UltraSE: Single-Channel Speech Enhancement Using Ultrasound

Ke Sun, Xinyu Zhang
University of California San Diego
kesun@eng.ucsd.edu,xyzhang@ucsd.edu

ABSTRACT

Robust speech enhancement is considered as the holy grail of audio processing and a key requirement for human-human and human-machine interaction. Solving this task with single-channel, audio-only methods remains an open challenge, especially for practical scenarios involving a mixture of competing speakers and background noise. In this paper, we propose UltraSE, which uses ultrasound sensing as a complementary modality to separate the desired speaker's voice from interferences and noise. UltraSE uses a commodity mobile device (*e.g.*, smartphone) to emit ultrasound and capture the reflections from the speaker's articulatory gestures. It introduces a multi-modal, multi-domain deep learning framework to fuse the ultrasonic Doppler features and the audible speech spectrogram. Furthermore, it employs an adversarially trained discriminator, based on a cross-modal similarity measurement network, to learn the correlation between the two heterogeneous feature modalities. Our experiments verify that UltraSE simultaneously improves speech intelligibility and quality, and outperforms state-of-the-art solutions by a large margin.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Hardware** → **Signal integrity and noise analysis**; **Noise reduction**.

KEYWORDS

Speech Enhancement, Ultrasound

ACM Reference Format:
Ke Sun, Xinyu Zhang. 2021. UltraSE: Single-Channel Speech Enhancement Using Ultrasound. In *The 27th Annual International Conference On Mobile Computing And Networking (ACM MobiCom '21), October 25–29, 2021, New Orleans, LA, USA*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3447993.3448626

1 INTRODUCTION

Human auditory system is remarkably capable of singling out a speech source amid a mixture of interfering speakers and noises, which remains a key challenge for machine hearing. The problem has witnessed a surge in today's digital communication systems for human-human and human-machine interaction. Examples include mobile VoIP, voice commands, post-production of live speech, *etc*.

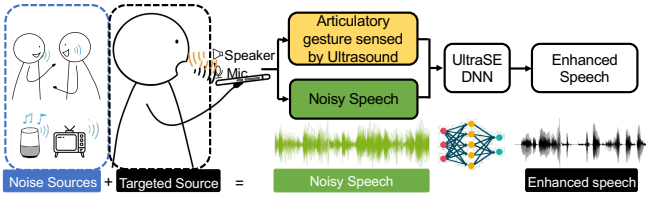


Figure 1: UltraSE targets the scenario where the user holds the smartphone to record the speech in a noisy environment. UltraSE uses ultrasound sensing as a complementary modality to separate the desired speaker's voice from interferences.

The related research problem of speech separation and enhancement (SSE) is often considered as the holy grail of audio processing.

Since the problem is inherently ill-posed, classical solutions need to rely on prior knowledge (*i.e.*, per-speaker feature engineering) [1] or directional microphone arrays [2] to isolate the desired source from ambient sounds. In the past several years, deep learning techniques have proliferated and significantly advanced the field, enabling single-microphone speaker-independent SSE [3]. State-of-the-art solutions have demonstrated around 10 dB improvement in average audio quality, in separating a mixture of 2 clean speeches [4]. However, the challenging scenario of more than 2 speakers mixed with background noise received little attention [5]. A very recent preliminary test [6] revealed that existing deep learning models often underperform in such cases, because the unstructured background noise compromises their ability to identify separable structures in the speech streams. In addition, existing audio-only approaches cannot solve the *label permutation problem*, *i.e.*, associating the model outputs to the desired speaker. Audio-visual algorithms [6] leverage video recordings of the speakers' faces to simultaneously solve the SSE and permutation problems. However, the need for a camera at specific view angle and under amenable lighting condition limits their practical usability [7].

In this paper, *we propose to utilize ultrasound sensing as a complementary modality to separate the desired speaker voice from noises and interferences.* Our method, called UltraSE, is applicable to commodity mobile devices (*e.g.*, smartphones) equipped with a single microphone and loudspeaker. Figure 1 illustrates our basic idea. During the voice recording, UltraSE continuously emits an inaudible ultrasound wave, which is modulated by the speaker's articulatory gestures (lip movement in particular) close to the smartphone. The signals recorded by the microphone thus contain both the audible sounds and inaudible reflections. As illustrated in Figure 1, whereas the audible sounds ("Green") mix the targeted clean speech ("Black") and other interferences plus background noise ("Blue"), the inaudible reflections ("Orange") only capture the targeted user's articulatory gesture motion which is correlated with the clean speech. UltraSE employs a DNN framework to capture such correlation and denoise the audible sounds.

ACM MobiCom '21, October 25–29, 2021, New Orleans, LA, USA

ACM ISBN 978-1-4503-8342-4/21/10.
https://doi.org/10.1145/3447993.3448626

UltraSE faces 3 core design challenges. *i) How to characterize the articulatory gestures by ultrasound despite interference?* It is challenging to capture the fine-grained articulatory gestures since they are fast ($-80 \sim 80$ cm/s) and subtle (< 5 cm displacement). Moreover, mutual interference exists between the speech and ultrasound due to harmonics and hardware artifacts. To address the challenge, we fully exploit the advantages of ultrasound, *i.e.*, high sampling rate and perfect alignment with the clean speech in the time domain. We design the transmitted ultrasonic waveform to capture the *short-term high-resolution Doppler spectrogram*, and apply a one-time transmission volume calibration to reduce the cross-modality interference.

ii) How to design a DNN model to fuse the two modalities and represent their correlation? Since the physical feature characteristics of the two modalities are different, we design a two-stream DNN architecture to process each and a self-attention mechanism to fuse them. Further, no existing method has addressed the cross-modal noise reduction problem which is fundamental to UltraSE, *i.e.*, using one modality (ultrasound) to reconstruct another modality (speech) which is polluted by noise/interference. We thus propose a conditional GAN (cGAN) based training model, with a novel cross-modal similarity measurement network, to enable this capability.

(iii) How to improve both intelligibility and quality for the enhanced speech? It is known that the amplitude of time-frequency (T-F) spectrogram is critical for speech intelligibility, whereas the phase determines the speech quality [8]. We thus expand UltraSE into a two-stage multi-domain DNN architecture, which prioritizes the optimization of intelligibility in the T-F domain, and then reconstructs phase in the T domain to improve speech quality. We place the multi-modal fusion network inside the T-F domain, based on the empirical observation that the articulatory gestures are more related to the speech intelligibility.

To evaluate UltraSE, we develop an Android app to collect a new speech dataset called UltraSpeech, which contains 22.2 hours of clean speech and corresponding ultrasound sensing signals from 20 users. We then combine UltraSpeech with the DARPA TIMIT speech corpus [9] and AudioSet ambient noise dataset [10] to create a 300 hours noisy speech dataset. Our evaluation results show that UltraSE can separate the targeted speech in a sophisticated environment with multiple speakers and ambient noise , improving SNR by 10.65 to 17.25 dB. UltraSE achieves an SNR gain of 6.04 dB on average over state-of-the-art single-channel speech enhancement methods, across various interference/noise settings. Its performance gain is even comparable to multi-channel (audio-visual) solutions.

UltraSE represents the first audio-only method to bring the SSE performance close to multi-channel solutions, while overcoming the label permutation issue. Through the UltraSE design, we make the following technical contributions:

- We design a *multi-modal multi-domain DNN framework for single-channel speech enhancement* which fuses the ultrasound and speech features, and simultaneously improves speech intelligibility and quality.
- We design a *cGAN-based cross-modal training model* which effectively captures the correlation between ultrasound and speech for multi-modal denoising.
- We collect a new speech dataset—UltraSpeech, and verify UltraSE's performance in comparison with state-of-the-art solutions.

2 RELATED WORK

2.1 Audio-only Speech Enhancement

Despite decades of research, speech enhancement remains a challenging open problem that attracts extensive research today [6, 11–13]. Classical model-driven solutions [14–16] typically build on various assumptions, such as stationarity of signals, uncorrelated clean-speech and noise, independence of speech and noise in the time-frequency domain, *etc*. Thus, they often lack robustness in real-world environment [3]. More recent solutions adopt supervised learning instead [3], and can be categorized by their domain of feature processing.

T-F domain methods: Time-Frequency (T-F) domain methods aim to learn a spectrogram *mask*, *i.e.*, a weighting matrix that can be multiplied with the noisy speech spectrogram to recover the desired clean speech [12]. The key problem lies in *i)* what type of mask should be used, and *ii)* how to use DNN to predict such a mask. Early stage solutions only estimate the amplitudes of a spectrogram by using real-valued Ideal Binary Mask (IBM) [17], Ideal Ratio Mask (IRM) [18] or Spectral Magnitude Mask (SMM) [19]. They then directly apply the original noisy phase on each T-F bin to generate the enhanced speech. Although these amplitude masking methods benefit speech intelligibility, they suffer from poor speech perceptual quality due to the unavoidable phase error. Complex Ideal Ratio Mask (cIRM) [20] and Phase-Sensitive Mask (PSM) [21] are then proposed to incorporate phase information. Recently, PHASEN [12] and Ni *et al*. [22] found that the estimated cIRM tends to downgrade to IRM, since the T-F domain phase is close to white noise especially for low-amplitude T-F bins. Thus, they proposed two-stream [12] or two-stage [22] networks to take both the IRM and cIRM and derive a combined training loss. For the model design, most T-F domain methods deem the T-F spectrogram as an image, and design DNN/CNN-based models [20, 23] to minimize the MSE/MAE loss between the estimated mask and ground truth. PHASEN [12] and Ouyang *et al*. [24] observed that the fundamental frequencies and speech harmonics are separated afar, and the correlation cannot be fully captured by CNN. So they adopt dilated convolution and frequency-domain attention instead. Unlike the hand-crafted MSE/MAE loss function, Soni *et al*. [25] further used GAN to discriminate whether the enhanced results are clean or noisy.

T domain methods: Time (T) domain methods divert around the error-prone phase prediction problem by processing the waveform directly. For example, Rethage *et al*. [26] modified the WaveNet; TCNN [27] proposed an encoder-decoder architecture with an additional temporal convolutional net; SEGAN [28] utilized a GAN-based network to generate the 1D waveform of clean speech. Yet the performance of such methods is not among the top tier, since the speech auditory patterns, such as proximity in time/frequency, harmonics, and common amplitude/frequency modulation, are more prominent on a T-F spectrogram [3].

