickok, and Lori L Holt. 2009. Reflections on mirror neurons and speech perception. *Trends in cognitive sciences* 13, 3 (2009), 110–114.

[40] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1466–1474.

[41] Daniel Michelsanti and Zheng-Hua Tan. 2017. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *arXiv preprint arXiv:1709.01703* (2017).

[42] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[43] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[44] Francis Nolan. 2020. Intonation. *The handbook of English linguistics* (2020), 385–405.

[45] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[46] Stavros Petridis, Zuwei Li, and Maja Pantic. 2017. End-to-end visual speech recognition with LSTMs. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2592–2596.

[47] Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic. 2018. Visual-only recognition of normal, whispered and silent speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6219–6223.

[48] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13796–13805.

[49] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc.

[50] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.

[51] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709* (2015).

[52] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. *arXiv preprint arXiv:1606.02891* (2016).

[53] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567* (2020).

[54] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[55] Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (2016), 213–227.

[56] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 581–593.

[57] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 160–173.

[58] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 4214–4217.

[59] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[60] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. 2018. SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–18.

[61] Kristin J Teplansky, Brian Y Tsang, and Jun Wang. 2019. Tongue and lip motion patterns in voiced, whispered, and silent vowel production. In *Proc. International Congress of Phonetic Sciences*. 1–5.

[62] Yu-Chih Tung, Duc Bui, and Kang G Shin. 2018. Cross-platform support for rapid development of mobile acoustic sensing applications. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 455–467.

[63] Jean-Marc Valin and Jan Skoglund. 2019. LPCNet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5891–5895.

[64] Guanhua WANG, Yongpan ZOU, Zimu ZHOU, Kaishun WU, and Lionel M NI. 2014. We can hear you with wifi!.(2014). In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 593–604.

[65] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. 2019. Rfid tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[66] Yi Yu, Abhishek Srivastava, and Simon Canales. 2021. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–20.

[67] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.

[68] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 57–71.

[69] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. 2022. WearSign: Pushing the Limit of Sign Language Translation Using Inertial and EMG Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[70] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[71] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.

[72] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.