NEC: Speaker Selective Cancellation via Neural Enhanced Ultrasound Shadowing

Hanqing Guo*, Chenning Li*, Lingkun Li, Zhichao Cao, Qiben Yan, Li Xiao
Department of Computer Science and Engineering, Michigan State University
{guohanqi, lichenni, lilingk1, caozc, qyan, lxiao}@msu.edu
*These authors contributed equally

***Abstract*—In this paper, we propose NEC (Neural Enhanced Cancellation), a defense mechanism, which prevents unauthorized microphones from capturing a target speaker's voice. Compared with the existing scrambling-based audio cancellation approaches, NEC can selectively remove a target speaker's voice from a mixed speech without causing interference to others. Specifically, for a target speaker, we design a Deep Neural Network (DNN) model to extract high-level speaker-specific but utterance-independent vocal features from his/her reference audios. When the microphone is recording, the DNN generates a shadow sound to cancel the target voice in real-time. Moreover, we modulate the audible shadow sound onto an ultrasound frequency, making it inaudible for humans. By leveraging the non-linearity of the microphone circuit, the microphone can accurately decode the shadow sound for target voice cancellation. We implement and evaluate NEC comprehensively with 8 smartphone microphones in different settings. The results show that NEC effectively mutes the target speaker at a microphone without interfering with other users' normal conversations.**

I. Introduction

Voice recording is an essential information-sharing approach, which is benefiting many aspects of our daily life. Nowadays, smartphones and Internet-of-Things (IoT) devices equipped with microphones allow people to record voice anytime and anywhere. However, the growing presence of unauthorized microphones has led to numerous incidences of privacy violations. Off-the-shelf microphones are widely available and can be deployed to steal users' biometric traits (e.g. voiceprints) or private conversations. Thus, unauthorized voice recording has become a serious societal issue [1].

Recent studies [1], [2] attempt to disrupt unauthorized voice recording by emitting an ultrasonic scrambling noise wave (i.e., a jamming signal) to obfuscate the superposed voice. However, the scrambling noise wave is generated using the low-level acoustic signal features that are irrelevant to the speaker identity. Consequently, other benign microphones in the reception range will also be jammed, most of the time undesirably. In fact, the use of such voice jammer in public spaces is prohibited and unlawful (violation of *47 U.S.C. § 333*), since it poses serious risks to critical public safety communication. Moreover, if the attacker learns the frequency pattern of the scrambling noise wave, the attacker can deploy an additional microphone to nullify the noises and record them illegally. To allow users to secure their voices lawfully without intervening in others' microphones/recorders usage,

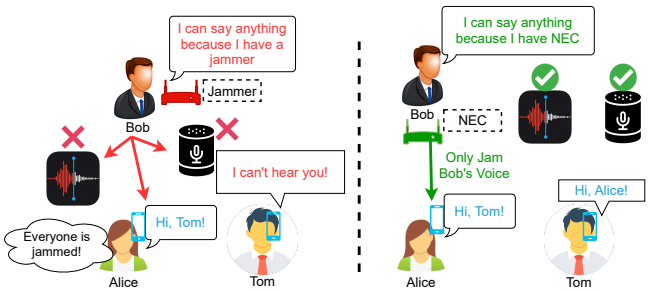


Fig. 1. NEC cancels the target speaker's (e.g. Bob's) voice without intervening other communications.

we propose *NEC (Neural Enhanced Cancellation)*, which only jams *a specific target speaker's voice* from the recording of any microphones nearby.

Figure 1 illustrates the necessity of deploying NEC instead of the commercial audio jammer. Consider that Bob is initiating a private conversation in a public area (e.g., cafe or work office), in order to prevent his speech from being leaked, he turns on a commercial jammer to obfuscate all the surrounding input devices. The left sub-figure shows that during the attack, other applications such as voice reminder, voice assistants, and phone calls are all effectively disabled by Bob's jammer, which is not only unlawful but also annoying to other users. In contrast, if Bob deploys NEC, only his speech is imperceptible by the others' microphone, while other users can still safely use their voice applications as usual.

Generally, NEC is composed of a microphone, a neural network model, and an ultrasonic speaker. Figure 2 entails the components of NEC, the red lines demonstrate the target speaker's voice (e.g., Bob's voice), while the green lines represent Bob's irrelevant voice (e.g., Alice's voice, background noise, and model processed voice). Our goal is to make Bob's voice unrecognized/unrealized on Alice's phone/recorder. At the very beginning, the microphone perceives both Bob's and Alice's voice. Then, we feed the mixed audio to our proposed deep neural network (DNN) model. Note that, comparing to the existing systems that utilize low-level acoustic signal (such as Gaussian noise or scrambling noise), we use the DNN model to extract the high-level speaker-specific vocal features for differentiating Bob's voice from the mixed recordings. The output signal of the DNN model is marked as *shadow sound*, which is then modulated to ultrasonic frequency to

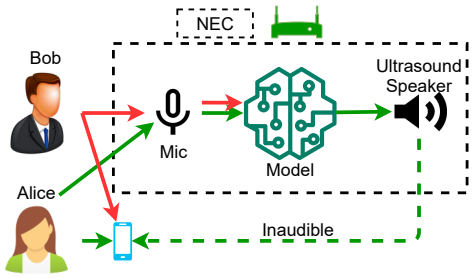


Fig. 2. The voice stream flow of NEC

make it inaudible to other users. Subsequently, Alice's phone will receive a combination of Bob's voice, Alice's voice, and the inaudible shadow signal generated by NEC. The signal combo will yield a mostly undisturbed sound for Alice.

We have four main design goals as follows:

- *Utterance-independent Vocal Feature Extraction.* For a target speaker, we need to train our DNN model with the speaker's reference audio before the deployment. To alleviate the training overhead across different scenarios, the speaker's vocal features should be independent of his/her utterances. As such, we can deliver a one-fits-all DNN model, which is trained once and easily transferred.
- *Microphone-aware End-to-end DNN Training.* The shadow sound is superposed on the speaker's voice at the microphone. To make the superposition more effective, we need to design an end-to-end training pipeline that aims to maximize the effectiveness of the superposed shadow sound.
- *Low-latency Shadow Sound Generation.* We will modulate a shadow sound onto the ultrasonic frequency to make it inaudible. However, the processing delay may degrade the shadowing efficiency due to the feature mismatch between the speaker's voice and the generated shadow sound. Thus, we need a DNN model that is computationally efficient.
- *Synchronization-free.* To cancel Bob's voice on other devices, it typically requires the synchronization of the arrival time of the shadow sound, Bob's sound, and Alice's sound. However, it is challenging to synchronize them (without modifying Alice's devices). Therefore, we need a synchronization-free approach for voice cancellation.

To achieve all four goals, we first explore the human vocal principle and observe the speaker-specific but utterance-independent formants of the audio spectrogram from ten speakers using various speech contents. Then, we design a DNN model to generate a shadow sound by imitating the superposition of multiple waves at the microphone. The DNN includes the speaker encoder and selector for feature reference and extraction. Moreover, we analyze the delay bound and compress the DNN layers to guarantee that the processing delay can meet the requirement on various devices (e.g., mobile, Raspberry-Pi).

We implement NEC using commercial off-the-shelf (COTS) ultrasound transceivers and evaluate its performance in different real-world scenarios. In the experiment, we run a benchmark testing using a public speech corpus dataset and two real-world case studies. The evaluation results demonstrate that NEC effectively mutes the target speaker at a microphone by causing a 200% word error rate under Google's voice-to-text service without interfering with others' conversations. Our contributions are summarized as follows:

- NEC is the first practical speaker selective cancellation system, which aims to protect the target speaker's voice without interfering with other microphones in presence.
- We explore the human vocal principles and design a DNN model to imitate the superposition of waves at the microphone, which produces the speaker-specific but utterance-independent shadow audio in real-time.
- We implement NEC and extensively evaluate its performance with the benchmark and user studies. The results show its superior performance in comparison with state-of-the-arts systems. The demos can be found on our project website: https://nec-app.github.io/.