Multi-domain methods: In recent concurrent work TFTNet [13], a learnable decoder replaces the iSTFT in the T-F domain to realize a joint T-F and T domain model for speech enhancement. Unlike TFTNet, our key insight is that the speech intelligibility is much more important than speech quality for speech enhancement. We thus design a two-stage multi-domain DNN network to prioritize

the optimization of speech intelligibility in the T-F domain, and then reconstruct phase in the T domain to improve the speech quality.

Speech source separation: Although most of the aforementioned approaches demonstrated acceptable performance for non-speech noise, they still can not handle the *cocktail party* scenario involving multiple interfering speakers. To resolve such speech separation problems, Deep clustering [29] trained speech embedding for each source and then uses clustering algorithms to separate them. PIT [30] iteratively changed the permutation of sources in the training process to train a permutation invariant speech separation model. *These methods still need to know the number of speakers a priori*, and do not work well for the case with more than 3 speakers plus noise [31]. Further, the label permutation problem persists—They can separate multiple sources of speech, but cannot automatically identify which is from the targeted speaker, which may hinder certain machine-operated back-end tasks (*e.g.*, voice assistant on a smartphone). UltraSE overcomes all these deficiencies.

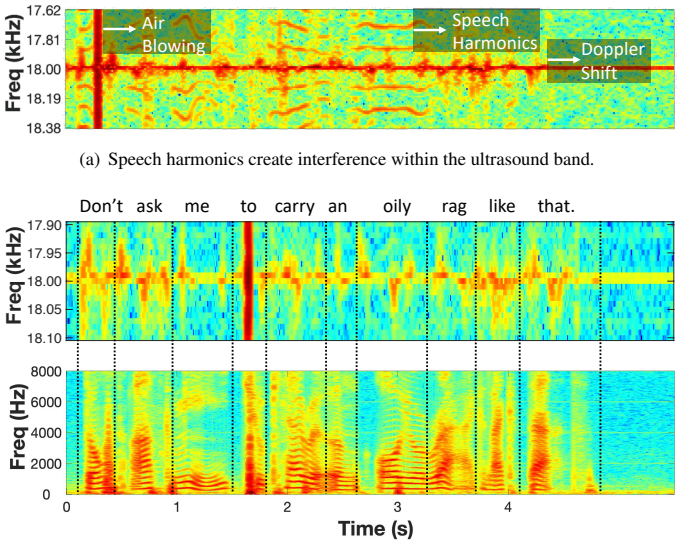
2.2 Multi-modal Speech Enhancement

To tackle the permutation issue, audio-visual (AV) methods use a video recording of the subject's face as a hint for the audio [32, 33]. Specifically, Ephart *et al*. [6] trained a speaker-independent speech separation model based on a large set of YouTube videos [6]. Afourasl *et al*. [7] found that even partially occluded videos of lip motion can assist speech separation. Nonetheless, AV approaches bear many drawbacks. Besides microphone, they need an additional camera pointing to the subject's face under good lighting conditions, which is inconvenient and even infeasible in many typical use cases. Moreover, camera is unusable in many privacy-sensitive locations.

The idea of using ultrasound as a complementary modality to enhance speech has been explored by previous works [34, 35]. However, these works [34, 35] all require special ultrasonic hardware. In comparison, UltraSE only needs the single audio channel on the smartphone and overcomes practical challenges such as mutual interference between modalities. Besides, they use traditional methods, *i.e*., non-negative matrix factorisation [35] and nonlinear regression [34], and only show the performance of speech enhancement on ambient noise rather than speech interference. UltraSE further pushes the limits of this idea by designing a multi-modal multi-domain DNN framework to achieve similar performance for speech separation and enhancement with the audio-visual methods.

2.3 Device-free Ultrasonic Sensing

Device-free ultrasonic sensing techniques can leverage the loudspeakers and microphones on commodity mobile devices to track the distance/direction changes of nearby objects [36]. State-of-the-art ultrasonic gesture tracking schemes [36–40] can achieve mm-level accuracy. Besides location and hand gesture tracking, recent studies also attempted to use ultrasonic sensing for lip reading [41]. However, due to insufficient spatial resolution, they only fit coarse sensing applications, *e.g*., liveness detection [42, 43]. SilentTalk [41] uses a model-based method to classify the Doppler shift features caused by 12 basic mouth motions and recognize specific short sentences. SilentKey [44], EchoPrint [45], LipPass [46], and VocalLock [47] use the ultrasounic sensing features introduced by mouth motion for biometric authentication. In contrast, UltraSE is the first to demonstrate that ultrasonic sensing can serve as a complementary modality



(a) Speech harmonics create interference within the ultrasound band.

(b) Doppler shift spectrogram of a single-tone 18 kHz transmitted signal and the corresponding T-F spectrogram of speech w/o interference.

Figure 2: T-F domain features of an example speech segment: "Don't ask me to carry an oily rage like that."

to solve the *cocktail party problem* and bring speech enhancement to the next level.

3 SENSING THE ARTICULATORY GESTURES

In this section, we first provide a primer on the relationship between speech and articulatory gestures. Then we introduce UltraSE's ultrasound sensing signal design, along with mechanisms to mitigate the mutual interference between speech and ultrasound.

Human speech generation involves multiple articulators, *e.g*., tongue, lips, jaw, vocal cords, and other speech organs [42]. Coordinated movement of such articulators, including lip protrusion and closure, tongue stretch and constriction, jaw angle change, *etc*., is used to define the phonological units, *i.e*. phoneme in phonology and linguistics [48]. Thus, assuming that we can fully capture and interpret the articulatory gestures, it would be possible to recover the speech signals. However, it is challenging to capture the fine-grained gesture motion of all articulators by using a single microphone [41]. First, the articulators are close to each other. Some are inside the mouth/throat. So it is hard to discriminate their motion. Second, the articulatory gestures are always fast and subtle. Each typically lasts 100 $\sim$ 700 ms and involves $<$ 5 cm moving distance for lip and jaw [49]. Thus, state-of-the-art sensing methods can only recognize a limited number of words or phrases by using COTS microphones [41], and the accuracy in the wild is typically quite low [50]. In UltraSE, we do not expect that the captured articulatory gesture features can directly synthesize the speech signals. We propose to take these features as coarse complementary information to facilitate the SSE.

3.1 Transmitted Ultrasound Signals Design

Modality advantages: Compared to other approaches such as camera, ultrasound possesses two advantages in sensing articulatory

gestures. First, the ultrasound sensing signals are captured by using the same sensor (*i.e.* microphone) as the speech signals. This introduces an automatic *"feature alignment"* in the time domain, which means the captured ultrasound sensing features are well synchronized and matched with the clean speech signals. Second, the *sampling rate* of the ultrasound sensing (typically 48 kHz or 96 kHz) is much higher than vision-based methods (typically $24 \sim 120$ fps), which enables finer time resolution when capturing the articulatory gestures.

Design goals: Compared to previous works on ultrasound based gesture sensing especially hand gestures [36, 37, 39, 51], UltraSE needs to satisfy the following additional design goals to fully exploit the modality advantages: ***(i)*** The extracted features require high sampling rate to achieve high T-F resolution. The velocity of users' articulatory gestures ranges from $-80 \sim 80$ cm/s ($-160 \sim 160$ cm/s for propagation path change) [49], which will introduce $-100 \sim 100$ Hz Doppler shift when the transmitted signal's frequency is 20 kHz. Meanwhile, each articulatory gesture corresponds to a single phoneme lasting $100 \sim 700$ ms [42], which is approximately 5 times shorter than hand gestures [52]. *Therefore, to characterize the articulatory gestures, the ideal way is to characterize the short-term high-resolution Doppler shift*. ***(ii)*** The extracted features need to be robust to different kinds of noises introduced by multipath and frequency-selective fading. UltraSE thus needs to remove the reflections from static objects, mitigate the multipath from moving objects (*e.g.*, body parts), and extract the signal features from articulatory gestures alone.

Ultrasonic sensing signal design: To satisfy these requirements, we choose multiple single-tone continuous waves (CWs) with linearly spaced frequencies as our transmitted signals. Although modulated CW signals, such as FMCW [53], OFDM [51] and Pseudo-Noise (PN) sequences [37, 39], can measure the impulse response to resolve multipath, they all suffer from the aforementioned low sampling rate problem. The fundamental reason is that the modulation processes signal in segments (*i.e.*, chirp period or symbol period). Thus, each feature point of the modulated CW signal characterizes the motion within a whole segment, which is typically longer than 10 ms (960 samples) at a sampling rate of 96 kHz. Thus, only $10 \sim 70$ feature points can be output for each articulatory gesture with typical duration of $100 \sim 700$ ms [42], which can hardly represent the fine-grained instantaneous velocity of gesture motion. In comparison, each sampling point of the single-tone CW can generate one feature point (Doppler shift estimation) to represent the micro motion with duration of 0.01 ms ($\frac{1}{96000}$) at a sampling rate of 96 kHz. To further resolve the multipath effect and frequency selective fading, we combine multiple single-tone CWs with equal frequency spacing, resulting in a transmitted waveform $T(t) = \sum_{i=1}^{N} A_i \cos 2\pi f_i t$, where N, A_i and f_i denote the number of tones, the amplitude and frequency of the i^{th} tone, respectively.

To alleviate the spectral leakage across different tones when generating the spectrogram in later stage, we ensure that the STFT window size (1024 points) is a full cycle of all the transmitted tones at the maximum sampling rate (48 or 96 kHz allowable by COTS microphones). We empirically set the first frequency $f_0 = 17.25$ kHz, the frequency interval $\Delta f = 750$ Hz, and the number of tones $N = 8$. We decrease the amplitude A_i of the sub-20kHz frequencies to make sure that the transmitted signals will not disturb users.

3.2 Mitigating Sensing Interference

Despite the orthogonality in frequency, mutual interference exists between speech and ultrasound in the following two cases, which causes ambiguity of Doppler features.

