# UltraSR: Silent Speech Reconstruction via Acoustic Sensing

Yongjian Fu, *Student Member, IEEE,* Shuning Wang, *Student Member, IEEE,* Linghui Zhong, *Student Member, IEEE,* Lili Chen, *Member, IEEE,* Ju Ren, *Senior Member, IEEE,* Yaoxue Zhang, *Senior Member, IEEE*

**Abstract**—Silent Speech Interface (SSI) have been developed to convert silent articulatory gestures into speech, facilitating silent speech in public spaces and aiding individuals with aphasia. Prior arts of SSI, either relying on wearable devices or cameras, may lead to extended contact requirements or privacy leakage risks. Recent advancements in acoustic sensing offer new opportunitis for gesture sensing. However, they typically focus on content classification rather than on reconstructing audible speech, leading to the loss of crucial speech characteristics such as speech rate, intonation, and emotion. In this paper, we propose UltraSR, a novel sensing system that supports accurate audible speech reconstruction by analyzing the disturbance of tiny articulatory gestures on the reflected ultrasound signal. The design of UltraSR introduces a multi-scale feature extraction scheme for aggregating information from multiple views, and 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 UltraSR using a portable smartphone and evaluate it in various environments. The evaluation results show that UltraSR can reconstruct speech with a (Character Error Rate) CER as low as 5.22%, and decrease the CER from 80.13% to 6.31% 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.

**Index Terms**—Acoustic sensing, Silent Speech Interface, Human-computer interaction.

## 1 Introduction

Speech interfaces play a crucial role in facilitating both human-to-human and human-to-machine interactions. Nonetheless, speech interfaces face challenges in various scenarios, including privacy concerns in public places and soundless requirements in silent places. Silent Speech Interface (SSI) is an advanced technology that enables silent voice interaction in public by reconstructing speech from silent articulatory gestures. It not only opens a covert way for voice-based interactions but also can help people who have acquired voice disorders (e.g., laryngectomy) to regain the ability to speak.

Accurately capturing articulatory gestures is an essential prerequisite for SSI. To achieve this goal, classical SSI solutions commonly use various wearable sensors (e.g., EMG, EEG) [1]–[3] to sense the movements of vocal organs (e.g., lips, tongue). Although promising, 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 [4]–[7]. However, cameras raise privacy issues and may perform poorly in low-light conditions. In order to overcome the limitations of vision-based methods, recent advances have explored how to use wireless signals (e.g., WiFi [8] and acoustic [9]–[12] ) for articulatory gesture sensing. Nevertheless, current wireless-based solutions are unsuitable, since the original intention of inferring speech content from reflected signals for command classification instead of reconstruction of audible speech, results in the loss of some vital speech information such as speech rate, intonation, and emotion. Consequently, how to acquire a manner that can enable silent speech reconstruction using wireless signals remains an open challenge.

In this paper, we develop a Ultrasound-based Silent speech Reconstruction (UltraSR) system that can be deployed on commercial mobile devices (e.g., smartphones). UltraSR is enabled by inaudible acoustic sensing, which is easy to deploy due to ubiquitous speakers/microphones and easier to capture fine-grained vocal gestures due to the slow speed of sound in air. Figure 1 illustrates the practical use case of UltraSR, including silent speech in public and assisting aphasics to speak with voice. UltraSR employs a regression model, which maps articulatory gestures directly to audible speech via ultrasound for restoring rich speech information.

- *Corresponding author: Ju Ren*
- *Yongjian Fu, Shuning Wang, Linghui Zhong is with the School of Computer Science and Engineering, Central South University, Changsha, China.*
  *E-mail: fuyongjian, shuning.wang, zlh2021@csu.edu.cn*
- *Lili Chen, Ju Ren, Yaoxue Zhang is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China.*
  *E-mail: lilichen,renju,zhangyx@tsinghua.edu.cn*


(a) Silent speech in public. (b) Assisting aphasics to speak with voice.

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

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 [13], 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, UltraSR 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. Going further than SVoice [14], in addition to considering the phase and amplitude changes of the reflected ultrasound signals, we also contemplate incorporating speed variations into the feature space. We design a Multi-Scale Feature Extraction (MSFE) strategy to capture high-resolution features from ultrasound echoes. This strategy integrates multi-view spectrogram features of articulatory gestures, encompassing changes in distance, energy, and speed. Following this, we deploy a triple-stream convolutional structure along with multiple attention heads to extract a multi-channel context from the ultrasound signals, aiming to reconstruct 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 UltraSR 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 UltraSR, a novel Silent Speech Interface that reconstructs audible audio from silent articulatory gestures via acoustic sensing, while preseving rich speech information.
- We propose a multidimensional spectral feature aggregation scheme, which enhances the representational capabilities of ultrasound for articulatory gestures from multiple scale perspectives.
- We design a cross-modal data augmentation mechanism that can generate virtual articulatory gestures from wild audio. Such a dual form can enable widely available data from the target domain to facilitate model training, which can be readily extended to other wireless 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 UltraSR, achieving a CER of 5.22%, and decreasing the CER from 80.13% to 6.31% on new users with only 1 hour of ultrasound signals provided.

## 2 Related Work

In this section, we first introduce the research about acoustic-based articulatory gesture sensing and then discuss silent speech reconstruction.

### 2.1 Acoustic-based Articulatory Gesture Sensing

SottoVoce [15] represents an innovative approach in speech technology, which employs a two-level CNN network to transform ultrasonic images into audible speech. Despite its high imaging accuracy, relying on specialized equipment limits its widespread applicability. In contrast, several studies have explored the use of commercial devices for sensing articulatory gestures via ultrasound [16]–[21]. For instance, Zhang et al. [22] authenticate live users by analyzing their unique articulatory gestures. At the same time, other research primarily focuses on discerning users' gesture differences when uttering passphrases. WaveVoice [19] and UltraSE [20] both leverage speech and ultrasonic features for enhancement, albeit treating ultrasound more as supportive data. SilentTalk [9] is a lip-reading system using the ultrasonic Doppler effect, capable of recognizing 12 fundamental lip motions. Endophasia [12] distinguishes 20

silent commands by creating a two-dimensional motion profile from ultrasonic signals. EchoWhisper [10] and SoundLip [11] extend this technology for word-level and sentence-level silent command recognition using smartphones. Our work stands distinct from these recent studies in three key respects. Firstly, we focus on establishing mapping relations between articulatory gestures and speech through regression tasks instead of classifying predefined words or sentences. Secondly, our methodology aims to recover speaker-dependent information from ultrasound, a facet not explored in existing studies. Lastly, we introduce a novel data augmentation mechanism that synthesizes ultrasound from wild audio, thereby broadening the system's scalability to encompass unseen sentences, not just pre-recorded or predefined. This approach significantly enhances the versatility and practicality of our system, making it a promising solution for real-world applications.

### 2.2 Silent Speech Reconstruction

Vision-based techniques have been increasingly recognized for their effectiveness in speech restoration, with numerous studies demonstrating their potential [4]–[6], [21]. For instance, Petridis et al. [5] pioneered an end-to-end visual speech recognition system employing LSTM networks, while Ephrat et al. [6] developed a CNN-based model to produce speech audio from silent video footage. These methods, however, are generally constrained by environmental factors such as lighting conditions and necessitate continuous camera surveillance of the user's face, leading to potential privacy issues. Parallel to these, wearable device-based approaches have emerged as alternatives for speech recognition [1], [2], [23]. Gaddy et al. [1] utilized EMG sensors to measure silent lip movements for audio synthesis. Gonzalez et al. [24] converted articulatory gestures detected by permanent magnet articulography (PMA) into audible speech. Anumanchipalli et al. [3] created a neural decoder reconstructing audible speech from electrocorticography (ECoG) signals. While promising, these approaches often involve using specialized equipment, which can be cumbersome and intrusive for users. In contrast, UltraSR offers a more user-friendly alternative. It stands out as a non-invasive, cost-effective solution, leveraging the widespread availability of smartphones. This approach does not require users to wear specialized devices, nor does it depend on specific environmental conditions like lighting. As a result, UltraSR is more convenient and comfortable for users and addresses privacy concerns associated with continuous facial monitoring. Its adaptability to existing smartphone technology makes it a practical and accessible option for a broader user base, aiming to revolutionize the field of speech restoration and recognition.

## 3 Sensing Articulatory Gestures via Ultrasound

In this section, we introduce the possibility of using ultrasound to capture the articlatory gestures.

### 3.1 Principles of Silent Speech

Human speech is a complex process involving the coordination of various organs. The pitch of sound is a crucial aspect

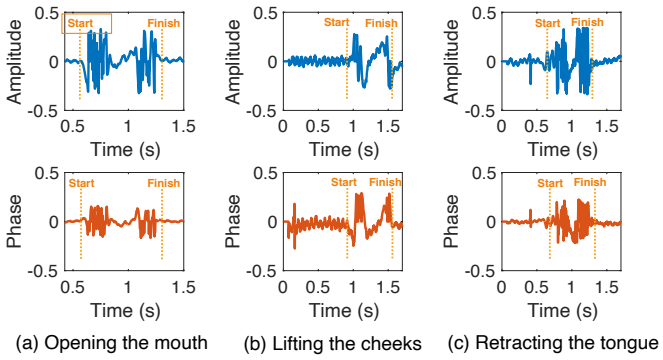



Figure 2: Capturing multiple vocal organs using CW signals, where (a), (b), and (c) is for mouth, cheek, and tongue, respectively.

of articulation and generated through unique vibration patterns of the vocal cords. Moreover, the intricate movements of several organs, such as the lips, tongue, and jaw, give rise to a range of articulatory gestures [25]. These gestures closely work with the vibrations of the vocal cord to produce a single phoneme [26], the fundamental units of speech.