The paper is organized as follows. §II summarizes the related work. We then introduce the background and preliminary of our vocal system in §III, followed by the system design of NEC in §IV. The implementation and evaluation are further presented in §V and §VI. We conclude the paper in §VIII.

II. Related Work

Microphone Jamming: Microphone jamming has been proposed [1]–[3] to protect private conversations. To avoid the recording of private conversations, a pre-configured audio jammer is deployed to emit the scrambling noise waves to disrupt the speech recording. Specifically, Chen et al. [3] adopt the white noise to distort the microphone recordings, while Tung et al. [2] explore the sound masking with the specially designed scramble noise to obfuscate the spoken sensitive information. Patronus [1] emits ultrasound to generate the scrambling waves at the recorder without introducing human-sensitive noise. In contrast, rather than canceling and jamming by the low-level signal features (e.g., frequency, phase), we use high-level human vocal features to generate a shadow sound for speaker-selective jamming.

AI-augmented Speaker Diarization: AI plays important role of processing signal [4]–[7]. Recent studies [8]–[10] propose AI-based speaker diarization, a process to partition multi-speaker audio into homogeneous single speaker segments based on the speaker identity. It effectively solves "who spoke when" in a multi-speaker scenario. Several audio embedding models have been proposed for speaker-specific feature extraction, including speaker factor [11], i-vector [12], [13], and d-vector [8], [9], [14]. Based on these features, a number of classification models have been designed to extract the speaker-specific embedded audios, such as clustering algorithms [8], [12], [13], DNN model [9], [14], and even an integrated model with visual information (e.g., lip movement and face recognition) [10], [15], [16]. However, these methods cannot be adopted in our scenario. First, all existing speaker diarization models are used for post-processing after the audio is recorded, but we need to deal with voice cancellation in an end-to-end fashion. Additionally, the processing delay is

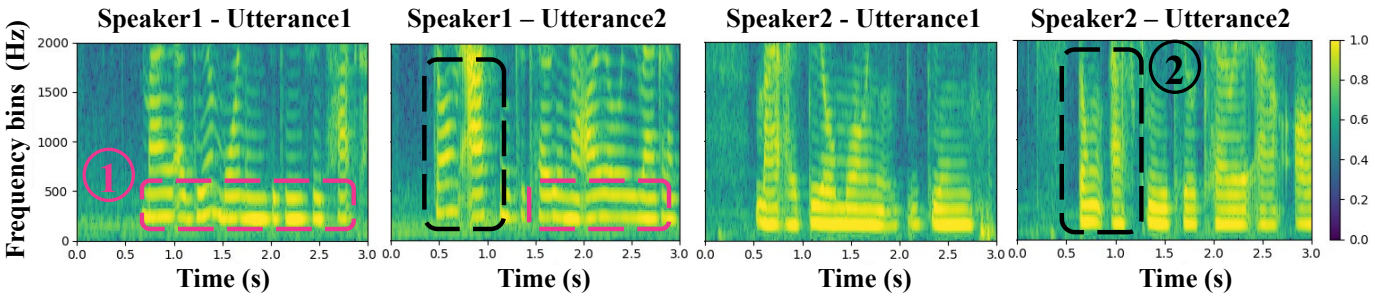


Fig. 3. Distribution of formants across spectrograms, representing the speaker-specific but utterance-independent timber pattern. Utterance1: "My ideal morning begins with hot coffee." Utterance2: "Don't ask me to carry an oily rag like that."

an important factor to guarantee an effective shadow sound generation, which has been ignored by these post-processing models. In this work, we design the adaptive features, DNN structures, and training methods to realize an end-to-end voice cancellation system to protect a target speaker's voice.

III. Background of Vocal System

Observations: To illustratively show the harmonic components of a sound induced by the physical vocal system, we first collect four audios from two volunteers. Each volunteer records two audios of their utterances of two sentences: "*my ideal morning begins with hot coffee*" and "*don't ask me to carry an oily rag like that*". For each audio, we derive the corresponding formants [17] via FFT for each frame with a duration of 20 ms. The rationale is that the duration of a typical phoneme is longer than 20 ms, representing the maximal frame length [18]. Thus, each frame is dominated by the harmonic components of sustained tones, i.e., the number and relative intensity of the upper harmonics in the sound.

The results are presented in Figure 3. We can observe the consistent formants of each speaker with various spoken contents. For example, the similarity of the resonant frequency and the relative intensity of formants of different utterances from the same speaker can be observed in area ①, shown in red boxes. Hence, these characteristics are utterance-independent. Conversely, area ② in black boxes implies the distinct distribution of speaker-specific formants, which can also be observed across multiple spectra of various frames.

Validation: Based on the observations above, the remaining challenge is to quantify the utterance-independent but speaker-specific feature in audio spectrograms (i.e., area ① and ②), namely timbre pattern [19]. To guarantee the phonetically balanced state in the timbre, we first average the dynamic influence of individual phonemes by computing the averaged spectrum for all frames, namely **Long-time Average Spectrum (LAS)** [18], [20]. LAS can average out the dynamic characteristics associated with various phonemes such as the motion of the articulators [21]. Suppose the spoken content for each person is divided into M frames with the duration T in time, the LAS $F(w)_{LAS}$ can be formulated by averaging the spectrum of each frame:

$$F(w)_{LAS} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{F}(f_m(t)), \tag{1}$$

where $\mathcal{F}$ denotes the FFT, and $f_m(t)$ is the frame waveform signal with the duration T.

To visualize the distinctive LAS features for different speakers, we compute the LAS of four speakers (e.g., A, B, C, D), with every speaker requested to read the same sentence (e.g., "don't ask me to carry an oily rag like that"). The results in Figure 4 show that every speaker's LAS feature is unique even when their speech contents are the same. The distinctiveness of LAS features demonstrates the potential of differentiating voices from multiple speakers. To further verify the utterance-

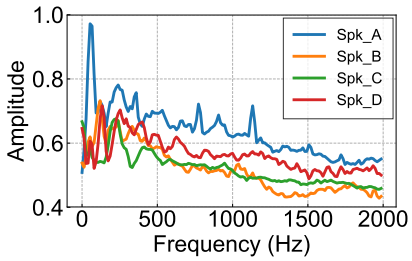


Fig. 4. LAS results from four speakers.

independent but speaker-specific timbre pattern in our computed LAS, we compute the Pearson correlation and deliver the correlation matrix across different speakers and spoken contents. Specifically, we first collect ten different utterances from four speakers (e.g., A, B, C, D) and compute the Pearson correlation across $F(w)_{LAS}$ [18]. As shown in Figure 5, the correlation coefficients for the same speaker with different utterances can reach up to 0.96 on average, whereas they are generally below 0.75 across speakers, even with the same utterances. The former implies the consistency of spectrum across various spoken contents for the same speaker, while the latter indicates the distinct timbre patterns of different speakers, which demonstrates the feasibility of using LAS to quantify the timbre patterns from audio spectrograms of different speakers.

IV. NEC System Design

As shown before, the voice signals from different human speakers present different spectrum features. Meanwhile, for the same speaker, the spectrum features are consistent across different spoken contents. The remaining challenge is to generate a speaker-specific shadow sound from these spectrum features.

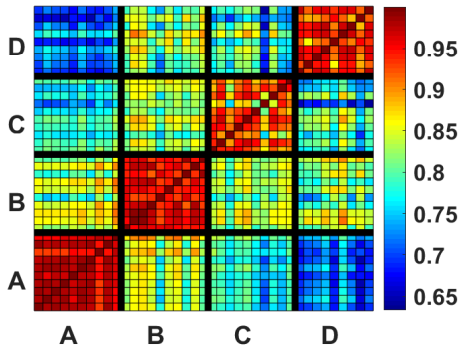


Fig. 5. Pearson correlation matrix of the long-time average spectrum of 10 different utterances from 4 speakers.