First, the speech harmonics may interfere the Doppler features due to non-linearity of microphone hardware. The speech and ultrasound signals generated in UltraSE are combined in the air, resulting in $S_{in}(t) = v(t) + \sum_{i=1}^{N} A_i \cos 2\pi f_i t$, where $v(t)$ represents the speech signals, and $\sum_{i=1}^{N} A_i \cos 2\pi f_i t$ is the high-frequency ultrasound sensing signals. Due to the microphone non-linearity [54–56], the captured signals can be modeled as $S_{out} \simeq A^1 S_{in} + A^2 S_{in}^2$ [54], which contains speech harmonics on the inaudible ultrasonic band, *i.e.*, $S_{noise} = \sum_{i=1}^{N} A_i^2 v(t) \cos 2\pi f_i t$. As shown in Figure 2(a), these speech harmonics often leak into the ultrasonic band, and will corrupt the articulatory gestures' Doppler features. Fortunately, when we decrease the amplitude of the ultrasound A_i, the second order term (harmonics' amplitude A^2) decreases faster than the first order term (Doppler shift amplitude A^1). Our empirical experiments reveal that, when the total amplitude of transmitted ultrasound is set to < 80 dBz (flat weighting) sound pressure level (measured at 5 cm away from the speaker), the interference effect becomes negligible. We thus always use this setting as the *default ultrasound amplitude* in UltraSE. It is worth noting that previous ultrasound based hand gesture sensing schemes [36, 37, 39] did not address this interference issue because they are typically tested without strong close-by speech interference.

Second, when a user speaks close to the microphone, some phonemes, *e.g.*, /p/ and /t/, may blow air flow into the microphone which generates high-volume noise. As an example, Figure 2(a) shows the T-F spectrogram introduced by the phoneme /t/. Amid the high-volume air flow, the microphone has to prevent saturation by calling on its auto gain control (AGC), which scales down all incoming signals and consequently renders the Doppler features negligible. In UltraSE, instead of removing the corrupted samples, we harness them as part of the ultrasonic sensing features, which helps characterize the sampling period corresponding to specific phonemes (*e.g.*, the /t/).

4 AN OVERVIEW OF ULTRASE DNN MODEL

For ease of exposition, we will first introduce the basic DNN architecture of UltraSE, and then discuss the challenges and design principles of each design component in the following sections. Our first step is to create the DNN input features from the raw signals (Section 5). Then, we design a *two-stage, multi-modal, multi-domain DNN model*, which comprises three key modules, as briefed below.

T-F domain multi-modal amplitude network (Section 6). This network module generates the amplitude Ideal Ratio Mask (aIRM), *i.e.*, the ratio between the magnitudes of the clean and noisy spectrograms, by using both speech and ultrasound as the input. It consists of two subnetworks.

Subnet (i) Two-stream feature embedding: Our model starts by using the noisy speech's T-F spectrogram and the concurrent ultrasound Doppler spectrogram as input (Section 5). We then design a

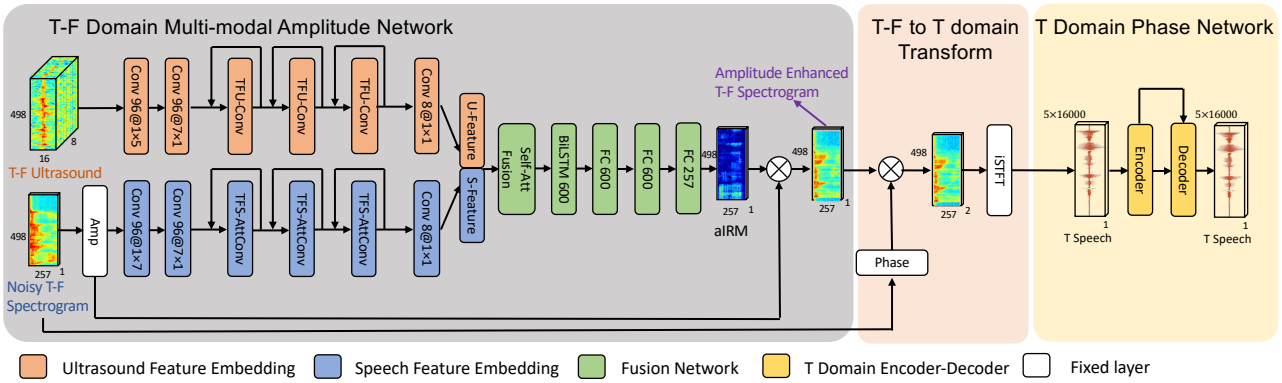


Figure 3: Overview of UltraSE's multi-modal multi-domain DNN design. Convolution layer notation: Channels@Kernel size

two-stream feature embedding architecture, to transform the different modalities into the same feature space, while maintaining their time-domain alignment.

Subnet (ii) Speech and ultrasound fusion network: Then, we concatenate the features of each stream in the frequency dimension. A self-attention mechanism is further applied to fuse the concatenated feature maps to let the multi-modal information "crosstalk" with each other. The fused features are subsequently fed into a BiLSTM layer followed by three FC layers. The resulting output is an *amplitude mask* which is multiplied with the original noisy amplitude spectrogram to generate the amplitude-enhanced T-F spectrogram.

cGAN-based cross-modal training (Section 7). As shown in Figure 4, we design a cGAN-based training method to further denoise the amplitude-enhanced T-F spectrogram. In our cGAN model, the generator is the above T-F domain multi-modal amplitude network; the discriminator is designed to discriminate whether the enhanced spectrogram corresponds to the ultrasound sensing features.

T domain phase network (Section 8). We use the iSTFT (a fixed 1D convolution layer) [57] to transform the amplitude-enhanced T-F spectrogram into T domain waveform. To fine-tune the phase of the enhanced signals, we design an encoder-decoder architecture to reconstruct the phase to be close to the clean speech in the T domain.

5 DNN INPUT FEATURE DESIGN

In this section, we discuss the preprocessing steps to generate the DNN input features for the two signal modalities. Figure 5 illustrates the workflow.

Speech feature extraction: Typical speech sound ranges from approximately 300 Hz to 3.4 kHz [58], and the signals above 8 kHz barely affect the speech intelligibility and human perception [59]. Thus, we first use a low-pass elliptic filter to extract the signals below 8 kHz. Then we resample the signals to 16 kHz by using a Fourier method. The final enhanced speech is also sampled at 16 kHz which suffices to characterize the speech signals. Higher sampling rate may unnecessarily increase the optimization space and model complexity.

The speech feature input for the DNN model is the T-F domain speech spectrogam, generated by applying STFT on the time domain waveform. The STFT uses a Hann window of length 32 ms, hop length of 10 ms, and FFT size of 512 points under 16 kHz sampling rate, resulting in $100 \times 257 \times 1$ complex-valued scalars per second.

Ultrasound sensing features: We first use a high-pass elliptic filter to isolate the signals above 16 kHz. Then, we create the ultrasound sensing features within the T-F domain, by extracting the Doppler spectrogram induced by articulatory gestures and aligning it with the speech spectrogram. The key consideration for this step is to balance the trade-off between time resolution and frequency resolution of the STFT under the limited sampling rate (96 kHz maximum). First, to guarantee time alignment between the speech and ultrasound features, their hop length in the time domain should be the same. The STFT uses a hop length of 10 ms to guarantee 100 frames per second, resulting in $10 \sim 70$ frames per articulatory gesture which is enough to characterize the process of an articulatory gesture (Section 3). Second, the frequency resolution, determined by the window length, should be as fine-grained as possible to capture the micro Doppler effects introduced by the articulatory gestures, under the premise that the time resolution is sufficient. A window length 85 ms is the longest length for STFT to make it shorter than the shortest duration of an articulatory gesture (100 ms) [42]. Overall, under the 96 kHz sampling rate, the STFT is computed using a window length 85 ms, hop length of 10 ms, and FFT size of 8192 points, resulting in 11.7 Hz ($\frac{96000}{8192}$) frequency resolution.

In addition, to mitigate the reflections from relatively static objects, we remove the 3 central frequency bins and leave $8 \times 2 = 16$ frequency bins corresponding to Doppler shift $[-11.7 \times 8, -11.7)$ and $(11.7, 11.7 \times 8]$ Hz. Finally, we run a min-max normalization on the ultrasound Doppler spectrogram. The resulting T-F domain ultrasound features is $100 \times 16 \times 8$ scalars per second, where 8 is the number of ultrasonic tones. The reason why we fuse the ultrasound sensing features in T-F domain instead of T-domain will be evident in the latter multi-domain design (Section 8).

The origin of the ultrasound feature and its correlation with the speech feature: Figure 2(b) uses one example speech segment to visualize the alignment between the ultrasound Doppler spectrogram and the clean speech spectrogram. The ultrasound sensing features mainly consist of the $-100 \sim 100$ Hz Doppler shift introduced by relatively large motion from the lip, tongue and jaw. It can not capture the high-frequency micro-vibration motions introduced by the vocal folds [60], since the vocal vibration displacements (about 20 μm [61]) are much shorter than the ultrasound wavelength (about 2 cm).

Some obvious characteristics in this example corroborate the correlation between the ultrasound sensing features and corresponding

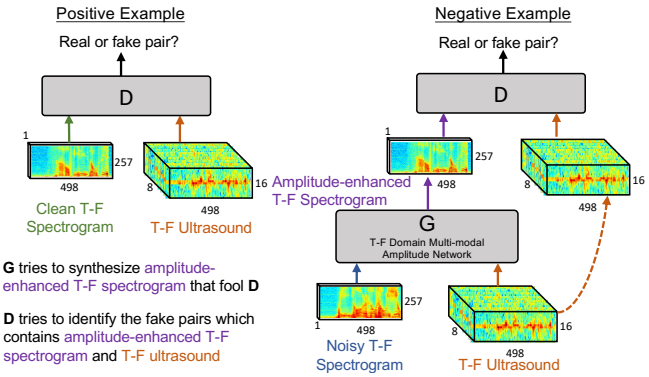


Figure 4: Overview of UltraSE's cGAN-based cross-modal training.