In silent speech, the vocal cord vibrations are notably subdued or absent while the movements of facial muscles such as the lips, jaw, and tongue remain. These articulatory gestures are essential for silent speech interfaces as they preserve the patterns vital for producing speech. By capturing and precisely interpreting these movements, audible speech can potentially be reconstructed. This offers innovative avenues for speech communication, instrumental in situations where vocal speech is impractical or unwanted. Understanding and utilizing the intricate link between these gestures and phonemes is critical to developing silent speech reconstruction technologies.

### 3.2 Signal Design

However, articulatory gestures are very subtle and rapid, typically lasting 100-700ms [20] and moving less than 5cm [27], making it challenging to capture the fine-grained gesture motion by using a single microphone [9]. Specific commonly employed modulation signals, such as Frequency Modulated Continuous Wave (FMCW) [28] and Orthogonal Frequency Division Multiplexing [29], are less conducive to capturing the continuous articulatory movements due to the presence of signal intervals. Although triangle wave modulation [30] ensures continuous signal output, it disperses energy across various harmonics, diminishing its capacity for high-precision distance resolution. Therefore, we seek fine-grained sensing using continuous wave (CW). It is noteworthy that CW signals can continuously track articulatory gestures, ensuring no critical information is missed.

Further, we validate the capability of capturing various articulatory movements using CW signals, including actions of the mouth, cheeks, and tongue. As illustrated in Figure 2, the start and end of these movements are clearly observable in the amplitude and phase changes. Distinct characteristic patterns emerge from different organ activities, and each organ's movement is continuously captured by the CW signals throughout its activity.

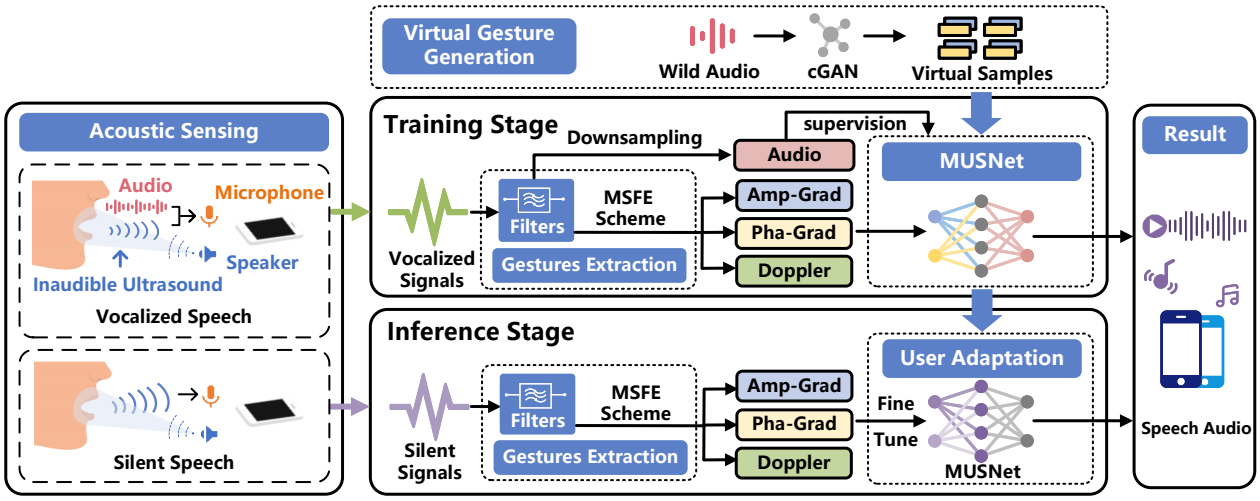



Figure 3: The system architecture of UltraSR mainly consists of acoustic sensing and gestures extraction modules, MUSNet for synthesizing speech from ultrasound, cGAN for generating virtual gestures from speech, and a fine-tuning mechanism using label-free data.

Specifically, we set the transmit signal as $Acos(2\pi ft)$, where $A$ is the amplitude and $f$ is the frequency of the signal. 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. Finally, 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. Notably, all of our design components are not tailored to any specific acoustic frequency band. The reason we select the 17kHz to 22kHz range for this paper is that it is largely inaudible to most people, thereby being more user-friendly. Furthermore, this frequency range is less susceptible to environmental interference, as the majority of acoustic noise in environments falls below 8kHz. If we expand the UltraSR to include audible frequencies (below 17kHz) or higher frequencies (above 22kHz), the UltraSR's performance is expected to benefit predictably from the increased bandwidth.

## 4 System Overview

UltraSR 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 UltraSR.

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 as well as Doppler and form training data pairs with audio. UltraSR utilizes the precise temporal alignment of ultrasound and clean audio to train the MUSNet 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 MUSNet model can be applied directly for inference. To achieve the above goals, UltraSR integrates four core components:

**Gestures Extraction** (Section 5) We design a series of signal processing to detect and segment vocal actions from echo signals, and propose a multi-scale feature extraction to aggregate information from multiple views.

**MUSNet** (Section 6) We design a tailored speech reconstruction model, named MUSNet, which can extract and fuse the phase, amplitude, and Doppler features of ultrasound and contextual information, and then quickly generate the corresponding clear and intelligible speech.

**Virtual Gesture Generation** (Section 7) 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 MUSNet to learn complicated mapping relationships and enhancing the generalization ability of MUSNet to unseen sentences.

**User Adaptation** (Section 8) 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, UltraSR 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 MUSNet model for new users.

## 5 Gestures Extraction

In this section, we detail the signal processing process, including gesture detection and extraction of features from multiple perspectives.

### 5.1 Gesture Detection

Before extracting features, we need to detect the occurrence of vocal movements to isolate pertinent data segments. To effectively identify silent speech events within echo signals, we utilize a combination of a Likelihood Ratio Test (LRT) and a Hidden Markov Model (HMM)-based event detection module. This approach is specifically designed to eliminate segments with undesirably low echo power density, thereby enhancing the accuracy and relevance of the data extracted for further analysis. This is accomplished by:

$$\begin{aligned} H_0 : absence : X = N \\ H_1 : presence : X = N + S \end{aligned} \tag{1}$$

where $S$, $N$, $X$ are Discrete Fourier Transform (DFT) coefficient vectors of reflected signals with movements, noise, and movements with noise, with their $kth$ elements $S_k$, $N_k$, and $X_k$, respectively.

Then, we have the probability density functions conditioned on $H_0$ and $H_1$:

$$\begin{aligned} p\left(X \mid H_0\right) &= \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \\ p\left(X \mid H_1\right) &= \prod_{k=1}^{L-1} \frac{1}{\pi\left[\lambda_N(k)+\lambda_S(k)\right]} \cdot \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\right\} \end{aligned} \tag{2}$$

where $\lambda_N(k)$ and $\lambda_S(k)$ denote the variances of $N_k$ and $X_k$. The likelihood ratio for the $kth$ frequency band is

$$\Lambda_k \frac{p\left(X_k \mid H_1\right)}{p\left(X_k \mid H_0\right)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{3}$$

where $\xi_k$ and $\gamma_k$ are called a priori and a posteriori SNR. The decision rule is obtained from the average likelihood ratio for each band, which is given by

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4}$$

As depicted in Figure 4, for a microphone-captured signal, our initial step is to ascertain the probability of events for each frame. Subsequently, we focus on segmenting those frames that exhibit high probabilities, specifically targeting frames that correspond to events such as articulatory gestures.

### 5.2 Multi-Scale Feature Extraction

To maximize the extraction of articulatory gestures from reflected signals, we propose a Multi-Scale Feature Extraction (MSFE) mechanism to benefit from multiple perspectives. As shown in Figure 5, we extract phase, amplitude, and Doppler from the echoes to represent changes in the distance, energy (reflected area), and velocity of the articulatory organs, respectively. This MSFE mechanism enhances the diversity of feature extraction, thereby enhancing the capability of capturing of vocal gestures.

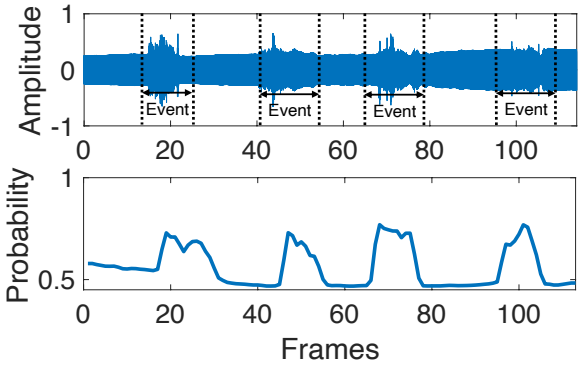



Figure 4: Gesture detection.