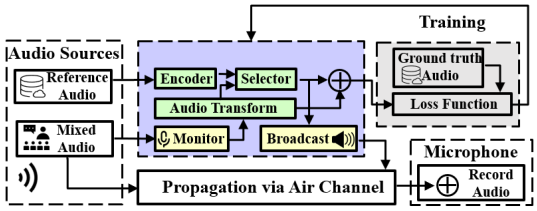


Fig. 6. Overview of NEC, which includes the software (green) and hardware (yellow) design as well as the training stage (grey) of our system.

A. System Overview

System Pipeline: The goal of NEC is to cancel Bob's voice in the wild (e.g., no one can record Bob's voice in their microphone, and no one is affected by Bob's NEC devices). However, passively canceling Bob's voice on Alice's recorder is very challenging. Prior work [22] takes a great effort to estimate the arrival of Bob's voice through a wireless channel, and compute the inverse signal of Bob's voice before the acoustic signals of Bob arrive. Next, they synchronize Bob's voice with the crafted inverse signal to perform the voice cancellation using rigorous procedures. However, such design relies on the speed difference between wireless signal and acoustic signal. In short-range scenario (e.g., Bob is close to Alice), their work will no longer be effective since the arrival time could be very close. Instead of generating the inverse signal by the prior knowledge of Bob's speech, we propose a *superposition* method to reduce the strength of Bob's sound signals received by Alice's microphone. In other words, NEC produces a shadow signal to be mixed with Bob's voice, which will distill Alice's sound on her microphone.

Figure 6 shows an overview of NEC's architecture from audio sources (left block) to the (Alice's) recording microphone (bottom right block), which serve as inputs and outputs, respectively. The reference audio is the historical recordings of the user, which is prepared to assist the DNN model to separate the voice stream of the user. The mixed audio refers to the audios containing Bob's voice and others' (background) voices. The output of NEC model is a shadow signal transmitted by an ultrasound speaker. The mixed audio and the shadow signal combined together to form the recordings on Alice's microphone. We assume that Alice receives the same mixed audio as the one collected by the NEC's microphone in proximity.

To create a general neural-enhanced framework, we first train NEC at the spectrogram level (top right block) in the offline training stage, where a $\bigoplus$ operation in the purple block represents the audio spectrogram superposition that combines the outputs of the **Selector** and **Audio Transform** modules. The functionality of **Selector** is to generate spectrogram that exclude Bob's sound, and the **Audio Transform** serves to transform the waveforms into spectrograms. Then, we convert the shadow spectrogram induced by our selector into inaudible ultrasound wave via **Broadcast**. The shadow wave will propagate through the air channel along with the mixed wave (§IV-C). $\bigoplus$ inside the Microphone block indicates the wave superposition of the mixed audio and broadcasting shadow sound at the microphone. Due to the equivalence of audio superposition for wave and spectrogram, the effectiveness of wave superposition is guaranteed for testing scenarios, as mixed audio and shadow sound arrive simultaneously at the microphone. The superposed wave corresponds to the recorded audio which effectively hides the target's (e.g. Bob's) voice.

Training Stage: The purpose of model training is to generate spectrogram that not caused by Bob's voice for any speech context with Bob. To achieve that, we manually craft mixed audios which contain Bob's voice and other speakers' voice, and use our selector to generate Bob's irrelevant spectrogram. To train NEC, we first provide a pre-trained **Encoder**, which generates the speaker-specific d-vector [8], [9] from the reference audio (e.g., 3 audio instances lasting 3 seconds) as reference input for the selector. Meanwhile, the mixed audio is processed by the audio transform, which generates a mixed spectrogram as another input of the selector. The rationale of using spectrogram has two folds. First, the **LAS** feature is effective in distinguishing different speakers based on our previous observation (§III); second, the calculation of LAS refers to the procedure of calculating the average spectrum for audio clips, which can be unfolded across multiple clips as a spectrogram. We directly feed the mixed spectrogram into our selector, along with the d-vector extracted from the reference input. This can boost the accuracy of DNN in extracting the high-level speaker-specific but utterance-independent vocal features from the mixed sound (§IV-B1).

Overshadow Stage: A key property of NEC is its generalization for deployment in the wild. First, rather than the cumbersome model-retraining and data collection, only 3 audio instances lasting 3 seconds are required by our one-fits-all model for a new user enrollment. Second, due to the linearity of the Fourier Transform (§IV-B2), we can transfer the spectrogram superposition into the wave superposition of audios at the microphone to guarantee the overshadowing performance. Finally, to avoid the disturbance during the overshadowing of NEC, we further convert the shadow spectrogram into inaudible ultrasound (§IV-C1).

B. Neural Enhanced Selective Speaker Cancellation

In this section, we present the design of NEC's DNNs, which aim to utilize the utterance-independent but speaker-

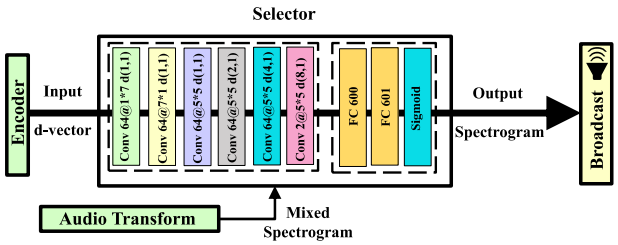


Fig. 7. NEC's DNN Selector generates the utterance-independent but speaker-specific shadow spectrogram by imitating the superposition of waves at the microphone.

specific features to generate the shadow sound. NEC incorporates an efficient **selector** to produce a shadow spectrogram, and further add Bob's voice through overshadowing onto the mixed spectrogram.

1) Architecture of DNN:

Encoder: The encoder module follows the design of **d-vector** in prior studies [8], [9], [23]. This module takes the reference audio of a target speaker as input and produces a speaker-specified embedding to allow the **selector** to filter out the target speaker's voice from the mixed audio spectrogram.

Selector: The purpose of the selector is to produce a shadow spectrogram and further hide Bob's voice by superposing the shadow spectrogram onto a mixed spectrogram. As shown in Figure 7, the selector takes d-vector and the mixed spectrogram as input. We formulate the mixture of spectrogram as follows:

$$S_{mixed} = |\sum_{n=-\infty}^{\infty} x_{mixed}[n]W[n-m]e^{-j\omega n}|, \qquad (2)$$

where the n-sample mixed audio in $\mathbb{C}^n$ is converted into a spectrogram with t *sampling points* and f frequency bins in $\mathbb{R}^{t\times f}$. $W[n-m]$ is the Hann window, and m is the window size. More specifically, the mixed spectrogram is composed of Bob's voice S_{Bob} and background voice S_{bk} (e.g., Alice's voice) as follows:

$$S_{mixed} = S_{Bob} + S_{bk}. \qquad (3)$$

In practice, the input audio lasts 3 seconds with a sampling rate of 16 kHz. The number of samples is 48,000. Also, we set the FFT size as 1,200, resulting in 601 frequency bins. The window length and hop length are 400 and 160, respectively, which generates 299 frames. Then, the shape of S_{mixed} is 601×299, denoted as (F, T), the frequency resolution and frame resolution are 13.31 Hz and 25ms with 15ms overlap. We transpose the mixed spectrogram for further processing and denote the shape of the transposed spectrogram as (T, F).

With the mixed spectrogram and d-vector in hand, we then utilize them to design a neural network based on our observation in §III. Revisiting Fig. 3, the frequency distribution of formants [17] and harmonic determine the identity of a given speech (i.e., LAS sufficiently captures the speaker characteristics). Our design goal of the selector is to capture these characteristics with multiple layers of CNNs. Prior to building the neural network structure, we propose the requirement for our DNN model as: **1)** the selector should be able to capture the formants and harmonic feature; **2)** the selector should consider the consistency of the frequency distribution within the same voice source.