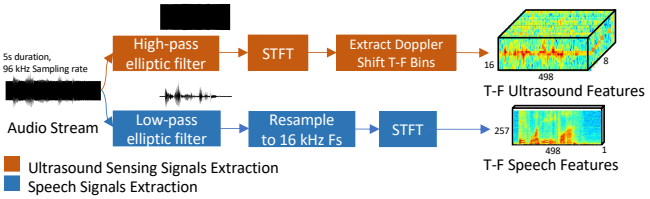


Figure 5: DNN input feature design

clean speech features. For example, each word in the speech signals is well aligned with a burst of Doppler shifts from the articulatory gestures. Meanwhile, negative Doppler shift is introduced by mouth open gestures slightly before the onset of each word. Our DNN model is designed to learn such cross-modality correlation for the purpose of SSE.

6 MULTI-MODAL FUSION DESIGN

The multi-modal fusion network aims to first appropriately learn the F domain features of the two modalities respectively, and then fuse them together to exploit the T-F domain correlation. The F domain of the ultrasound signal features represents the *motion velocity* (Doppler shift) of the articulatory gestures, while that of the speech sound represents the *frequency* characteristics such as harmonics and consonants. Meanwhile, the size of the two modalities' feature maps are different (Section 5). So one cannot straightforwardly concatenate these two feature maps into a scalar. We thus design a two-stream embedding architecture to transform them into the same feature space.

6.1 Two-stream Feature Embedding

Speech feature embedding: The input of the speech feature embedding subnetwork is the T-F domain amplitude spectrogram, denoted as $S^a_{noise} \in \mathbb{R}^{1\times T\times F^a}$. $F^a = 257$ is determined by the STFT window size. The "blue" part in Figure 3 shows the architecture of this subnetwork, which comprises traditional 2D convolution layers and 3 "TFS-Conv" blocks. The "TFS-AttConv" block, borrowed from PHASEN [12], employs both the ResNet [62] and self-attention mechanism [63] to learn the global correlation of sound patterns across T-F bins. In contrast, the small kernels of CNN cannot capture such long-range correlations. Figure 6 shows the structure of a single "TFS-AttConv" block. It contains 2 "CF-Att" blocks at the beginning and the end to learn the global correlation. In each "CF-Att",

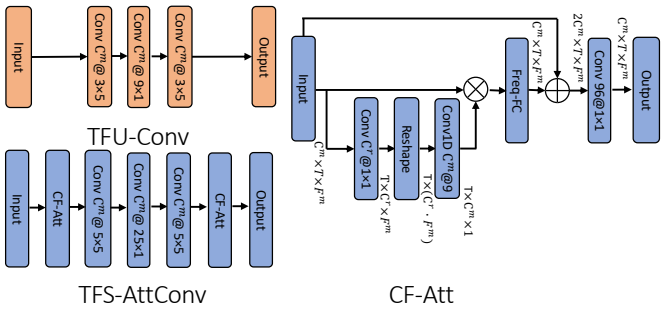


Figure 6: Two-stream feature embedding. Convolution layer notation: Channels@Kernel size

a self-attention mechanism is used to fuse the channel-wise information following a SENet-based design [63]. Then, the "Freq-FC" layer applies a learnable frequency transformation matrix to enable frequency-domain self-attention at each point in the T domain. We omit other details of this block which has been covered in PHASEN [12].

Ultrasound feature embedding: The input of the ultrasound feature embedding is $U^s \in \mathbb{R}^{T\times F^s\times C^s}$, where $C^s = 8$ is the number of ultrasound tones, and $F^s = 16$ is the maximum number Doppler shift frequency bins introduced by the articulartory gestures (Section 5). Since the motion speed always changes continuously, the F domain ultrasound features are mainly local Doppler shift features. Small kernels suffice to capture such feature correlation because the size of the F domain is only 16. Therefore, instead of the "TFS-AttConv", we design a "TFU-Conv" block which removes the attention layers and reduces the kernel size of the F domain in all the 2D convolution layers. To maintain the time alignment of the two modalities after feature embedding, we keep the T domain kernel size the same as in the "TFS-AttConv" block. For convenience of concatenating the two modalities' features, we choose the same output channel number for all the 2D convolution layers.

Finally, after 3 "TFU-Conv" and "TFS-AttConv" blocks respectively, the channel number of the two streams reduces to $C^s_r = 8$ and $C^a_r = 8$ by applying a 1×1 2D convolution.

6.2 Speech and Ultrasound Fusion Network

After the feature embedding, we concatenate the feature maps of the two streams: $S^f_{in} = concat(M^a_o, U^s_o)$, where $S^f_{in} \in \mathbb{R}^{C^r\times T\times F^{as}}$ and $F^{as} = F^a + F^s$. This concatenated feature map is then fed into the "Self-Att Fusion" to learn the relationship between the two modalities. The "Self-Att Fusion" block is similar to the "CF-Att" block, but the size of the feature maps differs. *First*, since the meaning of channel in ultrasound sensing and speech is different, we first use a channel self-attention to learn the correlation across different channels. *Second*, to enable these two modalities' features to "crosstalk" with each other in the F domain, the self-attention for the F domain is realized by using a learnable transformation matrix on the fused features. *Third*, the feature after self-attention fusion is concatenated with the original feature and fused by a 1×1 2D convolution.

Finally, the whole feature map is fed into a BiLSTM and 3 fully connected (FC) layers to predict the aIRM $\in \mathbb{R}^{T\times F_m\times 1}$ of the noisy speech. The predicted aIRM is then multiplied with the original noisy

	Conv1	Conv2	Conv3	Conv4	Conv5	Conv6	Conv7	Conv8	Conv9		BLSTM10	FC11	FC12	FC13
Num Filters	48	48	48	48	48	48	48	48	8	Hidden Size	300			
Filter Size	1 × 7	7 × 1	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	1 × 1	Output Size		600	600	5

Table 1: Layers comprising ultrasound subnetwork

	Conv1	Conv2	Conv3	Conv4	Conv5	Conv6	Conv7	Conv8	Conv9	Conv10	Conv11	Conv 12	Conv 13
Num Filters	48	48	48	48	48	48	48	48	48	48	48	48	8
Filter Size	1 × 7	7 × 1	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	1 × 1
Dilation	1 × 1	1 × 1	1 × 1	1 × 2	1 × 4	1 × 8	1 × 16	1 × 1	2 × 2	4 × 4	8 × 8	16 × 16	1 × 1

Table 2: Layers comprising speech subnetwork (BLSTM and FC layers parameters are the same as the ultrasound subnetwork.)

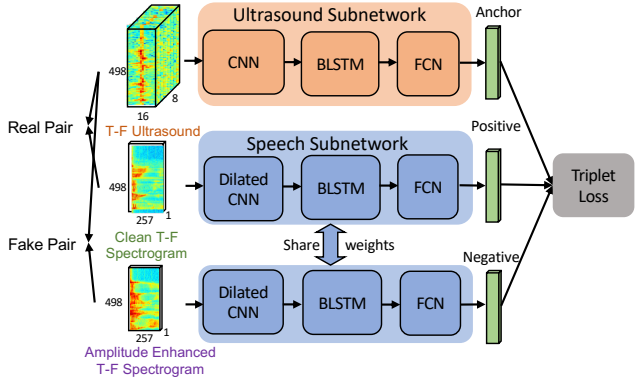


Figure 7: Architecture of the T-F domain cross-modal similarity measurement network (*i.e.*, the Discriminator).

speech's amplitude spectrogram to generate the *amplitude-enhanced T-F spectrogram*.

Note that all the convolutional layers in the multi-modal fusion network use zero padding, dilation= 1 and stride= 1 to make sure the output feature map size is the same as the input speech/ultrasound spectrogram. Also, each 2D convolutional layer is followed by batch normalization (BN) and ReLu activation.

7 CGAN-BASED CROSS-MODAL TRAINING

The fundamental problem for UltraSE is multi-modal noise reduction, *i.e.*, using one modality (ultrasound) to recover another modality (speech) which is polluted by noise/interference. The former modality has low sensing resolution but is interference-free and correlated with the latter. Although we intentionally maintain the time alignment between the two (Section 6), it is hard to force the multi-modal fusion network to "understand" such multi-modal correlation, because a traditional loss function (*e.g.*, MSE) can only train the network to clean up the T-F spectrum end-to-end. We thus propose a cGAN-based training method, which implicitly incorporates the maximization of cross-modal correlation itself as a training goal.

7.1 Cross-modal Similarity Measurement

A key element in any GAN design is to define the similarity metric used by the discriminator. Unlike traditional GAN applications (*e.g.*, image generation) which compare between the same type of features, our cross-modal cGAN needs to discriminate whether the enhanced T-F speech spectrogram matches the ultrasound Doppler spectrogram (*i.e.*, whether they are a "real" or "fake" pair). We propose a *cross-modal Siemese neural network* to meet this challenge.

A Siamese neural network uses shared weights and model architecture while working in tandem on two different input vectors to compute comparable output vectors. It is traditionally used to measure the similarity between two inputs from the same modality, *e.g.*, two images [64]. To enable a *cross-modal* Siamese neural network, we create two separate subnetworks (Figure 7), aiming to characterize the correspondence between the T-F domain features of the speech and ultrasound, respectively. The basic architecture for these 2 inputs is a CNN-LSTM model. Since human speech contains harmonics and spatial relationship in the F domain, the speech CNN subnetwork uses *dilated convolutions* for *frequency domain context aggregation*. The Doppler shifts from ultrasound sensing mostly encompasses local features. Thus, the ultrasound CNN subnetwork only contains traditional convolution layers. Following the convolution, a Bi-LSTM layer is used to learn the long-term *time-domain information* for both modalities. Finally, three fully connected (FC) layers are introduced to learn two comparable output vectors respectively. We emphasize that the architecture and parameters are not shared in this cross-modal design, which differs from the traditional Sieamese networks.