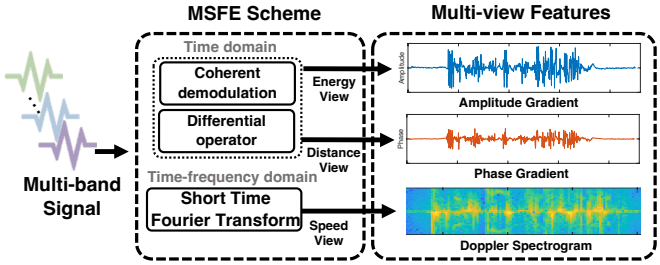



Figure 5: The Multi-scale feature extraction scheme.

#### 5.2.1 View of Distance and Energy

For the distance and energy perspectives, we design several steps for signal processing. Specifically, we fisrt separate the ultrasonic waves from the reflected signals by using 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 $cos(2\pi ft)$:

$$\begin{aligned} R_i \times cos(2_i t) &= A_i cos(2\pi ft + \phi) \times cos(2\pi f_i t) \\ &= \frac{A_i}{2} cos(4\pi ft + \phi) \times \frac{A_i}{2} cos(\phi). \end{aligned} \tag{5}$$

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

$$\begin{aligned} R_i \times sin(2_i t) &= A_i cos(2\pi ft + \phi) \times -sin(2\pi f_i t) \\ &= -\frac{A_i}{2} sin(4\pi ft + \phi) \times \frac{A_i}{2} sin(\phi). \end{aligned} \tag{6}$$

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. $R_i(t)' = R_i(t) - R_i(t-1)$.

#### 5.2.2 View of Speed

For the extraction of velocity perspective features, we use the short-time Fourier transform (STFT) to extract the time-frequency spectrum. To mitigate the impact of dominant center frequency components, we discard the information

from three bins in the central range of each subcarrier. It's important to note that our signal processing window size aligns with standard speech processing practices. Note that our signal processing window size remains consistent with speech processing. The 2D patterns are highly correlated with Vocal organs' movements. Furthermore, we define the spectrogram as the Power Spectral Density of the function:

$$\text{spec}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2 = \left| \sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-j\omega n} \right|^2 \tag{7}$$

where $x[n]$ represents input signal, while $\omega[n-m]$ denotes the overlapping Kaiser window function with an adjustable shape factor $\beta$ that enhance the resolution and minimize spectral leakage close to the sidelobes of the signal. To calculate the coefficients of the Kaiser window, we follow a specific computation method:

$$\omega[n] = \frac{I_0\left(\beta\sqrt{1-\left(\frac{n-N/2}{N/2}\right)^2}\right)}{I_0(\beta)}, 0 \le n \le N \tag{8}$$

#### 5.2.3 Correlation Between Ultrasound and Speech

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 6 shows the fluctuations of the amplitude gradients, phase gradients, and Doppler spectrum in channel 1 (17 kHz) caused by articulatory gestures. We observe a clear correspondence between ultrasound and speech. In addition, the fluctuation patterns of phase and amplitude gradients as well as Doppler 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.

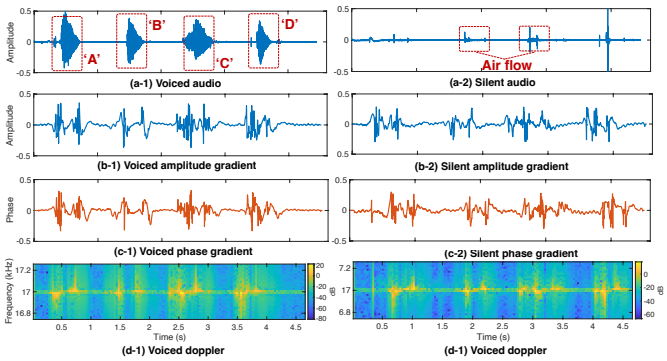



Figure 6: A subject is asked to utter "A, B, C, D" while remaining vocalized and silent, respectively. (a), (b), (c) are the speech and ultrasound signals collected during vocalization, while (d), (e), (f) are collected during silence. The fluctuations in the gradients of phase and amplitude are correlated with human speech and similar during vocalized and silent speech.

## 6 Speech Reconstructions

MUSNet is designed to reconstruct audio from ultrasound signals, as shown in Figure 7, which consists of three main modules: (1) *Feature Embedding Module*, which is to extract feature embeddings from the output of MSFE, including phase, amplitude, and Doppler fluctuations; (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.

### 6.1 Feature Embedding

The phase gradients, amplitude gradients, and Doppler of the ultrasound signal with time series length $T^u$ are sent into the feature embedding module, denoted as $U^p \in \mathbb{R}^{T^u \times C^u}$, $U^a \in \mathbb{R}^{T^u \times C^u}$ and $U^d \in \mathbb{R}^{T^u \times C^u}$ respectively, where $C^u = 8$ is determined by the number of subcarriers. The "blue" part in Figure 7 illustrates the basic structure of the module. Considering the different noise and fluctuation patterns of phase and amplitude, MUSNet first applies a dual-stream structure with 1D convolution Resblocks [31] 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 align with speech features' temporal resolution. Notably, the dual streams are trained independently without sharing weights. Then, MUSNet concatenates the outputs of the triple 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.

### 6.2 Speech Parameters Predictor

The primary challenge faced by MUSNet lies in reconstructing a high-dimensional audio waveform from low-dimensional sensing data, a process often hindered by significant information loss. To tackle this, MUSNet shifts its focus from directly predicting temporal waveforms to leveraging intermediate representations of speech features, precisely the Mel-frequency cepstral coefficients (MFCC) [32]. This approach simplifies the task by mapping feature embeddings at each time slot to their corresponding MFCC, which encapsulate vital amplitude envelope information of the audio signal in the time-frequency domain while discarding the phase information. However, the sensing feature at an individual time slot often falls short of accurately predicting MFCC owing to its limited information scope. To overcome this, we implement a bi-directional Transformer encoder [33]. Its multi-headed attention mechanism is adept at extracting contextual information from the feature embeddings across multiple dimensions and extending over a broader time horizon. This advanced method enhances

the accuracy and efficiency of MUSNet in reconstructing complex audio waveforms from limited sensing data.

Treat the output $U^o = (U_1^o, \ldots, U_{T^m}^o)$ of the feature embedding module as the input, where $U_i^o \in \mathbb{R}^d, i = 1, \ldots, T^m$ 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). MUSNet 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:

$$Z_i = \sum_j^{T^m} \alpha_{i,j}(U_j^o W^V + p_{i,j}^V), \tag{9}$$

where

$$\alpha_{i,j} = \text{softmax}(\frac{(U_i^o W^Q)(U_j^o W^K + p_{i,j}^K)^{\mathrm{T}}}{\sqrt{d/h}}). \tag{10}$$

$W^Q$, $W^K$, $W^V$ 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 $U_i^o$ and $U_j^o$ within a clipping distance $k$, which is only related to the relative distance of $i$ and $j$ and the learnable parameters $\omega$. $p_{(i,j)}^K$ and $p_{(i,j)}^V$ can be written as:

$$p_{i,j}^K = \omega_{clip(j-i,k)}^K, \tag{11}$$

where

$$\text{clip}(x, k) = \max(-k, \min(k, x)). \tag{12}$$

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.

### 6.3 Vocoder

The vocoder module is designed to reconstruct the speech waveform in the time domain from speech parameters, a process that critically influences the quality of the synthesized speech. Traditionally, vocoders are built upon the classical source-filter model of the human vocal mechanism, which conceptualizes speech production as two distinct and independent processes: the excitation source and the vocal tract response. While traditional vocoders relying on digital signal processing are fast, they often yield poor speech quality due to oversimplified speech modeling. On the other hand, vocoders fully based on neural networks produce high-quality speech but lack real-time processing capability. To strike a balance between synthesis speed and quality, we employ LPCNet [34] as our vocoder. LPCNet uniquely

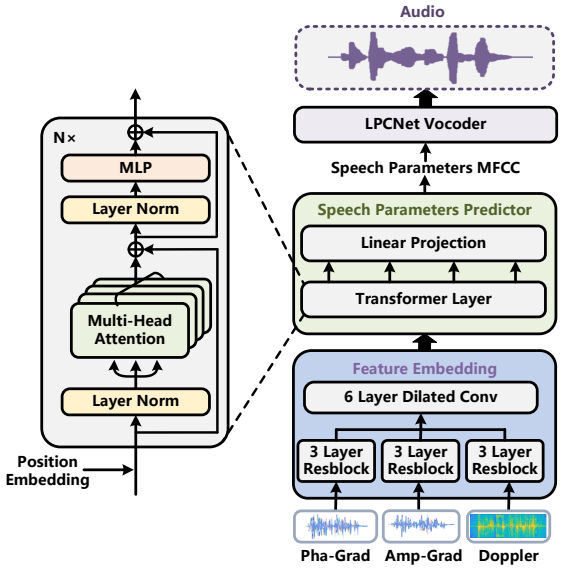



Figure 7: The structure of MUSNet to perform exhaustive extraction of speech information from ultrasound.

combines the nonlinear fitting prowess of neural networks with the simplicity of Linear Predictive Coding (LPC) filters. By using LPC filters to simplify vocal tract modeling and neural networks to fit the excitation sources, LPCNet effectively reduces the complexity of acoustic modeling. The generation process for the speech sample point $R_i$ at time $t$ is as follows:

$$R_t = \sum_{k=1}^{M} \beta_k s_{t-k} + e_t, \tag{13}$$

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 8, LPCNet is divided into three modules, where the LPC filter module computes $\beta$ from speech parameters. Notably, we opt for MFCC as LPCNet's input to reconstruct the time-domain waveform, rather than Bark-frequency cepstral coefficients (BFCC), which have limited low-frequency range and can lead to inaccurate estimations. The frame rate network, consisting of two convolutional and fully connected layers, provides a conditional vector input to the sample rate network. The sample rate network, forming the core of LPCNet, employs two GRU layers to autoregressively predict $e_t$.