In our DNN design, we only focus on the first three formants since we observed that the lower orders of harmonic have more energy and are more representative for a single speaker. As the bandwidth of the first three formants ranges from 33 Hz to 79 Hz [24], we design the first convolutional layer with 64 filters, whose size is 1×7. The rationale of using this flat filter is to convolve the frequency domain information (F). In particular, each filter covers 93.17 Hz, which is enough to cover the individual formant bandwidth as mentioned previously. Another 64 filters follow, whose size is 7×1, which can cover 115ms (determined by the frame resolution) time-domain feature (T). It is worth mentioning that the length of phoneme varies from 5 ms to 670 ms based on existing vocal research [25], and the average reading speed for an adult is 184±29 words per minute [26], i.e., 281∼387ms per word. So the second convolutional layer only serves to explore the detailed information of the phoneme level.

To further incorporate both F domain and T domain features, we apply a sequence of (5×5) convolutional layers with the dilation ranging from (1,1) to (8,1). The dilation setting on T domain extends the effective range of filters from (5×5) to (5×40), corresponding to 85ms to 610ms. This range covers a few words and meets our **R2** for considering the consistency of frequency distribution. While other studies [8], [9], [14] also add extra layers (e.g., LSTM, CNN with larger filter size and dilation shape) for speaker separation task, we consider that those layers play a less important role. For example, a larger filter will introduce irrelevant frequency information and long time span data, when the speaker merely adjusts his/her formants frequency when speaking a single word or a short sentence.

The output of CNNs has the shape of (T, 2×F) since we add a padding layer before the convolutional layer to maintain the shape of feature domain consistency, where 2×F comes from two filters in the last CNN layer. After that, the d-vector is repeatedly concatenated to the output of the last convolutional layer in every time frame. The fused feature embedding will be fed into two fully connected layers. As a result, we get a (T, F) shadow spectrogram. Figure 7 shows the detailed flowchart of our selector. In total, we only use 6 CNN layers and 2 Fully Connected (FL) layers for the selector model. Compared with the existing models such as [8], [9], [14], our model is computationally efficient by eliminating the redundant modules (e.g., LSTM, CNN with larger filter size and dilation shape) unrelated to our research goal.

2) Spectrogram-based Overshadowing: In the overshadowing process, we first feed mixed spectrogram and d-vectors into our selector. Then, we deliver the generated shadow spectrogram to be superposed with the received mixed audio at the microphone.

Shadow Spectrogram Generation: From the point of view of the microphone, the received mixed audio and shadow sound

should be superposed to imitate the over-the-air overshadowing at the microphone, formulated as $\mathbf{x}_{record} = \mathbf{x}_{mixed} + \mathbf{x}_{shadow}$. Those vectors represent time-series samples of mixed audio, shadow sound, and recorded audio, respectively.

Through crafting the shadow sound, our goal is to make the recorded audio as close as background audio (e.g., Alice's sound or environmental noise). A straightforward idea is to optimize the shadow sound directly with the audio-level superposition in the time domain. However, there are two drawbacks to the temporal wave superposition. First, the temporal waveform is less representative than a spectrogram. Second, since the output of our selector is the shadow spectrogram, an Inverse STFT module should be introduced to convert spectrogram to waveform ahead of the loss function, which results in the gradient vanishing issue for back-propagation based on our evaluations. Therefore, we use a shadow spectrogram from our DNN selector for the following overshadowing processing.

Superposition for Audio Wave and Spectrogram: The linearity of the Fourier Transform guarantees the equivalence of the temporal wave and spectrogram superposition, which can be denoted as follows:

$$\mathcal{F}[\sum_{i=1}^{n} a_i x_i(t)] = \sum_{i=1}^{n} a_i X_i(w), \textit{for } \mathcal{F}[x_i(t)] = X_i(w), \quad (4)$$

where $\mathcal{F}$ denotes the Fourier Transform and $x(t)$ is the temporal waveform signal. Given the linearity of Fourier transform with a coefficient a_i, we can convert temporal wave superposition into a linear combination of spectrograms as follows:

$$S_{record} = S_{mixed} + S_{shadow}. \quad (5)$$

The S_{record}, S_{mixed} and S_{shadow} correspond to the spectrogram of recorded audio, mixed audio and shadow audio, respectively. To avoid the gradient propagation issues and expedite the convergence of DNN, the shadow spectrogram from the speaker selector is first normalized before being superposed with the mixed spectrogram. To allow the recorded magnitude to eliminate Bob's voice while retaining other's (e.g., Alice's) voice components, we design the loss function:

$$Selector^*_{opt} = \underset{Selector*}{\operatorname{argmin}} ||S_{record} - S_{bk}||_2^2, \quad (6)$$

where the $Selector*$ denotes the model parameters of our DNN selector, and the S_{record} is the sum of mixed spectrogram and shadow spectrogram. Using the back-propagation with the L_2 norm loss, we can derive an optimal parameter $Selector^*_{opt}$ for our DNN selector, which will output an optimal shadow spectrogram S_{shadow}. This optimization ensures the resulting S_{record} to be as close to S_{bk} as possible.

C. Overshadowing Over the Air

1) Inaudible Shadow Sound Generation: Given the shadow spectrogram generated by the trained DNN selector, we can apply the inverse STFT on the shadow spectrogram to derive the shadow sound wave for further broadcasting. To make the shadow sound inaudible for privacy concerns and deployment convenience, we resort to the non-linear property of microphones [27], [28] to modulate the emitted shadow wave, via the **Broadcast** module in Figure 6.

Non-linearity of Hardware: The non-linearity property of microphone hardware represents the physical limitations of the diaphragm and the pre-amplifier, which amplify the signals in a non-linear manner. Mathematically, given an input signal V_{in} to microphone, the output signal V_{out} of the commercial amplifier within the microphone is not amplified linearly, i.e., $V_{out} \neq A_1 V_{in}$, where A_1 is the gain for input. Instead, the output signal is $V_{out} = A_1 V_{in} + A_2 V_{in}^2 + A_3 V_{in}^3 + \cdots$. We focus on $A_2 V_{in}^2$ of the non-linear V_{out} by ignoring (relatively small) higher-order components [27], [29]. Without loss of generality, let $m(t)$ be a simple tone, e.g., $m(t) = cos(2\pi f_m t)$. We then up-convert the baseband signal m_t onto a carrier with central frequency $f_c > 20kHz$. The modulated signal can be written as follows with the power coefficient α:

$$V_{in} = (cos(2\pi f_m t) + \alpha) cos 2\pi f_c t. \quad (7)$$

Since f_c is in the inaudible frequency range, the modulated signal V_{in} cannot be heard by humans. Given the non-linearity effect, the recorded signal V_{out} will not only contain the linear component $A_1 V_{in}$, but also the non-linear component $A_2 V_{in}^2$ representing the inaudible but recorded component, denoted as follows:

$$\begin{aligned} V_{in}^2 &= (cos^2(2\pi f_m t) + \alpha^2 + 2\alpha cos(2\pi f_m t)) cos^2(2\pi f_c t) \\ &= \sum_i \lambda_i cos(2\pi f_i t) + \mu, \end{aligned} \quad (8)$$

where f_i denotes frequency components at f_m, $2f_c$, $2f_m$, $2(f_m \pm f_c)$, $f_m \pm 2f_c$ and μ is a consequent constant. Given the low-pass filter in the COTS microphone, we can eliminate the high frequency components while retaining the f_m components, where f_m is the baseband frequency of $m(t)$ perceived by a microphone.

Shadow Sound Broadcast: Then, we can encode our shadow wave $\mathbf{x}_{shadow}$ into an inaudible frequency range by modulating it with a carrier whose central frequency is f_c. The broadcast shadow wave can be computed as follows:

$$\mathbf{b}_{shadow} = \mathbf{x}_{shadow} \times cos 2\pi f_c t, \quad (9)$$

where $\mathbf{b}_{shadow}$ refers to the inaudible shadow wave, and $\mathbf{x}_{shadow}$ is induced by S_{shadow} from inverse Fourier transform process. More details on modulation setting can be found in §VI-D which evaluates the impact of seven different types of mobile devices.