As shown in Figure 7, we use the Triplet loss [65] to train the cross-modal Siamese network. The triplet loss function accepts 3 inputs, *i.e.*, an anchor input U^s is compared to a positive input S^a_{gr} and a negative input S^a_{out}. It aims to minimize the distance between "real" pair U^s and S^a_{gr}, and maximize the distance between "fake" pair U^s and S^a_{out}. In our model, the anchor input U^s is the ultrasound sensing features, the positive input S^a_{gr} is the corresponding clean speech amplitude spectrogram, and the negative input S^a_{out} is the noisy speech amplitude spectrogram. Thus, our network model minimizes the following Triplet loss:

$$\mathcal{L}_{Triplet}(D) = \mathbb{E}_{U^s,S^a_{gr},S^a_{out}\sim p_{data}(U^s,S^a_{gr},S^a_{out})} \\ [(\|f_u(U^s) - f_s(S^a_{gr})\|^2 - \|f_u(U^s) - f_s(S^a_{out})\|^2 + \alpha, 0)] \quad (1)$$

where f_u is the ultrasound subnetwork, f_s is the speech subnetwork, and α is a margin distance between "real" and "fake" pairs.

We use a speech and ultrasound dataset collected on COTS smartphones (Section 9.1) to train the cross-modal Siamese network, and verify its effectiveness in a benchmark experiment. The training and testing sets contain 3 h and 0.5 h speech corpus for 15 and 5 users, respectively. The T-F domain speech feature input is a $1 \times 498 \times 257$ scalar (5 s segment), and the T-F domain ultrasound feature input is a $8 \times 498 \times 16$ scalar.

Figure 8 shows the probability density function (PDF) of outputs, where a smaller value indicates higher similarity. It is obvious that the output PDFs for the real pairs and fake pairs are perfectly separated, which means that *our similarity measurement network can effectively discriminate whether a pair of speech and ultrasound inputs are generated by the same articulatory gestures*.

	Enc1	Enc2	Enc3	Enc4	Enc5	Enc6	Enc7	Dec7	Dec6	Dec5	Dec4	Dec3	Dec2	Dec1
Num Filters	16	32	32	64	64	128	128	128	64	64	32	32	16	1

Table 3: Layers comprising T domain phase network. Kernel size = 32, Stride = 2, Padding = 15.

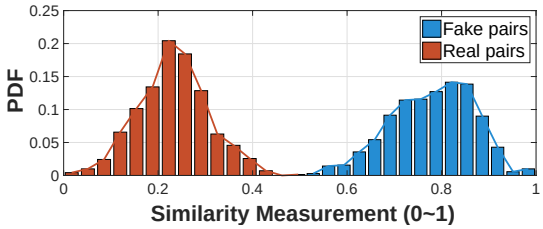


Figure 8: PDF of outputs from the cross-modal similarity measurement network.

7.2 cGAN-based Model Training

Now we discuss how to leverage such similarity measurement as a discriminator in a cGAN to further fuse the multi-modal information. Our cGAN model aims to not only minimize the MSE of the speech amplitude spectrogram (relative to the ground-truth), but also guarantee high similarity between the "fake" pair (*i.e.*, the enhanced speech and ultrasound sensing features) and the "real" pair (*i.e.*, the clean speech and ultrasound sensing features).

cGAN has been widely used to add a conditional goal to guide a generator to automatically learn a loss function which well approximates the goal [66]. Figure 4 shows the structure of the UltraSE cGAN model. The generator "$G(S^a_{noise}, U^s)$" is the aforementioned multi-modal network (Section 6), which takes the noisy speech amplitude spectorgram S^a_{noise} and ultrasound sensing spectrogram U^s as the input. $G(\cdot)$ is trained to output amplitude-enhanced T-F spectrogram of the speech S^a_{out}, which not only minimizes the traditional amplitude MSE loss [12], but also tries to "fool" an adversarially trained discriminator "$D(S^a_{out}, S^a_{gr}, U^s)$", which strives to discriminate the fake pair (S^a_{out}, U^s) from the "real" pair (S^a_{gr}, U^s) under the aforementioned triplet loss function. More specifically, The "D" loss is $\mathcal{L}_{Triplet}(D)$ (see Eq. (1)), and the "G" loss is

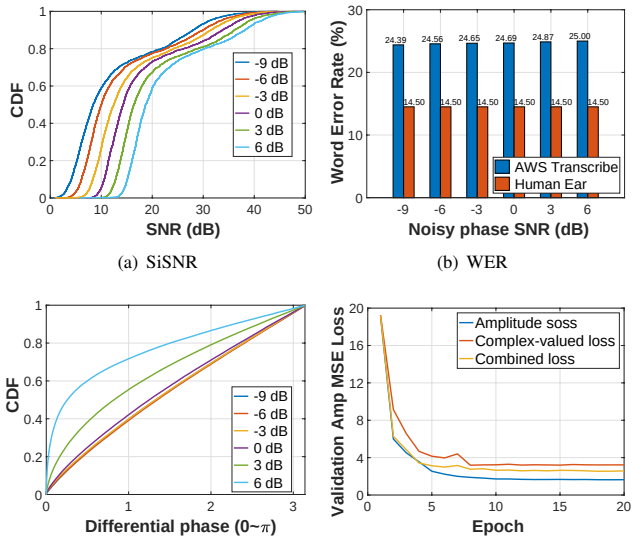
$$\mathcal{L}(G) = \mathbb{E}_{U^s, S^a_{gr}, S^a_{noise} \sim p_{data}(U^s, S^a_{gr}, S^a_{noise}), z \sim p_z}$$
$$[\mathcal{L}_{Triplet}(D(G(U^s, S^a_{noise}), S^a_{gr}), U^s)] + \lambda \|G(U^s, S^a_{noise}) - S^a_{gr}\|^2$$

where $\lambda\|G(U^s, S^a_{noise}) - S^a_{gr}\|^2$ is the traditional MSE amplitude loss. The reason why we use the amplitude MSE loss here rather than complex-valued loss or combined loss [12] will be clarified in Section 8.

Our cGAN design represents a general model for cross-modal noise reduction, which may be reused in other sensor fusion problems involving heterogenous sensing modalities.

8 MULTI-DOMAIN SPEECH ENHANCEMENT

In this section, we first investigate the pros and cons of T-F domain vs. T domain speech enhancement by using statistical analysis and experimental validation. Our key insight is that improving *intelligibility* is more critical than enhancing *quality*, since the top priority for speech enhancement lies in helping users/machines to understand the speech in noisy environment. This motivates us to expand the aforementioned T-F domain network into a two-stage multi-domain model, which first pushes the limits of intelligibility and then refines the speech quality.



(a) SiSNR (b) WER

(c) CDF of phase difference between the clean and noisy T-F domain spectorgram.

(d) Validation of amplitude loss by using different training loss functions.

Figure 9: Benchmark of the T-F domain methods

8.1 Understanding the Pros and Cons of T-F Domains Speech Enhancement

Speech sounds and interferences usually exhibit rich auditory patterns in the T-F spectrogram. In this section, we intend to understand the impact of phase in the T-F spectrogram to enlighten our multi-domain model design.

How does the T-F spectrogram phase affect the speech intelligibility and quality? We first conduct an experiment by using the UltraSpeech dataset (detailed in Section 9), where we keep the clean speech's amplitude in the T-F spectrogram while replacing its phase with the noisy speech phase, just as in aIRM (Sec. 2). We use two metrics to evaluate the impact. *(i) Scale-invariant Signal-To-Noise Ratio (SiSNR)* characterizes the speech *quality* [67]:

$$\mathcal{L}_{SiSNR} = 10\log_{10}\left(\frac{||\frac{\langle \hat{\mathbf{s}}, \mathbf{s}\rangle \mathbf{s}}{||\mathbf{s}||^2}||^2}{||\frac{\langle \hat{\mathbf{s}}, \mathbf{s}\rangle \mathbf{s}}{||\mathbf{s}||^2} - \hat{s}||^2}\right) \quad (2)$$

where $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the T domain clean speech and enhanced speech signals, respectively. *(ii) Word Error Rate (WER)*, representing speech *intelligibility*, is the probability that a word cannot be correctly recognized by an automatic speech recognition (ASR) algorithm [68] and human perception.

As shown in Figure 14(b), when applying the noisy T-F spectrogram phase directly, the SiSNR degrades slightly. On the other hand, phase does not affect the WER in a noticeable way. The noisy phase with a very low SNR of −9 dB only decreases the WER by 0.7% when using AWS Transcribe [68]. Meanwhile, human subjects can clearly understand the speech and only feel a little jittering effect. In

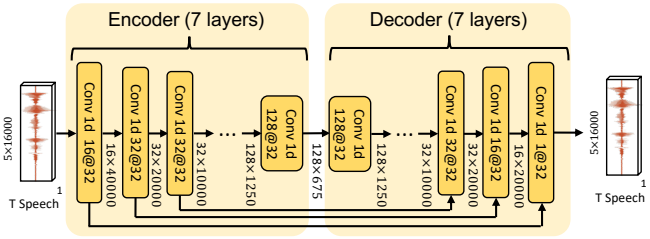


Figure 10: T domain phase network. Convolution layer notation: Channels@Kernel size

summary, *the phase in the T-F spectrogram barely affects the speech intelligibility and only slightly degrades the speech quality*.

What is the appropriate training loss function for recovering the speech intelligibility? Figure 9(c) plots the CDF of phase difference between the clean and noisy speech spectrogram across the T-F bins. We see that the phase difference is almost uniformly distributed for low-SNR speech. This means the phase values in all the T-F bins are distorted in the spectrogram which makes the phase recovery challenging. Since phase is not critical to intelligibility, we proceed to study the performance of different DNN loss functions in recovering the T-F spectrogram *amplitude*.

We examine 3 different loss functions. The first is the amplitude MSE loss which only considers the T-F spectrogram amplitude: $\mathcal{L}_a = \lambda\|G(U^s, S^a_{noise}) - S^a_{gr}\|^2$. The second is the complex-valued MSE loss which accounts for both the T-F spectrogram amplitude and phase: $\mathcal{L}_p = \|S^c_{out} - S^c_{gt}\|^2$ The third is a combined loss used in PHASEN: $\mathcal{L}_{combined} = 0.5\times\mathcal{L}_a + 0.5\times\mathcal{L}_p$, where S^a_{out}, S^a_{gt} and S^c_{out}, S^c_{gt} are the power-law compressed ($A^{0.3}$) amplitude spectrogram and complex-valued spectrogram. We apply these 3 training loss functions to the architecture in Section 6 and 7. Figure 9(d) shows the validation amplitude MSE loss. Obviously, upon convergence, training with amplitude MSE loss leads to lower validation error in amplitude MSE, and hence better speech intelligibility, than the two alternative loss functions.