## 7 Virtual Gesture Generation

In this section, we introduce how to generate virtual gesture from wild audio, where cGAN is used as the cross-modal generator.

### 7.1 Intuitive Insight

Establishing complex mapping relationships between speech and ultrasound signals necessitates substantial amounts of paired training data, leading to impractically high data collection costs. Therefore, inspired by back-translation from machine translation [35]–[38], we design

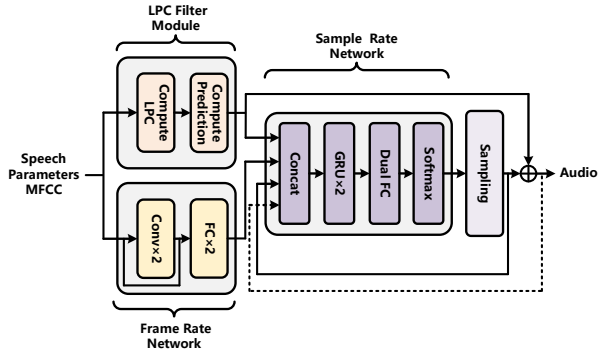



Figure 8: The structure of LPCNet Vocoder, which synthesizes speech from MFCC.

a novel virtual gesture generation scheme that generates ultrasound signals from wild audio.

In machine translation tasks, back-translation hinges on the concept that abundant pseudo-data pairs can be synthesized by training a model to convert from the target to the source language, particularly when the target corpus is more accessible [39]. In a similar vein, leveraging the ease of collecting audio data from everyday activities (e.g., phone calls or voice chats), we can develop a reverse model to generate ultrasound signals from audio. This approach allows for the creation of additional data pairs using solely new audio. Our key insight is that fabricating virtual gesture data from audio enhances the diversity of sentence sets and acts as a denoising training method. This helps in clarifying ambiguities in forward models, ultimately leading to improved generalization performance.

It's important to note that, in contrast to prior work in sign language translation [40], which implemented "back translation" by rearranging words in the original sentence, our approach focuses on generating entirely new sentences. We posit that this novel insight holds potential for broader application in various sensing tasks, offering a substantial reduction in data collection overhead.

## 7.2 Model Design

The transformative capability of cGAN (Conditional Generative Adversarial Network) [41] across domains has been widely demonstrated in fields like image generation [41]–[43], speech enhancement [44]–[46], and melody synthesis [47]. We present our cGAN framework for audio-to-ultrasound cross-domain translation in Figure 9. This framework comprises two pivotal elements: a generator $G$ and a discriminator $D$. Consider $(U, M)$ as a pair of an ultrasound vector and its corresponding audio vector. The model, augmented with additional audio information $M$, trains the generator $G$ to mimic the real distribution of the target ultrasound data $U^r$, while the discriminator $D$ aims to differentiate the authentic data $U^r$ from the generated pseudo-data $U^g$ produced by $G$. The objective function $V(G, D)$ is defined as follows:

$$\min_G \max_D V(D, G) = \mathcal{L}_D(G, D) + \mathcal{L}_G(G, D). \quad (14)$$

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

### 7.2.1 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 [19], [20], we use the Siamese network [48] and Triplet loss [49] 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 gradients, amplitude gradients, and Doppler 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:

$$\begin{aligned}&\mathcal{L}_{\boldsymbol{D}}(G, D) = \mathbb{E}_{M,U^r,U^g \sim p_{\text{data}}(M,U^r,U^g)} \\ &[\|f_a(M) - f_u(U^r)\|_2^2 - \|f_a(M) - f_u(U^g)\|_2^2 + \alpha]_+,\end{aligned} \quad (15)$$

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.

### 7.2.2 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. The feature extractor and ultrasound predictor modules are developed in G based

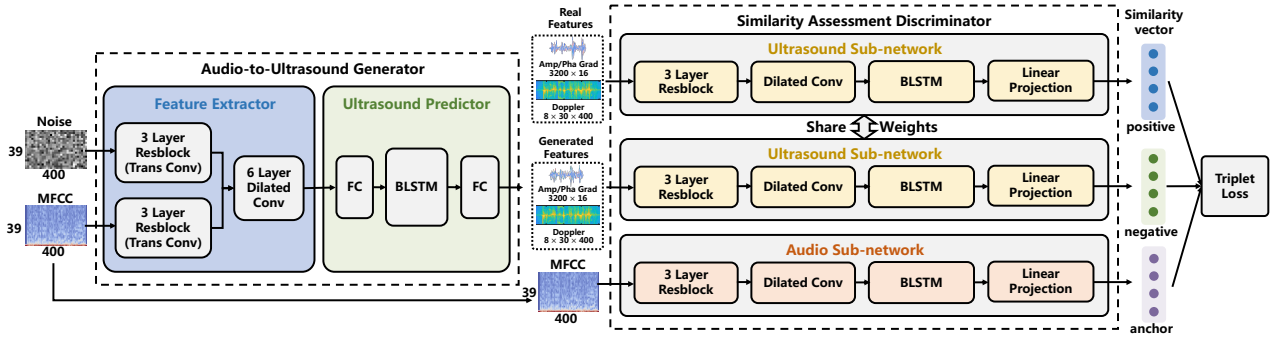



Figure 9: The architecture of cGAN-based virtual gesture generation network, which generates ultrasound from audios.

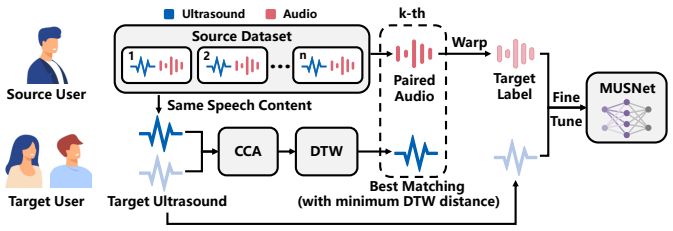



Figure 10: The workflow of target user adaptation design, which utilizes audio of the source user data to achieve label-free model fine-tuning.

on our intuition that pairwise tasks can use a similar model structure. In order to extract features, we likewise apply a Resblock-based triple-stream structure on the input sequence and then feed the concatenated output into a 6-layer 1D dilated convolution. In particular, contrary to MUSNet (Section 6), we utilize transposed convolution [50] in the Resblock structure rather than general convolution to upsample audio parameters (MFCC) to reconstruct ultrasound parameters (amplitude, phase and Doppler) with much smaller temporal resolution. Finally, we use a Bi-directional LSTM layer [51], sandwiched between two linear projections, to jointly predict the amplitude gradients, phase gradients, and Doppler of the ultrasound.

$G$ 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 [52]. The final loss function for $G$ is shown in Equation 8:

$$\mathcal{L}_G(G,D) = \mathbb{E}_{M,U^g,U^r\sim p_{\text{data}}(M,U^g,U^r),N\sim p_N} [\mu(\|f_a(M) - f_u(G(M,N))\|_2^2 - \|f_a(M) - f_u(U^r)\|_2^2 + \alpha)_+ + \lambda\|G(M,N) - U^r\|_1], \tag{16}$$

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

## 8 User Adaptation

This section details our strategy to enhance the user adaptability of UltraSR. Given the variations in individual timbre, a substantial amount of training data from new users is typically required for model fine-tuning, rendering this approach impractical for real-world deployment. Furthermore, aphasics are unable to provide audible speech for labeling purposes. Our critical insight to overcome this challenge recognizes that while audio is rich in individual characteristics like timbre, ultrasound signals, reflected by articulatory gestures, predominantly convey the semantic information of speech, which remains relatively consistent across different individuals. This understanding leads us to fine-tune our model by collecting unlabeled ultrasound signals from target users while leveraging labeled data from an existing source training dataset. The workflow for this adaptation process is depicted in Figure 10.