2) Latency Tolerance for Overshadowing: **Offset Issue in the Wild:** Ideally, the broadcasting shadow wave and the mixed wave arrive at the recorder simultaneously. However, in the real-world scenario, the shadow sound may arrive at the recorder with a slight timing offset due to the propagation delay and system delay, which can be formulated as below:

$$t_{offset} = t_{AB} + t_p + t_{BC} - t_{AC}, \quad (10)$$

where t_{offset} is composed of the propagation delay $t_{AB} + t_{BC} - t_{AC}$, and system delay t_p refers to system process delay,

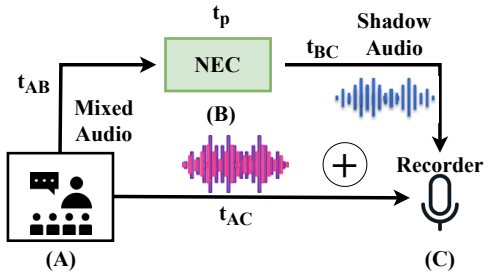


Fig. 8. The illustration of the influence with time offset between the mixed audio and the shadow sound.

illustrated in Figure 8. Note that the shadow sound received by the recorded is $\mathbf{x}_{shadow}$, originating from $\mathbf{b}_{shadow}$ based on the non-linearity of hardware (§IV-C1).

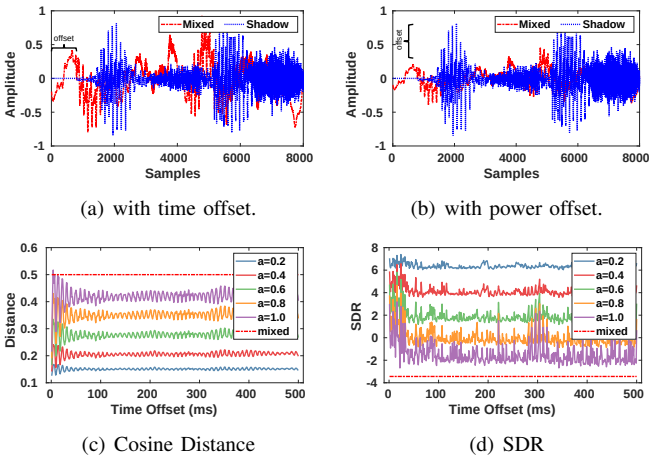
Besides the time offset, we still have the power offset between $\mathbf{x}_{mixed}$ and $\mathbf{x}_{shadow}$, which is introduced by the different attenuations for over-the-air transmission and determined by the original power of mixed audio source and NEC system. To analyze the impact of both offsets, we reformulate the temporal wave superposition to represent the recorded signal as follows:

$$x_{record}[n] = a x_{mixed}[n] + x_{shadow}[n - t_{offset}],$$
$$x_{shadow}[n] = 0 \quad for \quad n < 0, \tag{11}$$

where $x_{record}[n]$ is the value of n_{th} sample, a is the coefficient to represent power ratio between x_{mixed} and x_{shadow}. We set x_{shadow} equal to zero when the shadow sound does not arrive at the recorder side. Figure 9 demonstrates the time offset and power offset between the mixed signal and shadow signal, respectively. We first collect the mixed signal, which is a 16 kHz recorded audio from a speaker in a noisy car, and extract the first $8,000$ samples. The shadow sound is transmitted through an inaudible frequency carrier and captured by the recorder. Specifically, Figure 9(a) illustrates the 800 samples offset, corresponding to 50 ms in time delay. We can observe that the mixed audio will be superposed by the shadow wave if a is 0.5, as shown in Figure 9(b).

Tolerance Analysis: Theoretically, we expect our time offset to be within the one-word duration, such that the offset will not be perceived by the user. Since the average reading speed for an adult is 184 ± 29 word per minute [26], i.e., $281 \sim 387$ms per word, our theoretical analysis shows that if the real-time offset is limited within around 300ms, the recorded audio would be clear to humans. For the power offset, a recent study [30] shows that performing wireless overshadowing attack requires a power difference between overshadow signals and legitimate ones to be as small as 3 dB, i.e., $a = 0.5$.

Quantitative Analysis: We also measure the cosine distance and Source to Distortion Ratio (SDR) [9], [31] along with different settings. Figure 9(c) shows the cosine distance between the recorded signal at different time offsets, and different power coefficients a of the background (e.g., Alice's sound) at the same time offset. For comparison, we also compute the cosine distance for mixed signal and background audio. The shorter cosine distance indicates a greater similarity between record audio and background audio. We can observe that: first,



(a) with time offset. (b) with power offset.

(c) Cosine Distance (d) SDR

Fig. 9. Time offsets of overshadowing between the mixed and shadow audios.

the lower power coefficient leads to shorter cosine distance; second, the time offset within 500ms does not affect the cosine distance significantly; third, if we emit higher power shadow audio than the power of mixed audio (e.g., $a < 0.6$), then it is guaranteed that the recorded audio has high cosine similarity with the background sound; fourth, by applying our shadow audio, the similarity between record audio and background audio increases. The mixed audio without the addition of shadow audio has the largest cosine distance.

Similarly, Figure 9(d) calculates the SDR between different record signals (with different power offset) and the background audio. The higher SDR indicates less distortion of the audio. A reference result is generated by the SDR of mixed signal and background audio. This result not only supports our previous observations in Figure 9(c) that the lower power coefficient is better but also reveals that the smaller time offset (within 50ms) results in higher SDR.

Based on our theoretical and quantitative analysis, we identify the requirements for successfully implementing NEC in a real-world scenario. That is, *the time offset introduced by propagation delay and system latency should be limited to* 300*ms. To superpose the shadow signal on the mixed signal, the power coefficient is expected to be lower than* 0.6*, in which case the desired record audio will be perceived by the recorder.* We further justify the requirements in §VI.

V. Implementation

Experimental Setup: Figure 10 presents the implementation and experimental settings of NEC. In NEC, the input mixed audio is first collected and processed by our trained encoder and selector DNNs, which produces the corresponding shadow spectrogram in Figure 7. Then, we transform it into audios and up-convert it into the ultrasound carrier frequency, making it inaudible during broadcasting (§IV-A). We run NEC on a local laptop (§VI-C) to generate the shadow spectrogram, which is sent to a Keysight 33500B waveform generator, followed by an ultrasonic power amplifier [32] to amplify the inaudible shadow wave. Being transmitted through the air by a wideband dynamic ultrasonic speaker, Vifa [33], the shadow wave is superposed with the mixed audio at a COTS smartphone's

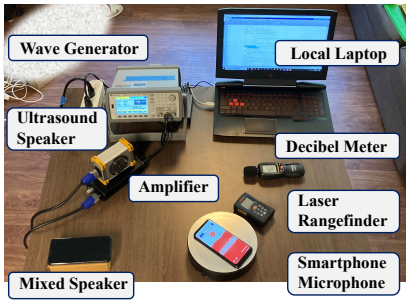


Fig. 10. Implementation and experimental settings.

microphone. We use a loudspeaker to play mixed audios, i.e., the "Mixed Speaker" in Figure 10 emulates a mixed conversation from Alice, Bob, and others. The target's voice will be effectively muted in the final recorded audio.

Dataset Compilation: Table I summarizes our testing dataset. First, we conduct the **System Benchmark** by testing the target speaker with the public datasets in controlled environments to verify whether our target speaker's voice can be hidden in the presence of real-world noises. Then, we deploy our system in the wild for a real attack scenario: the target volunteer wants to avoid being recorded while talking in public scenarios, but the COTS microphone can record others' voices normally.