8.2 Two-stage Multi-domain Network Design

Based on the above studies, we derive 3 design principles for our multi-domain architecture: *(i)* The T-F spectrogram amplitude contributes to the speech intelligibility whereas the phase is related to the speech quality. *(ii)* The T-F spectrogram phase is hard to predict by using DNN models. *(iii)* Training DNN models with aIRM MSE loss in the T-F domain optimizes speech intelligibility. We now elaborate on the detailed design, which follows the flow in Figure 3.

Stage 1: T-F domain multi-modal amplitude speech enhancement. The DNN architecture and training model of this stage has been covered in Sec. 6 and Sec. 7. The amplitude-enhanced T-F spectrogram output is multiplied with the original noisy phase to generate a complex-valued T-F spectrogram. Then, the iSTFT [57] is used to transform the T-F spectrogram to the T domain waveform and output the *amplitude-enhanced T-domain waveform*.

Stage 2: T domain speech phase enhancement. The goal of this stage is to fine tune the T-domain waveform to further improve the speech *quality*, using the SiSNR (Eq. (2)) as the training loss function. Inspired by SEGAN [28], our T domain network is an encoder-decoder network as shown in Figure 10. The encoder contains 7 1D convolution layers to transform 5 s of time domain waveform to a 128×675 scalar. The decoding stage reverses the encoding operation by means of fractional-strided transposed convolutions. We connect each encoding layer to its homologous decoding layer to fully capture the low-level details of the original features. The network parameters are listed in Table 3. All the 1D convolutional layers are followed by parametric rectified linear units (PReLUs) [69]. We also tried a cGAN training model similar to Section 7 in this stage, but observed negligible performance gain. Thus, we only enforce the cGAN training in the T-F domain.

Notably, the first and second stage output can be used to satisfy different applications, *e.g.*, for ASR and human listener, since they are trained for speech intelligibility and quality, respectively.

9 ULTRASE IMPLEMENTATION

9.1 UltraSpeech Dataset

Traditional speech datasets only contain raw speech without ultrasound sensing signals [9, 70]. To evaluate UltraSE, we thus create a new dataset called UltraSpeech which comprises both.

Data collecting: We recruited 20 fluent English speakers (4 female, 16 male, average age 25) to collect the UltraSpeech dataset. Each participant was asked to say at least 300 sentences in the TIMIT speech corpus [9] by using 2 typical phone holding styles ("Phone Call" mode and "Towards Mic" mode, shown in Figure 12(b)) in quiet environment. Meanwhile, we use a custom-built Android app called UltraRecord, to emit the ultrasonic signals and capture the audio segments at 96 kHz sampling rate, through the bottom speaker and microphone on a smartphone. Note that we do not constrain the user to hold the smartphone at a specific distance from the mouth. In total, we collected 8k 5-second clean speech segments for each holding style.

We follow existing SSE work [6, 12] to generate the noisy speech dataset through synthetic mixture. The interfering speech comes from the TIMIT data set [9], which contains 6300 different English sentences, generated by 630 speakers lasting 3.5 hours in total. The ambient noise dataset comes from AudioSet [10] which contains more than 1.7 million 10-second segments of 526 types of noise from real-life, including a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds.

Training/testing dataset generation: Each segment of training/testing data is synthesized by a linear combination of 3 pieces: $\langle S_j, U_j, S_{i_{noise}}\rangle$, where S_j and U_j are the clean speech segment and corresponding ultrasound features from UltraSpeech; $S_{i_{noise}}$ is the i_{th} noisy sound segment.

Besides, we generate a training set where the interfering speech and clean speech come from the same speaker. This is widely recognized as the *most challenging case* of SSE [71], since the interference bears the same auditory patterns that are indistinguishable from the desired speech. We add this into the training dataset to force the model to exploit the ultrasound features in addition to the audible features.

Our training dataset contains 15 participants' clean speech collected by the Samsung Galaxy S8 smartphone. Each participant's clean speech is mixed with 20 different noise settings. For each noise

setting, the number of interfering speakers n is uniformly distributed in $[0, 4]$, and the SNR is uniformly distributed in $[-9, 6]$ dB (-1.5 dB average). In total, the training data contains 120k 5-second segments of noisy speech (300 hours).

9.2 UltraSE DNN Implementation

We implement the UltraSE DNN model in Pytorch. The dimension of feature maps and the parameters of each layer are shown in Figure 3, 6, 10 and Table 1, 2, 3. ReLU activations follow all layers except for the last layer, where a sigmoid is applied. For training, we use Adam optimizer with a $1e-04$ initial learning rate, dropping by 25% every 5 epochs for a total of 20 epochs. UltraSE has 15.5 M and 3.1 M parameters for the first and second stage DNN.

10 EXPERIMENTAL EVALUATION

We evaluate UltraSE using 4 metrics commonly adopted in SSE research.

• SDR [72]: Signal-to-distortion ratio, which considers not only noise/interference, but also acoustic artifacts (*e.g.*, burbling sound) as distortion to the ground-truth speech;

• SiSNR [73]: Scale-invariant signal-to-noise ratio (Sec. 8.1) which, unlike the classical SNR, ensures rescaling the estimated signal will not unfairly improve the metric;

• STOI [74]: Short-time objective intelligibility measure (from 0 to 1);

• PESQ [75]: Perceptual evaluation of speech quality, which models the mean opinion score ranging from 1 (bad) to 5 (excellent);

10.1 Micro Benchmark Comparison

In this section, our default testing dataset includes another 5 participants' clean speech in the "Towards mic" mode, collected using Samsung S8. Our testing environment includes 6 different interference plus noise settings: $1s+a$, $2s+a$, $3s+a$, $>3s+a$, $2s$ ("s" and "a" denotes interfering speaker and ambient noise) and the hardest case $>=2$ same-speaker intererences plus noise ($>=2ss+a$). The SNR level of noisy speech signals is uniformly distributed in $[-9, 6]$ dB. All the results of UltraSE are from a single model generated from the training dataset.

We compare UltraSE with 4 state-of-the-art SSE methods, PHASEN [12] (T-F domain method), SEGAN [28] (T domain method), AVSPEECH [6] (Audio-visual method), Conv-TasNet [4] (Speech separation method). For a fair comparison, we reimplemented PHASEN, SEGAN and Conv-TasNet and train and test them on the UltraSpeech dataset. PHASEN and SEGAN only use the $1s+a$ training set, since they are designed for speech enhancement, not separation. The results for PHASEN and SEGAN under $1s+a$ (see Table 4) is similar to the original work, which shows the correctness of our implementation. For the speech separation method, *i.e.*, Conv-TasNet, we first train and evaluate it in the "$2s$" environment to check the correctness of our implementation. Then, we use the "$2s+a$" dataset to train the model with the 2 speakers' clean speech as ground truth, and compare the results in other environments in Table 4. For AVSPEECH, since our data set does not have the video recordings, we directly use the results in [6] as baselines.

Compared to the state-of-the-art speech enhancement methods, UltraSE significantly improves the speech quality and intelligibility in both noisy and multi-speaker environments. Table 4 shows the testing results under all input SNR levels uniformly distributed in $[-9, 6]$ dB. UltraSE outperforms PHASEN and SEGAN across all the 4 metrics. In the $1s+a$ environment, UltraSE achieves an average 17.25 SiSNR (18.75 ΔSiSNR) and 3.50 PESQ. In other environments with multi-speaker interference, the ultrasound sensing modality plays a more prominent role, improving SiSNR by 6.04 dB and 9.77 dB on average over the 2 baselines respectively. Even for the hardest case $>=2ss+a$, UltraSE still achieves 8.97 dB SiSNR and 2.52 PESQ. In addition, UltraSE achieves slightly higher performance than AVSPEECH, likely because the ultrasonic features are sampled at finer time granularity than video frames, and can better align with the speech signals.

Environment	Methods	SDR	SiSNR	STOI	PESQ
$1s+a$	**UltraSE**	**17.14**	**17.25**	**0.87**	**3.52**
	PHASEN	15.63	15.20	0.82	3.05
	SEGAN	5.48	5.50	0.64	2.32
	AVSPEECH	16.0	/	/	/
	Conv-TasNet	12.23	12.58	0.76	2.48
$2s+a$	**UltraSE**	**10.55**	**10.65**	**0.76**	**2.80**
	PHASEN	5.20	5.22	0.65	2.23
	SEGAN	2.01	1.96	0.54	1.69
	AVSPEECH	10.1	/	/	/
	Conv-TasNet	10.23	10.38	0.74	2.40
$3s+a$	UltraSE	**10.88**	**10.94**	**0.76**	**2.81**
	PHASEN	5.14	5.15	0.66	2.15
	SEGAN	1.74	1.78	0.55	1.68
	Conv-TasNet	6.31	6.50	0.71	2.11
$>3s+a$	UltraSE	**12.10**	**12.17**	**0.78**	**2.66**
	PHASEN	5.13	5.13	0.67	2.14
	SEGAN	0.71	0.72	0.53	1.67
	Conv-TasNet	6.23	6.41	0.71	2.15
$>=2ss+a$	UltraSE	**8.90**	**8.97**	**0.72**	**2.52**
	PHASEN	5.03	5.05	0.62	2.10
	SEGAN	1.27	1.29	0.56	1.69
	Conv-TasNet	5.69	5.93	0.73	2.21
$2s$	UltraSE	14.85	14.86	**0.86**	**3.35**
	AVSPEECH	10.3	/	/	/
	Conv-TasNet	**14.98**	**15.02**	0.85	2.97

Table 4: UltraSE Micro Benchmark

Most of the existing speech separation methods can only work with limited number of interfering speakers ($2 \sim 3$) and without ambient noise [29, 30, 73, 76]. As shown in Table 4, when training the Conv-TasNet by using the "$2s+a$" dataset, Conv-TasNet achieves good performance in the "$2s+a$" and "$2s$" setup, but is not general in other sophisticated environments. In comparison, UltraSE outperforms Conv-TasNet by around 6 dB of SDR or SiSNR, 10% in STOI and 24% in PESQ, under the $>3s+a$ setup.