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. We cannot use traditional fine-tuning schemes for $U^t$ collected from target users due to the lack of time-aligned audio label $A^t$. Thus, we focus on $\mathcal{D}^s$. 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 UltraSR 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:

$$\begin{aligned}\mathcal{L} &= \|A^s - \hat{A}^t\|_2 \\ &= \|A^t - \hat{A}^t\|_2 + \mathcal{L}_{\text{time}}\,,\end{aligned} \tag{17}$$

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 tackle this issue, we employ Dynamic Time Warping (DTW) [32] to develop a loss function that mitigates the timing distortion effects between $A^s$ and $U^t$. Specifically, we use a pre-trained model to initially predict a lower-quality version of $\hat{A}^t$ from $U^t$. We then apply DTW to temporally align $A^s$ to $\hat{A}^t$, resulting in a time-aligned version $\bar{A}^s$. With $\bar{A}^s$ and $\hat{A}^t$ now synchronized in the time dimension, the

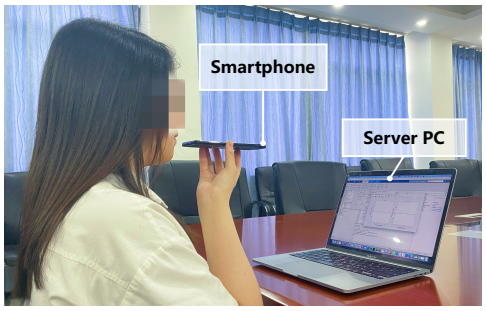



Figure 11: Experimental setup. The target sits in a normal posture and holds the smartphone naturally to record ultrasound signals during the target's speech with/without audio.

timing distortion loss, $\mathcal{L}_{\text{time}}$, is effectively nullified. This leads to the formulation of a new loss function:

$$\mathcal{L} = \|\bar{A}^s - \hat{A}^t\|_2. \tag{18}$$

To address the issue of non-differentiability in DTW when transforming $A^s$ to $\bar{A}^s$, it is essential to note that the loss induced by DTW in this process does not contribute to the model's back-propagation. This limitation of DTW can result in some loss of accuracy in obtaining $\bar{A}^s$ due to timing discrepancies, which we denote as $\delta_A$. To mitigate this accuracy reduction in the application of DTW for $A^s$, we introduce a more precise matching scheme rather than randomly selecting $A^s$ from $\mathcal{D}^s$. Given that the magnitude $\delta_A$ of depends on the extent of distortion between $A^s$ and $A^t$, and since $A^t$ cannot be directly acquired, we use the timing matching loss $\delta_U$ between $U^s$ and $U^t$ as a proxy for $\delta_A$. This indirect representation is valid due to the complete temporal synchronization between ultrasound and audio signals during recording.

To more accurately discern the differences between $U^s$ and $U^t$, we employ Canonical Correlation Analysis (CCA) [53] to identify more correlated components in the magnitude and phase of multiple subcarriers of $U^s$ and $U^t$. We then apply DTW to calculate the corresponding timing matching loss $\delta_U$. Finally, in the source dataset $\mathcal{D}^s$ , we select the $A^s$ that corresponds to the minimum $\delta_U$ as the label for $U^t$, thereby minimizing the $\delta_A$ loss.

## 9 Implementation

**UltraSR Prototype:** We implement a prototype of UltraSR for comprehensive evaluation. Figure 11 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 UltraSR on the commercial smartphone platform. We configure and control the smartphone using the LibAS [54] development toolkit run on a PC (i.e., MacBook Pro laptop) to synchronously collect both ultrasound and audio from the bottom microphone. We utilize a server equipped with an NVIDIA GTX 2080 Ti to train and test our neural network models.

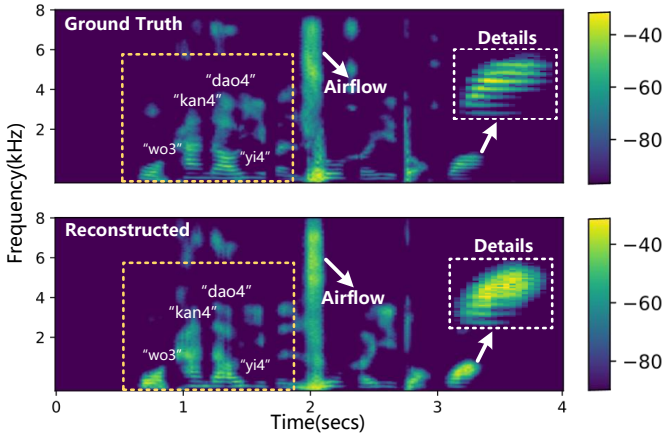



Figure 12: The T-F spectrogram of ground truth and the reconstructed speech. The speech contents are "wo3 kan4 dao4 yi4 ben3 shu1 he2 yi4 duo3 hua1." (I saw a book and a flower.).

**Data Collection:** In our experiments, we select 250 unique sentences covering all Chinese phonemes from the open-source dataset AISHELL-3 [55] 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. The experiment was approved by the Internal Review Board (IRB) of the Central South University for human subjects.

**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, UltraSR is trained and evaluated using the final training/test set.

**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 [56]:

$$CER = \frac{I_c + S_c + D_c}{N_c}, \tag{19}$$

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 [57], 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 [59]:* Extended short-time objective intelligibility ranges from 0 (poor) to 1 (excellent).

*PESQ [60]:* 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).

## 10 Evaluation

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

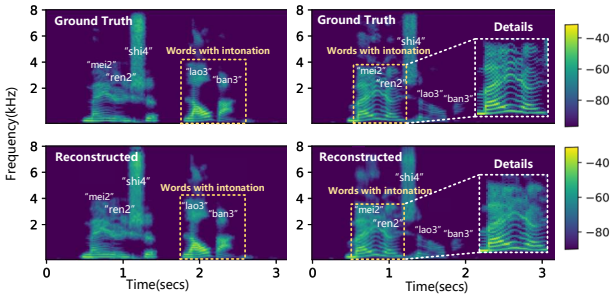
### 10.1 Overall Performance

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.

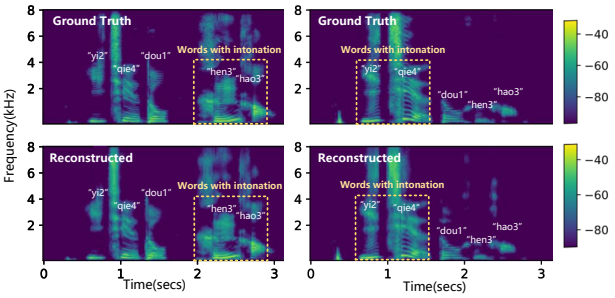
#### *10.1.1 Speech reconstruction ability*

To give an intuitive awareness of the speech reconstruction performance of UltraSR, we show the T-F spectrograms of the speech synthesized by UltraSR and the ground truth in Figure 12. 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 UltraSR 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 UltraSR 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 UltraSR can even capture the airflow that usually appears in some plosives, which further confirms the feasibility of the proposed UltraSR in feature extraction.

To further evaluate the ability of UltraSR 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 [61]. As shown in Figure 13, The similarity of the energy configuration and the pattern




(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 13: 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.

Table 1: Objective speech quality, intelligibility, and CER for baseline method and UltraSR.

| Method | STOI | ESTOI | PESQ | CER |
|---|---|---|---|---|
| Lip2Wav(Vision-based) | 0.73 | 0.54 | **1.77** | 14.08% |
| SoundLip [11] | / | / | / | 12.56% |
| SVoice | **0.77** | **0.72** | 1.53 | **5.72%** |
| **UltraSR (ours)** | **0.79** | **0.73** | 1.69 | **3.89%** |

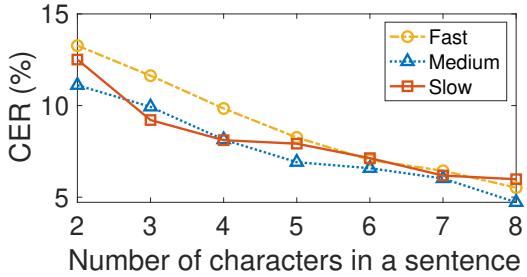
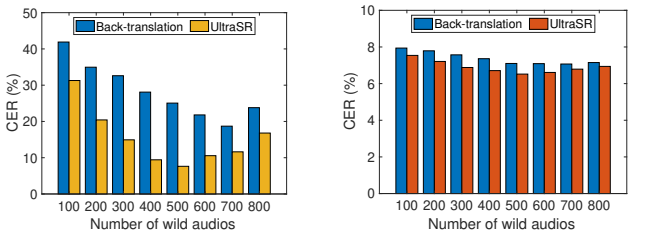



Figure 14: The CER on sentences of different lengths at different speech rates.

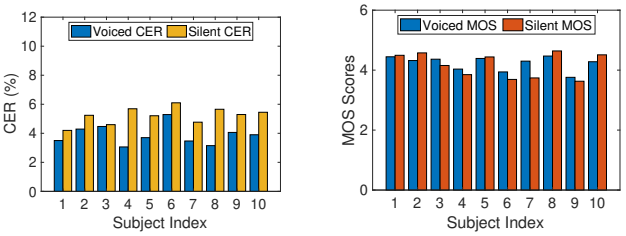
of pitch variations between the reconstructed speech and the ground truth indicates UltraSR 's ability to preserve intonation information. Similarly, the temporal alignment of the reconstructed speech and the ground truth in multiple repetitions demonstrates the preservation of speech rate information.