- **Model Training:** Prior to the evaluation of NEC, we train a one-fits-all DNN model for all the defensive scenarios in public. The training dataset is constructed by mixing audios of two different speakers from LibriSpeech [34], and mixing target speaker audios with different noises from NOISEX-92 [35]. We provide the background audio that excludes the target speaker and train our model to hide the target speaker's voice, given the mixed audio and reference audios of the target speaker.
- **System Benchmark:** Using the public dataset LibriSpeech [34] as the corpus source, we first select **10 target speakers**, we collect 3 audios for each target speaker as their reference audio, and the rest audios of the speaker are treated as normal speech in a real scenario (e.g., Bob's speech). To measure the robustness of NEC, we simulate different environments with different types of noises. In order to cover different frequencies of noises (e.g., high-frequency speech and low-frequency ambient noises), the noises from 5 application scenarios are then mixed with the target speakers' voices, which results in 3,190 mixed audios in total. Then, we randomly mix the 10 target speakers' voices with the ones from the other 40 speakers, which generates 560 total instances for the joint conversation.
- **User Case Studies-1:** We further collect the user study dataset from **10 target volunteers**, covering 3 females and 7 males. All volunteers are required to speak 25 sentences, respectively. Analogous to the dataset for the benchmark, we select the reference and test audios randomly, then mix test audios with 4 sources of noise. In total, 160 mixed audios are produced. Then, we randomly mix the audios of 10 target volunteers with the ones from another 18 volunteers to derive the joint conversation dataset.

TABLE I
TESTING DATASET FOR BENCHMARK AND USER CASES

Scenario	Source	Freq.	Type	Instance
Joint[a]	LibriSpeech	0-8k	40/18	560/-
Conv.	/ Volunteers	0-8k	40/-	560/40
Babble[b]		0-4k	-	690 / 40
Factory[c]	NOISEX-92	0-2k	-	690/40
Vehicle[d]		0-500	-	690/40

[a]Two speakers talk jointly. [b]100 people whispering.
[c]a production hall. [d]a vehicle running at 120 km/h.

- **User Case Studies-2:** We conduct another user case study to justify the feasibility of NEC in the real world. As shown in Figure 12, Bob carries the NEC device to hide his sound in the wild. We ask Bob and Alice to speak normally, with volume at $77dB_{SPL}$ from our decibel meter placed at $5cm$ away from their lips. Then, we record the loudness, SONR, and the proportion of Bob's sound on Alice's recorder (a Moto Z4 phone) at different distances for different cases (with or without NEC).

Note that our testing dataset is disjoint from the training one and reference audios. Thus, the two trained models can be deployed directly with only three arbitrary reference audios from the new target speaker volunteers, avoiding the cumbersome deployment costs (e.g., model re-training and data re-collection) [36], [37].

Quantitative Metrics: To measure the quality of NEC, we consider four main metrics:

- **Source to Distortion Ratio (SDR)** [9], [31] measures the ratio of energy (in dB) between the energy of the target signal and the errors (induced by the interfering speakers and artifacts) in the mixed signal. It should be low for Bob's voice and high for Alice's voice.
- **Word Error Rate (WER)** is adopted broadly to evaluate the machine translation systems [38]. We compute the WER by employing Google's speech-to-text service to transform the acoustic signals into texts. NEC aims to enlarge the WER for the target speaker and minimize it for other speakers (e.g., Alice).
- **User Rating Score (URS)** is the rating for recordings, in which **10 reviewers** rank the raw mixed and recorded audios of NEC with score 1-5, along with Bob's clean voice as the ground truth. Specifically, score 5 denotes the best performance, in which reviewers cannot recognize any words of the target speaker (e.g., Bob).
- **Sound Noise Ratio (SONR)** is used to evaluate the proportion of Bob's sound in the recorded sound. We regard the mixed audio as useful sound and treat Bob's voice as noise. By computing the power ratio between the mixed audio and Bob's sound at different distances, we validate the efficacy.

VI. EVALUATION

In this section, we comprehensively evaluate NEC in different environments with different settings and devices.

A. Overall Performance

System Benchmark: We first evaluate NEC on the public dataset and provide SDR and WER across multiple scenarios

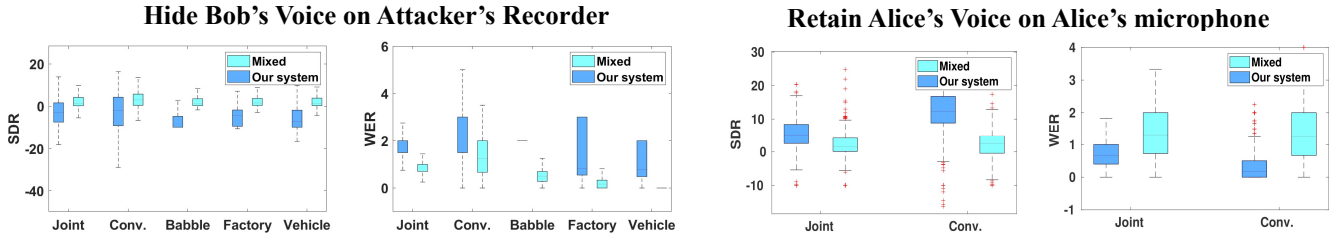


Fig. 11. Overall system performance of our system on three setups across multiple sources of noises.

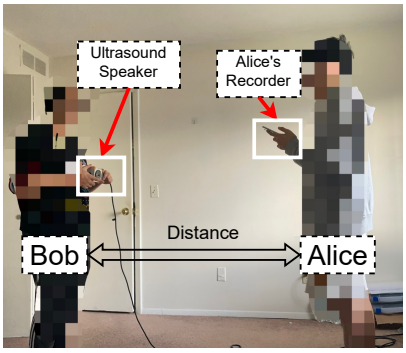


Fig. 12. Hiding Bob's voice from Alice's recording in a real-world scenario.

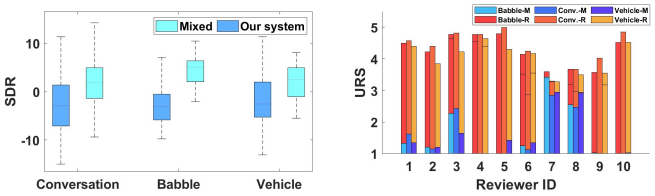


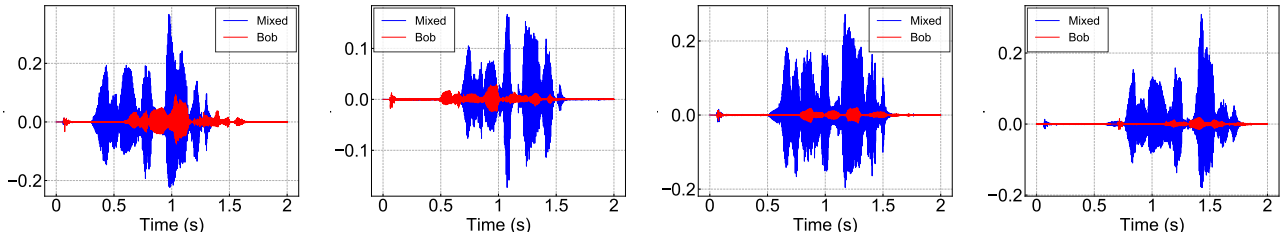
Fig. 13. User study results.

in Figure 11. When the target speaker, i.e., Bob, expects his voice to be hidden in the recordings, the recorded audios achieve a lower SDR and higher WER compared with the mixed audios. This shows that our shadow audios can hide Bob's voice reliably, making it unrecognizable by the Google service. Specifically, the median WER increases from 0.894 to 1.798, while the SDR reaches -4.918 dB from 0.997 dB. Note that the WER of the mixed audio is too high to be recognized by the Google service due to the background speeches from other people. Yet, it can still be recognized by humans. Conversely, NEC achieves a higher WER by hiding Bob's voice using the shadow wave, making it even unrecognizable for humans. We further verify its efficacy in the user studies below.