The scatter plot in Figure 11 shows the input and output SiSNR for each sentence in the testing dataset which includes all 6 environments. UltraSE consistently achieves high performance across different environments and sentences, with an average 14.75 dB SiSNR gain. Even in the worst case with -9 dB input, the enhanced speech achieves 8.86 dB SiSNR on average.

10.2 Ablation Study

We conduct an ablation study to better understand the performance of different design components in UltraSE. The testing dataset here

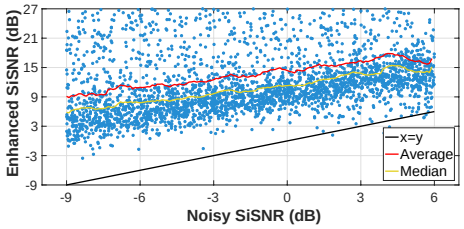


Figure 11: Noisy SiSNR v.s. Enhanced SiSNR

includes all the environments except the ">= $2ss + a$" which is not very common in practice. Table 5 summarizes the results.

"No T domain" represents the DNN model without the "T domain waveform speech enhancement". The results indicate that this module barely influences the STOI, a metric for speech intelligibility. But it helps gaining 0.46 dB SDR, 0.58 dB SiSNR, 0.12 PESQ respectively, which proves it can further improve the perceptual quality of the speech generated from the T-F domain multi-modal network.

"No cGAN" represents the model without the "cGAN-based cross-modal model training". All the metrics significantly improves when applying the cGAN, since our cGAN design forces the network to learn the correlation between the ultrasound and speech, which is the key principle behind the UltraSE design.

"No Fusion Network" means that the feature maps of ultrasound and speech signals are directly concatenated in the T-F domain without the fusion block. The performance slightly decreases, since the fusion block helps the multi-modal features to "cross-talk" with each other.

"No Ultrasound" represents the network without the ultrasound stream at the beginning of the network. The result becomes close to the traditional speech enhancement method without ultrasound sensing, *e.g.*, PHASEN.

10.3 System Efficiency

Time Consumption: We evaluate the run-time processing latency of UltraSE on 3 platforms, including a NVIDIA GTX 2020 (GPU), an Intel i9-9980 3.00GHz (CPU) and Samsung Galaxy S8 with Qualcomm Snapdragon 835 CPU (Smartphone). The first two correspond to the case where UltraSE is offloaded to a trusted cloud or edge server. Table 6 summarizes the results. The GPU server only experiences 14.85 ms latency which is acceptable for VoIP applications (150 ms maximum [77]). The smartphone case is measured by using Pytorch Mobile [78] on Samsung Galaxy S8. Note that the latest version of Pytorch Mobile [78] only supports *single-CPU processing without any GPU/NPU support*. Thus, the latency is relatively high (25.08 s to process 5 s speech), which is acceptable only for offline processing applications, *e.g.*, audio message and audio recording. There exists a rich literature [79] on improving DNN efficiency on smartphones, which demonstrated more than 50× latency reduction by using mobile GPU/NPU. We will explore such solutions for our future work. Also note that UltraSE needs to process the input in segments of 5 s due to the use of Bi-LSTM blocks. This means its SSE starts taking effect after a 5 s initial bootstraping period.

Energy Consumption: Our experiments show that a typical smartphone (Samsung S8) can continuously use UltraSE to record speech while emitting ultrasound signals for 60.57 hours with display off. Our measurement using Android Profiler [80] reveals that UltraSE's CPU load is 48.7% on average, and power consumption is

	SDR	SiSNR	STOI	PESQ
UltraSE (Testing data 96 kHz)	**13.10**	**13.21**	**0.80**	**3.01**
UltraSE (Testing data 48 kHz)	**13.08**	**13.18**	**0.79**	**2.99**
- No T domain	12.64	12.63	0.80	2.89
- No cGAN	10.80	10.85	0.77	2.60
- No Fusion Network	9.96	10.00	0.76	2.54
- No Ultrasound	7.78	7.68	0.70	2.39

Table 5: UltraSE ablation study.

	Preprocess	Stage 1	Stage 2
GPU	0.55 ms	12.02 ms	2.28 ms
CPU	0.05 s	1.38 s	0.26 s
Smartphone	0.25 s	23.02 s	1.81 s

Table 6: Inference time for processing 5 s speech.

at the level of "1" in between the scale of 0 to 3. When offloading to servers, the computational energy consumption becomes negligible. The only overhead is that UltraSE needs to upload the original 48/96 kHz sampling rate audio stream with both audible sounds and ultrasounds to the server, and then download the enhanced speech from the server. Our experiments show that Samsung S8 can continuously run UltraSE and upload/download the audio streaming via WiFi in the offloading mode for 10.82 hours. Server offloading may incur additional issues such as security, but this is beyond the scope of our current work.

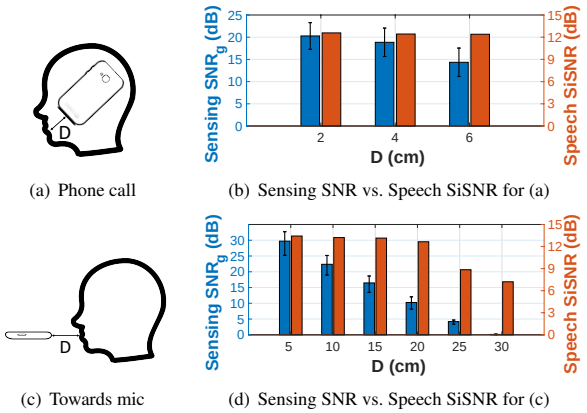
10.4 Generalization

Sampling Frequency: UltraSE model trained by 96 kHz sampling rate dataset can be directly used to enhance the testing speech recorded at 48 kHz sampling rate. The feature resolution at 48 kHz sampling rate is identical to the case at 96 kHz sampling rate as long as the FFT window length and hop length of ultrasound sensing features keep 85 ms and 10 ms respectively. Table 5 shows a negligible performance degradation when testing the 48/ 96 kHz sampling rate dataset on the 96 kHz sampling rate trained model.

Holding Styles: In the "Phone call" mode (Figure 12(a)), the user's face partially occludes the ultrasonic signals, so we train a model which is different from the "Towards mic" mode (Figure 12(c)). UltraSE can automatically select the model using the IMU-based holding style detection algorithm built into smartphones [81]. Our experiments show that, under −1.5 dB average input SNR, the performance of "Phone call" (12.47 dB SiSNR) is slightly lower than the "Towards mic" (13.12 dB SiSNR) due to the occlusion.

We further evaluate the sensitivity of each model under different mouth-to-mic distances. Figure 12(b) and Figure 12(d) show the average SNR of ultrasound (SNR_g) vs. the SiSNR of enhanced speech. For both holding styles, SNR_g well exceeds 10 dB, and speech SiSNR stays around 12 dB within 20 cm distance. *The experiment implies that the UltraSE model performs consistently as long as the mouth-to-mic distance remains within 20 cm.*

Motion interference: We measure the impacts of interference from 3 major motion artifacts, *i.e.*, respiration, hand gestures and walking. The experiments were conducted when the mouth is 15 cm and 2 cm away from the microphone in the "Towards mic" and "Phone call" mode, respectively. ***(i)*** The respiration frequency (∼30 bpm) is far less than the articulatory motions (> 10 Hz), so it creates negligible impacts on UltraSE. ***(ii)*** Hand gestures introduce similar Doppler effect as the articulatory motion [36, 40, 51], which may



(a) Phone call (b) Sensing SNR vs. Speech SiSNR for (a)

(c) Towards mic (d) Sensing SNR vs. Speech SiSNR for (c)

Figure 12: SNR of articulatory gestures.

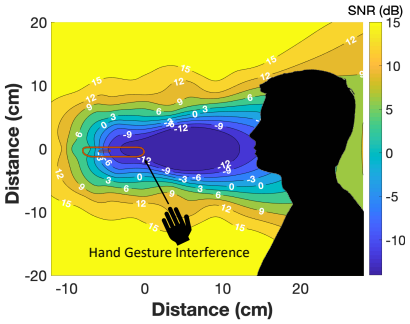
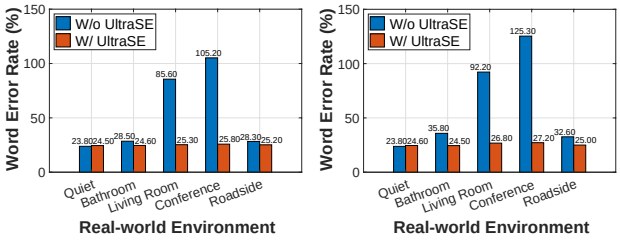


Figure 13: SNR_g under hand gesture interference

cause non-negligible interference. We measure the articulatory gestures' SNR_g under the pushing hand gesture interference. The SNR_g is sampled for each 2 cm in 7 different angles from 0° to 90°, at a step of 15°, close to the user's mouth. Figure 13 shows the spatial distribution [82] of SNR_g for the "Towards mic" mode. As long as the hand gesture is > 25 cm away from the mouth (which is typical in daily usage scenarios), the SNR_g remains above 10 dB which suffices for UltraSE (Figure 12). A microphone array can be used to focus on the user's mouth region to further mitigate interference [40], but this is beyond the scope of UltraSE. We omit the "Phone call" mode since the microphone is much closer to the mouth and the sensing SNR_g stays high. ***(iii)*** When other people walk nearby (0.8 m away), we found that SNR_g is barely impacted since the ultrasound volume is relatively low, and the user's mouth is much closer.

Overall, the articulatory gestures' SNR_g is sufficiently high (> 10 dB), and the UltraSE model is unaffected by the motion artifacts in daily usage scenarios.