We also compare UltraSR with three state-of-the-art techniques, vision-based Lip2Wav [62], acoustic-based SoundLip [11], and SVoice [14]. We reproduce the exper-



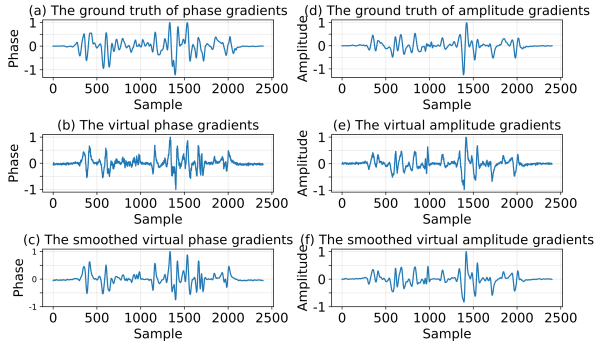

(a) Effect of the number of wild audio on unseen sentences. (b) Effect of the number of wild audio on seen sentences.

Figure 15: Performance of UltraSR's virtual gesture generation scheme.

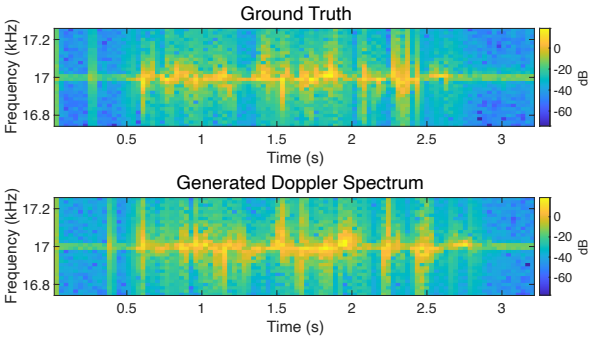



(a) CER for 10 subjects when vocalized/silent. (b) MOS for 10 subjects when vocalized/silent.

Figure 16: The CER and MOS of each subject during vocalization/silence.




(a) The generated gradients of amplitude and phase.




(b) The generated Doppler spectrum.

Figure 17: The virtual gesture generated by cGAN where speech content is "fan3 zheng4 wo3 ting3 han2 xin1 de5" (Anyway, I am quite chilled).

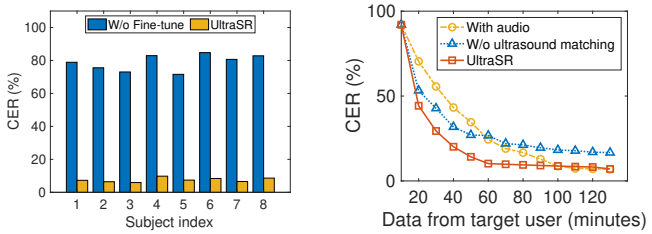
imental results of Soundlip using our dataset. Note that since we cannot obtain the corresponding vision dataset, we directly use the results claimed in the paper. From the comparisons shown in Table 1, we can see that our system UltraSR achieves better intelligibility and CER for speech reconstruction than Lip2Wav. In contrast, the speech quality of UltraSR 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, 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 poor, it does not hinder people from getting the critical information. This is evidenced by UltraSR achieving a STOI value of 0.79, which indicates excellent intelligibility since listeners are likely to understand the speech without much effort when the value is above 0.75. 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 [63]. In addition, UltraSR outperforms SoundLip on CER since the latter focuses more on the task of voice command classification rather than building complex mappings. More importantly, we find that the average CER of the UltraSR decreases from $5.72\%$ to $3.89\%$ compared with SVoice due to the multi-scale feature extraction mechanism that makes the representation of articulatory gestures more efficient. In other words, the speed of vocalization is also an indispensable feature in speech production, offering a speed perspective that provides additional information gain for capturing articulatory characteristics.

Speaker-dependant characteristics (e.g., speech rate) and speech content (e.g., sentence length) may also affect the performance of UltraSR. As shown in Figure 14, 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 UltraSR decreases as the number of content characters increases, which is expected since MUSNet relies on contextual information. UltraSR 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, sufficient for everyday conversation in Mandarin. Sentences with less than four characters are usually common phrases, which can be considered separately for building multitasking networks [11]. Although maintaining a fast speech rate (above 260/min) causes more phoneme overlap between characters, resulting in a decrease in CER, MUSNet can use contextual information to recover the correct information when the number of characters is more than 5. The above results show that UltraSR 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 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. As the silent

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

| Method | STOI | ESTOI | PESQ | CER |
|---|---|---|---|---|
| Single Frequency | 0.62 | 0.48 | 1.34 | 43.21% |
| W/o Gradient | 0.74 | 0.66 | 1.27 | 13.97% |
| W/o Transformer | 0.72 | 0.64 | 1.51 | 12.96% |
| W/o RPE | 0.71 | 0.63 | 1.45 | 9.45% |
| W/o Triple-stream | 0.74 | 0.66 | 1.27 | 8.82% |
| W/o Vocoder | 0.61 | 0.43 | 1.15 | 7.31% |
| Doppler only | 0.77 | 0.67 | 1.41 | 6.41% |
| **UltraSR** | **0.79** | **0.73** | **1.69** | **3.89%** |




(a) CER of diverse targets is basically stable.

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

Figure 18: Performance of UltraSR's user adaptability.

test set, we re-collected ultrasound information for each target in silence. 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) [64], 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 the listener does not know the sentence's content in advance. As shown in Figure 16, the speech reconstruction quality of using the silence test set is slightly lower than that of vocalization. All subjects' average CER and MOS scores during vocalization and silence are 3.89%, 4.25 and 5.22%, 4.17, respectively. This is expected due to the auditory feedback effect [3], where humans unconsciously suppress the movement of vocal organs (e.g., the tongue) during silent speech. Although the performance of UltraSR degrades in silence, its average CER is still lower than that of the start-of-the-art methods. Additionally, by collecting additional silent data, the voiced/silent data discrepancies can be eliminated using the techniques described in section 8.

#### 10.1.2 Effect of virtual gesture generation

This experiment is designed to investigate the effectiveness of the virtual gesture generation scheme. As shown in Figure 17, the directly generated ultrasound signal is doped with a certain amount of noise compared to the ground truth. We apply a Savitzky-Golay filter [65] 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 the corpus' diversity, we compare the performance of our method and the original back-translation on the various wild audios. Figure 15 displays the impact of the number of wild audios on the CER of UltraSR. Introducing 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 UltraSR decreases continuously. Especially for unseen sentences, the new data generated by cGAN dramatically reduces the CER of UltraSR from 40.19% to 7.55% 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 more audio.

#### 10.1.3 Adaptability to different users

We investigate UltraSR's 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 18 depicts the experimental results. UltraSR 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, UltraSR converges faster, requiring only one hour of data collection to reduce the CER to 6.31%. However, as the amount of data used for fine-tuning increased, the original scheme eventually outperforms UltraSR, as timing calibration errors are always present. Since our fine-tuning scheme aims to leverage unlabeled data for user adaptation, we use only naive fine-tuning methods, resulting in the need for one hour of data from the new user. Advanced transfer learning [66] methods can reduce the dependence on data, which is one of our future directions. In addition, some recent works combine new sensing frameworks with large language models (LLM) to bring opportunities for open-world SSI. Specifically, LLM can create diverse user datasets [67] or serve as evaluation indicators to transfer knowledge to the SSI model [68]. We firmly believe that integrating UltraSR as part of LLM applications will bring new perspectives in the future.

### 10.2 Ablation Study

In this section, we conduct ablation experiments to investigate performance in speech reconstruction quantitatively. To simultaneously compare the accuracy and quality of speech reconstruction, we use the test dataset when subjects remain

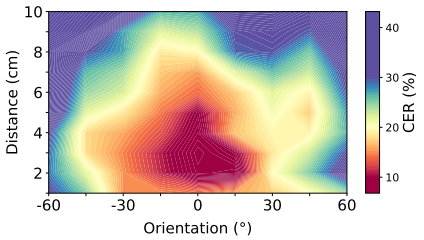



Figure 19: CER on different distance and orientation.

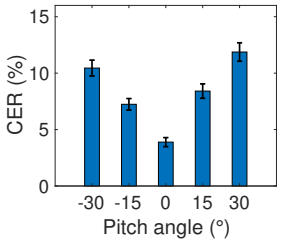



Figure 20: CER on various pitch angles.

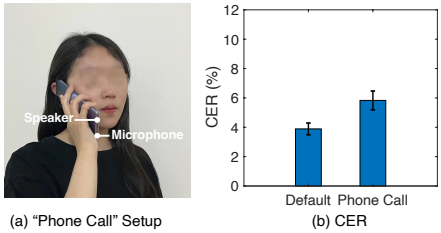



Figure 21: Impact of holding style.

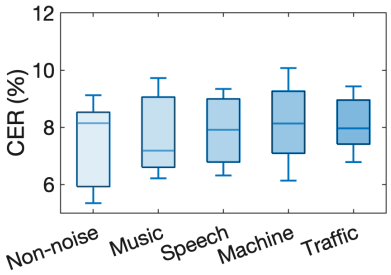



Figure 22: CER on different ambient noise.