Also, we evaluate the effectiveness of NEC to retain others' voice (e.g., Alice) in Figure 11(right). We set the ground truth as Alice's clear voice, and calculate the SDR and WER for the recorded audio and the ground truth audio. The result shows that, compared to the mixed audio which contains Bob's voice, we can achieve higher SDR and lower WER for capturing Alice's sound when Bob deploys NEC. **User Case Study-1:** Figure 13 shows the performance of SDR and URS for hiding target volunteers' voices in the wild. We observe a consistent declination in SDR of the recorded audios compared with raw mixed ones. We can hardly recognize the target volunteer's voice in the recorded audios, as the median SDR reaches -4.374 dB, much lower than the SDR of mixed audios at 2.798 dB. To evaluate the recorded audios comprehensively, we ask 10 reviewers to score the recorded audios and the mixed ones with ranking scores from 1-5. We expect a higher score when fewer utterances of the target volunteer can be recognized. It shows that the average score of the recorded audios can reach 4.034 for different reviewers. All 10 reviewers give 4 for most of the recorded audios, while most scores of 1 are given to the mixed audios, except the reviewers 7 and 8. **User Case Study-2:** As depicted in Figure 12, in this user study, we evaluate how much of Bob's voice will be leaked to Alice's recorder with/without deploying NEC. We ask Bob and Alice to speak simultaneously, and also record Bob's sole speech with the same speech content. The mixed audio and Bob's individual speech audio are recorded by Alice's phone (Moto Z4), with varying distances between Alice and Bob (from 0.5 to 3 meters).

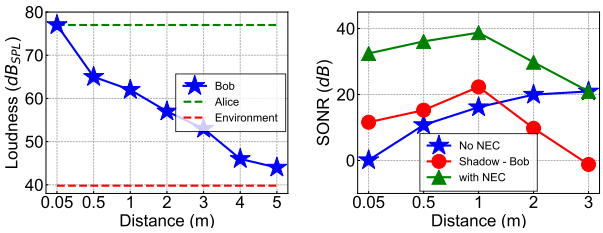
Figure 14 visualizes the waveforms of Bob's audio and the mixed audio. We can see that with the increasing distance, Bob's audio contributes less to the mixed one. We further record Bob's sound pressure level (SPL) at Alice's position and present the result in Figure 15(a). The result shows that the SPL of Bob's audio attenuates with the increasing distance, and its loudness reaches $43dB_{SPL}$ at the $5m$ distance (between Alice and Bob) with an environmental noise level of $39.8dB_{SPL}$. In comparison, the SPL of Alice's voice recorded by her own recorder remains at $77dB_{SPL}$. Given the large gap between the SPL of Alice and Bob's voices across different distances, and the attenuation of Bob's voice with the increasing distance, we can see that Bob only needs to cancel his voice over a short range (e.g., $2m$). Next, we further justify whether NEC can effectively overshadow Bob's sound across the distance.

Figure 15(b) presents the SONR results with/without NEC. When NEC is not deployed, the SONR between the recorded mixed audio and Bob's voice stays below 20dB, which implies that Bob's voice can be effectively captured by Alice's recorder. However, when Bob deploys NEC, even with a close distance ($< 2m$), Bob's voice can be mostly overshadowed, with SONR reaching 30dB. As mentioned above, the strength of Bob's voice signals drops significantly beyond $2m$. Therefore, although the recorded shadow audio strength also degrades dramatically beyond $2m$, the effectiveness of NEC within $2m$ makes it a viable solution for target voice cancellation.



(a) Distance = 0.5m (b) Distance = 1m (c) Distance = 2m (d) Distance = 3m

Fig. 14. Waveform of mixed audio and Bob's sole speech audio

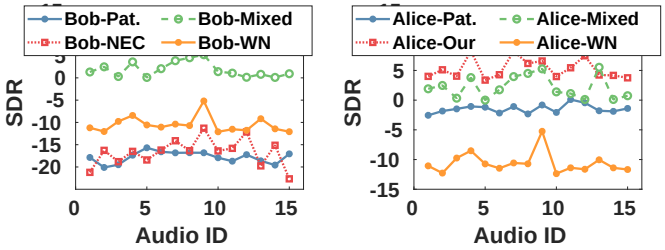


(a) Loudness vs. Distance (b) SONR vs. Distance

Fig. 15. Effectiveness of NEC across different distances

B. Comparison Study

Next, we perform a comparison experiment between NEC and two systems. The first one uses white noise to jam unauthorized recordings, which is commonly applied to commercial ultrasonic jammers. To simplify the jamming process, we manually add 10dB white noise over the recording sound to simulate this type of jamming system. Notice that the volume of white noise is usually determined by different jammers, we use 10dB based on our previous observation of the shadow sound volume on the same phone. The second one is a scrambling-based voice hiding system called Patronus [1], which can hide the target recordings by scrambling with specially designed white noises and recover the target recordings at an authorized device. Given the mixed (joint) audios (e.g., two volunteers, one of which is target), we reproduce the scrambling algorithms of Patronus to hide a speaker's voice.



(a) Hide Bob's voice. (b) Retain Alice's voice.

Fig. 16. Comparison study.

We first compare the performance of voice hiding by computing the SDR of the target voice. Figure 16(a) shows all of the three systems: NEC (Bob-NEC), White Noise (Bob-WN), and Patronus (Bob-Pat.) achieve a low SDR by effectively hiding the target voice in the mixed audio (Bob-Mixed). We find that, compared to NEC and Patronus, the white noise solution results in higher SDR, which means it retains more target voice than the other systems. Patronus and NEC can reduce the SDR of the mixed audios from 3 dB to nearly -20 dB. Therefore, the voice hiding performance of NEC is on par with that of the specially designed scrambling-based Patronus, and better than the white noise scrambling approach. Next, we evaluate the reception quality of Alice's voice in the presence of the three systems. As shown in Figure16(b), among the three systems, the White Noise approach cannot recover the disrupted voice, and therefore results in the lowest SDR for Alice's voice. For comparison, Patronus can recover a limited portion of scrambled sound by its recovery algorithm, and achieve low SDR for Alice's voice (i.e., -2.5 dB). The quality of Alice's voice after recovery is even lower than that of the raw mixed audios due to the influence of the scrambling noise. In comparison, NEC achieves a 5 dB gain compared with the mixed audios in recovering Alice's voice, since NEC carefully nullifies Bob's voice in the mixed audio. This experiment result demonstrates that NEC could selectively hide a target speaker's voice without interfering with other speakers. Surprisingly, NEC can even improve the reception quality of others' recording.

C. Running Time Analysis

To demonstrate the efficiency of our system, we measure the time consumption of each system module in Table II. Given 100 1s mixed audios, we evaluate the latency in two different hardware platforms: 1) desktop with a single NVIDIA 1080Ti GPU; 2) Raspberry Pi 4. The total processing time of the DNN module in NEC is around 1.51 ms, and the ultrasound modulation consumes 11.96 ms on average, well below the lasting period of the 1s chunks. In comparison, it takes 2.4× more time for VoiceFilter to process the same mixed audio. On the Raspberry Pi 4, the overhead of the selector is 293.7 ms, which is faster than 446.2 ms of VoiceFilter. The achieved latency (< 300) ms on the edge deployment using Pi 4 is less than the time offset tolerance of overshadowing, as discussed in §IV-C2, which further corroborates the feasibility of NEC.

TABLE II
TIME CONSUMPTION OF NEC WITH AN AUDIO SAMPLE LASTING 1S

Platform	System	Encoder	Selector	Broadcast
PC (1080Ti)	NEC	0.467ms	1.51ms	11.96ms
	VoiceFilter [9]	0.467ms	3.65ms	11.96ms
Rasp	NEC	12.7ms	293.7ms	11.96ms
	VoiceFilter [9]	12.7ms	446.2ms	11.96ms

D. Parameter Study

Diversity of Hardware Dependence: The variance of the non-linearity for the hardware (e.g., microphones, amplifiers,

filters) on smartphones can influence the optimal selection of the modulation parameters [28], which in turn impacts the performance of our system. Here, we evaluate our system using 7 different mobile devices listed in Table III. Specifically, the carrier frequency f_c is the dominant factor that affects the effectiveness of the non-linearity effect. All the tested smartphones have a range of acceptable frequency settings, and the best carrier frequency is listed in the brackets.

TABLE III
SMARTPHONES USED FOR TWO USER STUDIES.