Generalizations across smartphones: Different smartphones may have different speaker-mic layout. For example, the distances between the bottom microphone and speaker are 5 mm, 25 mm and 25 mm for Samsung S8, LG G8S ThinQ and VIVO X20 respectively. The high-frequency response of the speaker and microphone may also vary across phone models [83]. When applying the DNN model trained by the Samsung S8 dataset directly to LG G8S ThinQ and VIVO X20, the SiSNR of enhanced speech becomes 9.21 dB and 9.53 dB, respectively. which is lower than the same-phone case



(a) Phone call (b) Towards mic

Figure 14: Real-world Usage WER.

(13.21 dB), but still higher than the SiSNR without ultrasound sensing (7.68 dB). To maintain the optimal performance, a straightforward way is to perform a one-time training data collection for each phone model. Alternatively, we can enrich the UltraSpeech dataset with a diverse set of smartphones that cover the typical hardware configurations. This is left for our future work.

Real-world Usage Experiments: We asked the users to use UltraSE across 4 different real-world environments, *i.e.* 1) a bathroom environment with exhaust fan and running water noise (75 dBA on average); 2) a living room environment with television noise (55 dBA on average); 3) an indoor conference environment with conversation noise (60 dBA on average); 4) an outdoor roadside environment with vehicle noise (60 dBA on average). Unlike synthetic noisy speech, we can not capture the ground truth clean speech and evaluate the metrics like SDR, SiSNR, STOI and PESQ in these scenarios. Thus, to evaluate the performance of UltraSE for real-world usage, we use the ASR Word Error Rate $WER = \frac{S+D+I}{N}$ as the metric, where S, D, I, and N are the number of substitutions, deletions, insertions, totals of targeted user's spoken words respectively. Specifically, we asked the users to speak at least 50 sentences in the TIMIT speech corpus [9] across different environments. Figure 14 shows the WER with and without UltraSE across different environments. In non-speech noisy environments, *i.e.*, bathroom and roadside, UltraSE slightly improve the ASR speech recognition rate since ASR itself has the ability to mitigate background ambient noise interference. In speech noisy environments, *i.e.*, living room and conference, WER is higher than 100% since there exists many word insertions and substitutions introduced by non-targeted user's speech. UltraSE achieves significantly improvement in such cases since it is able to separate the desired speaker voice from noises by using ultrasound sensing.

11 CONCLUSION

We have demonstrated that ultrasonic sensing can serve as a complementary modality to solve the cocktail party problem. Our UltraSE system introduces general DNN mechanisms to enable such capabilities, *e.g.*, multi-modal multi-domain fusion network and cGAN-based training model based on a novel cross-modal Siamese network. UltraSE points to a novel direction that fuses wireless sensing capabilities to bring machine perception to a new level.

ACKNOWLEDGMENT

We would like to thank the anonymous shepherd and reviewers for their valuable comments. This work is partially supported by NSF CNS-1901048, CNS-1925767, Google Faculty Award, and a Samsung collaboration grant.

REFERENCES

[1] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proceedings of Interspeech*, 2019.

[2] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux. Improved mvdr beamforming using single-channel mask prediction networks. In *Proceedings of Interspeech*, 2016.

[3] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[4] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

[5] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending Speech Separation to Noisy Environments. *CoRR*, abs/1907.01160, 2019.

[6] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *Proceedings of ACM SIGGRAPH*, 2018.

[7] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. In *Proceedings of Interspeech*, 2019.

[8] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2015.

[9] John S Garofolo et al. Darpa timit acoustic-phonetic speech database. *National Institute of Standards and Technology (NIST)*, 1988.

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE ICASSP*, 2017.

[11] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[12] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of AAAI*, 2020.

[13] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng. Joint time-frequency and time domain learning for speech enhancement. In *Proceedings of IJCAI*, 2020.

[14] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 1984.

[15] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 1979.

[16] Yariv Ephraim and Harry L Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 1995.

[17] Guoning Hu and DeLiang Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 2004.

[18] Arun Narayanan and DeLiang Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceedings of IEEE ICASSP*, 2013.

[19] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2014.

[20] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2015.

[21] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of IEEE ICASSP*, 2015.

[22] Zhaoheng Ni and Michael I Mandel. Mask-dependent phase estimation for monaural speaker separation. In *Proceedings of IEEE ICASSP*, 2020.

[23] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. In *Proceedings of Interspeech*, 2017.

[24] Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, and Benoit Champagne. A fully convolutional neural network for complex spectrogram processing in speech enhancement. In *Proceedings of IEEE ICASSP*, 2019.

[25] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *Proceedings of IEEE ICASSP*, 2018.

[26] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *Proceedings of IEEE ICASSP*, 2018.

[27] Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *Proceedings of IEEE ICASSP*, 2019.

[28] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. In *Proceedings of IEEE ICASSP*, 2018.

[29] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Proceedings of IEEE ICASSP*, 2016.

[30] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proceedings of IEEE ICASSP*, 2017.

[31] Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, and Yuki Mitsufuji. Recursive speech separation for unknown number of speakers. In *Proceedings of Interspeech*, 2019.

[32] John R Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. In *Proceedings of NeurIPS*, 2002.

[33] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 2014.

[34] Ki-Seung Lee. Speech enhancement using ultrasonic doppler sonar. *Speech Communication*, 2019.

[35] Tom Barker, Tuomas Virtanen, and Olivier Delhomme. Ultrasound-coupled semi-supervised nonnegative matrix factorisation for speech enhancement. In *Proceedings of IEEE ICASSP*, 2014.

[36] Wei Wang, Alex X. Liu, and Ke Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of ACM MobiCom*, 2016.

[37] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of ACM MobiSys*, 2017.

[38] Ke Sun, Wei Wang, Alex X. Liu, and Haipeng Dai. Depth aware finger tapping on virutal displays. In *Proceedings of ACM MobiSys*, 2018.

[39] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of ACM MobiCom*, 2018.

[40] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. Rnn-based room scale hand motion tracking. In *Proceedings of ACM MobiCom*, 2019.

[41] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. Silenttalk: Lip reading through ultrasonic sensing on mobile phones. In *Proceedings of IEEE INFOCOM*, 2017.

[42] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of ACM CCS*, 2017.

[43] Yeonjoon Lee, Yue Zhao, Jiutian Zeng, Kwangwuk Lee, Nan Zhang, Faysal Hossain Shezan, Yuan Tian, Kai Chen, and XiaoFeng Wang. Using sonar for liveness detection to protect smart speakers against remote attackers. *Proceedings of ACM IMWUT (UbiComp)*, 2020.

[44] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. Silentkey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of ACM IMWUT (UbiComp)*, 2018.

[45] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. Echoprint: Two-factor authentication using acoustics and vision on smartphones. In *Proceedings of ACM MobiCom*, 2018.

[46] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *Proceedings of IEEE INFOCOM*, 2018.

[47] Li Lu, Jiadi Yu, Yingying Chen, and Yan Wang. Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones. *Proceedings of ACM IMWUT (UbiComp)*, 2020.

[48] Catherine P Browman and Louis Goldstein. Articulatory gestures as phonological units. *Phonology*, 1989.

[49] Kristin J Teplansky, Brian Y Tsang, and Jun Wang. Tongue and lip motion patterns in voiced, whispered, and silent vowel production. In *Proceedings of ASSTA ICPhS*.

[50] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of IEEE CVPR*, 2017.

[51] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of ACM CHI*, 2016.

[52] Xun Wang, Ke Sun, Ting Zhao, Wei Wang, and Qing Gu. Dynamic speed warping: Similarity-based one-shot learning for device-free gesture signals. In *Proceedings of IEEE INFOCOM*, 2020.

[53] Wenguang Mao, Jian He, and Lili Qiu. Cat: high-precision acoustic motion tracking. In *Proceedings of ACM MobiCom*, 2016.

[54] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of ACM MobiSys*, 2017.

[55] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *Proceedings of Usenix NSDI*, 2018.

[56] Ke Sun, Chen Chen, and Xinyu Zhang. "Alexa, Stop Spying on Me!": Speech Privacy Protection against Voice Assistants. In *Proceedings of ACM SenSys*, 2020.

[57] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.
[58] Ingo R Titze and Daniel W Martin. Principles of voice production, 1998.
[59] Brian B Monson, Eric J Hunter, Andrew J Lotto, and Brad H Story. The perceptual significance of high-frequency energy in the human voice. *Frontiers in psychology*, 2014.
[60] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of ACM MobiSys*, 2019.
[61] Fuming Chen, Sheng Li, Yang Zhang, and Jianqi Wang. Detection of the vibration signal from human vocal folds using a 94-ghz millimeter-wave radar. *Sensors*, 2017.
[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE CVPR*, 2016.
[63] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE CVPR*, 2018.
[64] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.
[65] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of IEEE CVPR*, 2006.
[66] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE CVPR*, 2017.
[67] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *Proceedings of IEEE ICASSP*, 2019.
[68] Amazon Transcribe, 2019. https://aws.amazon.com/transcribe/.
[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of IEEE ICCV*, 2015.
[70] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of IEEE ICASSP*, 2015.
[71] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *Proceedings of Interspeech*, 2018.
[72] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 2006.
[73] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *Proceedings of IEEE ICASSP*, 2018.
[74] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of IEEE ICASSP*, 2010.
[75] ITU-T Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.
[76] Yuan-Kuei Wu, Chao-I Tuan, Hung-yi Lee, and Yu Tsao. Saddel: Joint speech separation and denoising model based on multitask learning. *arXiv preprint arXiv:2005.09966*, 2020.
[77] Chris Lewis and Steve Pickavance. Implementing quality of service over cisco mpls vpns. *Selecting MPLS VPN Services*, 2006.
[78] Pytorch Mobile, 2020. https://pytorch.org/mobile/home/.
[79] Siqi Wang, Anuj Pathania, and Tulika Mitra. Neural network inference on mobile socs. *IEEE Design & Test*, 2020.
[80] Android Profiler, 2020. https://developer.android.com/studio/profile.
[81] Mayank Goel, Jacob Wobbrock, and Shwetak Patel. Gripsense: using built-in sensors to detect hand posture and pressure on commodity mobile phones. In *Proceedings of ACM UIST*, 2012.
[82] John D'Errico. Surface fitting using gridfit. *MathWorks file exchange*, 643, 2005. https://www.mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit.
[83] Best smartphones for audio, 2020. https://www.soundguys.com/best-smartphones-for-audio-16373.