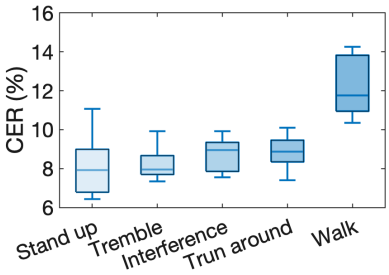



Figure 23: CER on typical body motions.

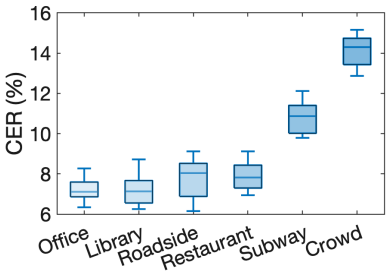



Figure 24: CER on various scenarios.

vocalized. We validate 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 a typical bidirectional LSTM module replaces the transformer structure. The results show that the Transformer can better capture the context information in the time series and improve the sensing accuracy.

**W/O RPE** uses 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 Triple-stream** means that we remove the triple-stream structure and stitch the differential phase, amplitude, and Doppler directly along the channel dimension as the input of the feature embedding module. The results indicate that the triple-stream structure positively affects extracting and fusing the phase, amplitude, and Doppler 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 speech quality while having a slight impact on content accuracy.

**Doppler** represents the use of Doppler shift instead of differential phase and amplitude. The results show that using Doppler cannot achieve the best performance because the information loss from a single perspective.

### 10.3 Robustness Analysis

In this section, we analyze the robustness of UltraSR under different user-smartphone topologies, holding style, ambient noises, body motions, and real application scenarios. In all experiments, subjects are asked to hold the smartphone and remain silent.

#### *10.3.1 Distance and Orientation*

We evaluate the performance of UltraSR 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 19. We can see that UltraSR can achieve the CER within $10\%$ stably from $2$ $8cm$ and $-15$ $15$ 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 UltraSR decreases continuously due to the decay of ultrasound signals in the air and the weak signal fluctuations caused by articulatory gestures. Nevertheless, UltraSR can keep the CER below $20\%$ in the range of less than $8cm$. Notably, a distance of less than $2cm$ also degrades the performance, as the ultrasound signal cannot fully capture articulatory gestures at such a close distance.

Furthermore, we also verified that UltraSR is robust to pitch angles. In this scenario, we keep the bottom of the phone $5cm$ away from the mouth and change the pitch angle between it and the mouth in the range of $-30$ to $30$ degrees. The experimental results are shown in the figure 20, which is consistent with our expectation that as the angle range increases, the system performance degrades. Our system can maintain stable performance within $-15$ to $15$ degrees, with a CER of less than $10\%$.

Considering the fact that users are used to sending voice to a smartphone within the range of $2cm$ to $8cm$ within

an angular offset of 15 degrees in most cases, especially in public places, we believe that UltraSR can be easily integrated into the voice applications of the smartphone and can maintain stable performance without changing users' habits.

#### 10.3.2 Holding Style

Users may have more than one gesture of holding the phone. A typical situation is the holding style of making a phone call, as shown in Figure 21. We evaluate the UltraSR's performance in this scenario by additionally collecting paired data from one user while simulating a phone call. The results show that our system can effectively capture vocal gestures and reconstruct speech in this scenario, with a CER of 5.83%. This motivates us to implement some exciting applications, such as silent phone calls, which is very useful for confidentiality scenarios.

#### 10.3.3 Ambient Noise

To assess the performance of UltraSR under varying ambient noise conditions, we employed an additional speaker positioned 40 cm away from the subject, serving as a noise source. Our experiments incorporated four distinct types of noise: music, speech, traffic, and machine noise, each maintained at a noise level of approximately 65 dB. The outcomes of these experiments are illustrated in Figure 22. Notably, UltraSR demonstrates remarkable robustness to these diverse noise types despite our training data being acquired in a controlled, quiet setting. This resilience is largely attributed to the frequency band of everyday noise signals typically falling below 10kHz, a range that UltraSR's coherent detector effectively filters out during the signal processing phase.

#### 10.3.4 Body Motion

To understand how UltraSR performs amidst common body movements that users might engage in while using a smartphone, we evaluate its robustness against several typical motions anticipated to produce significant signal interference. Four participants are asked to perform movements such as Stand Up, Hand Tremble, Hand Interference, Turn Around, and Walk while using UltraSR in silence. Here, Hand Interference' refers to disruptions caused by the other hand, while Hand Tremble' describes involuntary shakes or movements of the hand holding the device. We assess each subject's Character Error Rate (CER) under these varied motions. As Figure 23 indicates, UltraSR maintains an average CER below 10% for all tested motions except Walk', demonstrating its effectiveness amidst daily life activities. This resilience can be attributed to the diversity in body postures and handheld device orientations captured during data collection. The relatively higher CER observed during 'Walk' is likely due to more complex dynamic disturbances caused by extensive body movements. We recognize that addressing these disturbances may be possible by incorporating a broader dataset of body motions, which we aim to explore in our future work.

#### 10.3.5 Environmental Disturbance

To thoroughly assess UltraSR beyond the confines of a controlled laboratory setting, we also examine its performance in real-world environments teeming with unpredictable factors. We select six typical scenarios for this purpose: a quiet office as a baseline, a busy library, a roadside with heavy traffic, a noisy restaurant, a running subway, and walking through a crowded subway station. The experimental outcomes, as shown in Figure 24, reveal that the average CER in these scenarios is respectively 7.22%, 7.21%, 7.75%, 7.90%, 10.81%, and 14.11%. These results demonstrate that UltraSR maintains robust performance in most open environments, achieving a CER of less than 9% in most scenarios. It is noteworthy, however, that UltraSR's performance slightly declines in exceptionally crowded settings, likely due to external interference from surrounding individuals, and more notably when combined with user movement, such as walking. Addressing these challenges could involve the use of neural networks or advanced modeling techniques to filter out irrelevant user and device motions [69], [70], which, while crucial for wireless sensing tasks, falls outside the primary scope of this paper. Overall, the experimental results validate UltraSR's ability to accurately reconstruct audible speech from a user's silent speech across various real-life situations, underscoring its potential for facilitating silent speech in public spaces.

### 10.4 System Efficiency.

We evaluate the UltraSR's efficiency from two perspectives: computational time and energy consumption. Regarding computational time efficiency, the UltraSR is structured in a client/server format: the Samsung Galaxy S8 smartphone is responsible for the acquisition and uploading of ultrasound signals, while signal processing and model inference take place on the server. In this setup, the server provides remote voice conversion services to the user's smartphone. Once the server receives the uploaded data from the user, it requires an average time delay of 214ms to process the signal. This includes operations such as event detection (46ms), filtering (21ms), and Fourier transform (147ms). After the signal processing is complete, the neural network inference consumes an additional 287ms of time overhead. However, numerous studies have indicated that leveraging a mobile GPU/NPU can markedly decrease latency [19], [20]. We aim to develop a lighter network that supports full inference on mobile devices. Considering the mobile device's role is mainly to transmit and collect ultrasonic signals, its energy consumption is relatively low (16.84 mAh), especially when compared to the typical battery capacity of a smartphone (3000 mAh) [19]. Therefore, UltraSR's current energy usage on mobile devices is sufficiently low for everyday use, making it a viable solution for daily applications.

## 11 DISCUSSION

**Device Portability.** Smartphones often exhibit variability in the configuration of speakers and microphones. For instance, the Samsung S8 has its bottom microphone and speaker on the same side, whereas the VIVO X20 features them on opposite sides. Additionally, the frequency response of microphones and speakers can differ significantly

among various smartphone models [71]. Consequently, deploying a DNN model trained on data from the Samsung S8 onto other devices might result in a decline in UltraSR's performance. One strategy could be to compile larger datasets encompassing a wide range of phone types to facilitate broader implementation across diverse smartphone models. Alternatively, fine-tuning the model with a small data set from new devices is another effective method, as demonstrated in previous research [11]. This approach allows for model adaptation to the specific acoustic characteristics of different smartphone models, enhancing the versatility and effectiveness of UltraSR in various hardware contexts.

**Timbre and Language Adaptability.** While UltraSR offers a tailored solution to enhance user adaptability in SSI across different individuals, new users still need to provide some data for model fine-tuning. This requirement may not align with the broader expectations for general SSI usability. One viable approach to address this is through crowdsourcing, aiming to gather training datasets encompassing diverse users. Subsequently, a universal model could be developed using our proposed timing warping loss function. Another aspect to consider is the potential privacy concerns associated with vocal timbre, which could be used for identity authentication. A possible solution here is the integration of a speaker embedding module [72]. This module would focus on extracting speaker-independent features, thereby enabling the training of a generic model capable of altering vocal timbre. Furthermore, there is an opportunity to explore inter-language generality with UltraSR. Although it has been primarily trained and assessed on Mandarin data, the underlying relationships between speech and articulatory gestures are similar across different languages. Therefore, we believe that UltraSR could potentially extend its applicability to multiple languages if we can ensure the availability of datasets that span a broader linguistic spectrum.

**Limited Sensing Range.** The sensing range of UltraSR is limited due to the fast attenuation of the acoustic signal in the air. To address this challenge, one possible solution is leveraging multiple microphones(i.e., both top and bottom microphones) to enhance the signal-to-noise ratio (SNR) on mobile devices.

## 12 CONCLUSION

This paper presents UltraSR, a silent speech interface that can reconstruct audible speech from silent articulatory gestures using an ultrasonic signal generated by a portable smartphone. It contributes four tailored techniques: a multi-scale feature extraction scheme that aggregates information from multiple perspectives, 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 UltraSR has great potential to support silent speech in public and restore voice for aphasics.

## REFERENCES


[1] D. Gaddy and D. Klein, "Digital voicing of silent speech," *arXiv preprint arXiv:2010.02960*, 2020.

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

[3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[5] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2592–2596.

[6] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 455–462.

[7] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6219–6223.

[8] G. WANG, Y. ZOU, Z. ZHOU, K. WU, and L. M. NI, "We can hear you with wifi!.(2014)," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 593–604.

[9] J. Tan, C.-T. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[10] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin, "Echowhisper: Exploring an acoustic-based silent speech interface for smartphone users," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–27, 2020.

[11] Q. Zhang, D. Wang, R. Zhao, and Y. Yu, "Soundlip: Enabling word and sentence-level lip interaction for smart devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–28, 2021.

[12] Y. Zhang, W.-H. Huang, C.-Y. Yang, W.-P. Wang, Y.-C. Chen, C.-W. You, D.-Y. Huang, G. Xue, and J. Yu, "Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–26, 2020.

[13] A. J. Lotto, G. S. Hickok, and L. L. Holt, "Reflections on mirror neurons and speech perception," *Trends in cognitive sciences*, vol. 13, no. 3, pp. 110–114, 2009.

[14] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang, "Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 622–636.

[15] N. Kimura, M. Kono, and J. Rekimoto, "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*, 2019, pp. 1–11.

[16] S. Wang, L. Zhong, Y. Fu, L. Chen, J. Ren, and Y. Zhang, "Uface: Your smartphone can" hear" your facial expression!" *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–27, 2024.

[17] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–18, 2018.

[18] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1466–1474.

[19] Q. Zhang, D. Wang, R. Zhao, Y. Yu, and J. Shen, "Sensing to hear: Speech enhancement for mobile devices using acoustic signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–30, 2021.

[20] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 160–173.

[21] C. Cai, R. Zheng, and J. Luo, "Ubiquitous acoustic sensing on commodity iot devices: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 432–454, 2022.

[22] L. Zhang, S. Tan, and J. Yang, "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*, 2017, pp. 57–71.

[23] J. Wang, C. Pan, H. Jin, V. Singh, Y. Jain, J. I. Hong, C. Majidi, and S. Kumar, "Rfid tattoo: A wireless platform for speech recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, pp. 1–24, 2019.

[24] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[25] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

[26] R. L. Diehl, A. J. Lotto, L. L. Holt *et al.*, "Speech perception," *Annual review of psychology*, vol. 55, no. 1, pp. 149–179, 2004.

[27] K. J. Teplansky, B. Y. Tsang, and J. Wang, "Tongue and lip motion patterns in voiced, whispered, and silent vowel production," in *Proc. International Congress of Phonetic Sciences*, 2019, pp. 1–5.

[28] C. Cai, Z. Chen, J. Luo, H. Pu, M. Hu, and R. Zheng, "Boosting chirp signal based aerial acoustic communication under dynamic channel conditions," *IEEE Transactions on Mobile Computing*, vol. 21, no. 9, pp. 3110–3121, 2021.

[29] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 1515–1525.

[30] Y. Zhu, S. Zhang, H. Zhao, and S. Chen, "Suppression of noise amplitude modulation interference in triangle frequency modulation detector based on frft," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 16 107–16 117, 2021.

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

[32] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[34] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[35] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *arXiv preprint arXiv:1711.00043*, 2017.

[36] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," *arXiv preprint arXiv:1710.11041*, 2017.

[37] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," *arXiv preprint arXiv:1804.07755*, 2018.

[38] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for wmt 16," *arXiv preprint arXiv:1606.02891*, 2016.

[39] ——, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[40] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "Wearsign: Pushing the limit of sign language translation using inertial and emg wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–27, 2022.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[42] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[43] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2089–2093.

[44] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[45] D. Baby and S. Verhulst, "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, 2019, pp. 106–110.

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

[47] Y. Yu, A. Srivastava, and S. Canales, "Conditional lstm-gan for melody generation from lyrics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–20, 2021.

[48] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015, p. 0.

[49] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[50] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[51] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[52] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[53] H. Harold, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[54] Y.-C. Tung, D. Bui, and K. G. Shin, "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*, 2018, pp. 455–467.

[55] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[56] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[57] "Microsoft azure speech-to-text api," 2022. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

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

[59] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

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

[61] F. Nolan, "Intonation," *The handbook of English linguistics*, pp. 385–405, 2020.

[62] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 796–13 805.

[63] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[64] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[65] X. Li, L. Chang, F. Song, J. Wang, X. Chen, Z. Tang, and Z. Wang, "Crossgr: accurate and low-cost cross-target gesture recognition using wi-fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–23, 2021.

[66] C. Feng, N. Wang, Y. Jiang, X. Zheng, K. Li, Z. Wang, and X. Chen, "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*, vol. 6, no. 3, pp. 1–27, 2022.

[67] X. Chen and X. Zhang, "Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and

generative diffusion models," in *ACM Conference on Embedded Networked Sensor Systems (SenSys '23)*. Istanbul, Turkiye: ACM, New York, NY, USA, 2023, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3625687.3625798

[68] T. Benster, G. Wilson, R. Elisha, F. R. Willett, and S. Druckmann, "A cross-modal approach to silent speech with llm-enhanced recognition," *arXiv preprint arXiv:2403.05583*, 2024.

[69] Z. Chen, T. Zheng, C. Cai, and J. Luo, "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*, 2021, pp. 392–405.

[70] J. Liu, D. Li, L. Wang, F. Zhang, and J. Xiong, "Enabling contact-free acoustic sensing under device motion," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–27, 2022.

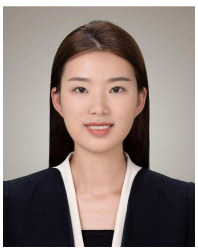
[71] "Best smartphones for audio," 2020. [Online]. Available: https://www.soundguys.com/best-smartphones-for-audio-16373

[72] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," *arXiv preprint arXiv:1804.04862*, 2018.

**Yongjian Fu** received the B.Sc. in Computer Science from Central South University in 2021, China. He is currently working toward a Ph.D. degree in Computer Science at the School of Computer Science and Engineering from Central South University. His research interests include wireless sensing, mobile computing, and Internet-of-Things.

**Shuning Wang** received the B.Sc. in Computer Science from Central South University, China. Since Sept. 2022, she has been pursuing the Ph.D. degree in Computer Science from Central South University, China. Her research interests include wireless sensing, mobile computing, and Internet-of-Things.

**Linghui Zhong** received the B.Sc. degree from Central South University, Changsha, China, in 2021. She is currently pursuing a Ph.D. at the School of Computer Science and Engineering, Central South University, China. Her research interests include Internet of Things, wireless sensing, and mobile computing.



**Lili Chen** received her Ph.D. degree from Northwest University. She is currently a Postdoctoral Research Associate in Tsinghua University. Her current research interests include wireless systems and smart IoT.

**Ju Ren** (Senior Member, IEEE) received the BSc, MSc, and PhD degrees all in computer science, from Central South University, China. Currently, he is an associate professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include Internet-of-Things, edge computing, edge intelligence, as well as security and privacy. He currently serves as an associate editor for many journals, including IEEE Transactions on Cloud Computing and IEEE Transactions on Vehicular Technology, etc. He also served as the general co-chair for IEEE BigDataSE'20, the TPC co-chair for IEEE BigDataSE'19, the publicity co-chair for IEEE ICDCS'22, the poster co-chair for IEEE MASS'18, a symposium co-chair for IEEE/CIC ICCC'2319, I-SPAN'18 and IEEE VTC'17 Fall, etc. He received many best paper awards from IEEE flagship conferences, including IEEE ICC'19 and IEEE HPCC'19, etc., the IEEE TCSC Early Career Researcher Award (2019), and the IEEE ComSoc Asia-Pacific Best Young Researcher Award (2021). He was recognized as a highly cited researcher by Clarivate (2020-2022).

**Yaoxue Zhang** (Senior Member, IEEE) received the BSc degree from the Northwest Institute of Telecommunication Engineering, Xi'an, China, in 1982, and the PhD degree in computer networking from Tohoku University, Sendai, Japan, in 1989. He is currently a professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include computer networking, operating systems, and transparent computing. He has published more than 200 papers on peerreviewed IEEE/ACM journals and conferences. He is the editor-in-chief of Chinese Journal of Electronics and a fellow of the Chinese Academy of Engineering.