Model	Brand	Carrier f_c (kHz)	Max Dis. (m)
Moto Z4	Motorola	24-28 (28.0)	3.2
iPhone 7 P	Apple	21-29 (27.8)	0.49
iPhone SE2	Apple	23-28 (25.2)	1.77
iPhone X	Apple	27-32 (25.3)	0.43
iPad Air 3	Apple	22-31 (28.0)	3.72
Mi 8 Lite	Xiaomi	24-32 (27.4)	1.65
Pocophone	Xiaomi	22-29 (26.3)	0.7
Galaxy S9	Samsung	25-31 (27.2)	3.64

Diversity of Effective Distance: Our system can be deployed with various maximum effective distances with different smartphone recorders, ranging from 49 cm to 3.72 m, as shown in Table III. The result also shows a great variance across recorders. We attribute this diversity to the difference in frequency response of these recorders, and the non-linearity of audio processing circuits.

Multiple Recorders: Since the performance of NEC can be affected by the variance of hardware, we investigate whether NEC system can be used to support multiple recorders simultaneously. To conduct this experiment, we use Moto Z4, Mi 8 Lite, POCOPHONE, and Galaxy S9 as recorders to eavesdrop on Bob's voice. With the collected recorded audios, we compute the SDR for recorded audios. For comparison, the SDR of mixed audio is also calculated to reveal the effect of NEC. We define that, if the SDR of recorded audio is less than the mixed audio, NEC is successfully performed. Our experiment result is presented in Table IV. For three different carrier center frequency settings, we played 20 crafted mixed audios and run NEC to superpose shadow audio to affect three recorders' recording. The column named **1+**, **2+**, **3** means at least **1**, **2**, or **3** devices are affected simultaneously by NEC. And the reported values such as 20/20 denote that all the 20 recorded audios are unable to recognize Bob's voice. This result provides the supportive evidence that NEC is capable of operating in public and affecting multiple recorders by carefully tuning the system parameters.

TABLE IV
NEC'S PERFORMANCE WITH MULTIPLE RECORDERS.

Number of Recorder		1+	2+	3
f_c (kHz)	26.3	20/20	9/20	4/20
	27.2	20/20	15/20	11/20
	27.4	20/20	14/20	8/20

VII. DISCUSSION & LIMITATION

Limitation of non-linear effect: The success of NEC relies on the imperfection of the receivers' (e.g., Alice's) microphone. However, when the non-linear effect is not present due to two reasons: 1) the great precision of Alice's microphone or 2) the improper modulation parameter settings, our selective voice protection will no longer be effective.

Limitation of protecting conversation: Although prior benchmark and user case experiments demonstrate that NEC can protect the target speaker's voice in the wild, it is a challenge to protect a conversation that involves multiple speakers while not disrupting other users (e.g., Alice). We failed to train a Selector model that is applicable to multiple target speakers with the current system architecture. In future work, we will figure out how to integrate the multiple speakers' embeddings and re-design the Selector model to avoid removing Alice's voice in the private conversation.

Directional of Ultrasonic Speaker: In our prototype shown in Figure 10, we assume the ultrasound speaker has the shadow audio ready before playing it. However, when we integrate the monitor, DNN models, and ultrasound speaker into one device and run it in a real-time manner, the shadow audio is dependent on the incoming mixed audio. In this case, one critical concern of NEC is whether the current mixed audio will be affected by the current shadow audio, therefore impacting the quality of future shadow audio. Fortunately, we can avoid it by putting the monitor and ultrasound speaker in *opposite direction*. We find that by exploiting the directional property of the ultrasound speaker, the shadow audio is barely sensed by the NEC's monitor as it produces limited amplitude in its back direction.

VIII. CONCLUSION

We present NEC, a lightweight AI-augmented voice protection system to protect the target speech without interfering with others' audio conversations. As an end-to-end processing system, NEC first actively emits specially designed ultrasound signals to a recorder. Due to the non-linearity effect, a shadow sound is generated and superposed onto the received mixed sound at the recorder, which effectively cancels the target speaker's voice in the recordings. To determine the frequency composition of the shadow sound, NEC leverages a tailored Deep Neural Network (DNN) to extract high-level speaker-specific but utterance-independent vocal features from the mixed sound. By imitating the overshadowing in the air, we superpose the shadow audio with the mixed audio in the training stage of the DNN model and deliver a one-fits-all model, which can be trained only once and deployed directly for new users. Our experimental evaluations demonstrate NEC's efficacy in a wide variety of real-world scenarios. The results show that NEC effectively disables the microphones from recording the target speaker's voice.

ACKNOWLEDGMENT

We would like to thank our shepherd Michael Paulitsch and anonymous reviewers for providing valuable feedback on our work. This work was supported in part by National Science Foundation grants CNS-1950171, CNS-1909177 and CCF-2007159.

REFERENCES

[1] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, "Patronus: preventing unauthorized speech recordings with support for selective unscrambling," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 245–257.

[2] Y.-C. Tung and K. G. Shin, "Exploiting sound masking for audio privacy in smartphones," in *Proceedings of ACM Asia Conference on Computer and Communications Security*, 2019.

[3] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Z. 0001, "Wearable Microphone Jamming." *CHI*, 2020.

[4] C. Li, M. Liu, and Z. Cao, "Wihf: Enable user identified gesture recognition with wifi," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 586–595.

[5] C. Li, Z. Cao, and Y. Liu, "Deep ai enabled ubiquitous wireless sensing: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.

[6] H. Guo, N. Zhang, S. Wu, and Q. Yang, "Deep learning driven wireless real-time human activity recognition," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[7] S. Zhu, J. Xu, H. Guo, Q. Liu, S. Wu, and H. Wang, "Indoor human activity recognition based on ambient radar with signal processing and machine learning," in *2018 IEEE international conference on communications (ICC)*. IEEE, 2018, pp. 1–6.

[8] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[9] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proceedings of Interspeech*, 2019.

[10] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *ArXiv*, 2020.

[11] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[12] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[13] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[14] A. Zhang, C. Wang, J. Paisley, Q. Wang, and Z. Zhu, "Fully supervised speaker diarization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[15] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, 2018.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[17] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, 2015.

[18] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proceedings of ACM CCS*, 2019.

[19] F. Winckel and T. Binkley, "Music, sound and sensation : a modern exposition," 1967.

[20] A. Löfqvist and B. Mandersson, "Long-time average spectrum of speech and voice analysis." *Folia phoniatrica*, 1987.

[21] A. Jongman, "Acoustics of american english speech: A dynamic approach," *Language and Speech*, 1995.

[22] S. Shen, N. Roy, J. Guan, H. Hassanieh, and R. R. Choudhury, "Mute: Bringing iot to noise cancellation," in *Proceedings of ACM SIGCOMM*, 2018.

[23] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[24] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomechanics and modeling in mechanobiology*, vol. 14, no. 4, pp. 719–733, 2015.

[25] M. Igras, B. Ziółko, and M. Ziółko, "Length of phonemes in a context of their positions in polish sentences," in *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. IEEE, 2013, pp. 59–64.

[26] S. Trauzettel-Klosinski and K. Dietz, "Standardized assessment of reading performance: The new international reading speed texts irest," *Investigative ophthalmology & visual science*, 2012.

[27] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of ACM MobiSys*, 2017.

[28] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided wave," in *Proceedings of Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[29] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of ACM CCS*, 2017.

[30] H. Yang, S. Bae, M. Son, H. Kim, S. M. Kim, and Y. Kim, "Hiding in plain signal: Physical signal overshadowing attack on LTE," in *Proceedings of USENIX Security Symposium (USENIX Security)*, 2019.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[32] A. Bioacoustics, "Portable ultrasonic power amplifier," in *http://www.avisoft.com/playback/power-amplifier/*, Retrieved by July 19th 2021.

[33] "Ultrasonic dynamic speaker vifa," in *http://www.avisoft.com/playback/vifa/*, Accessed on April 13, 2022.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[35] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, 1993.

[36] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wifi," in *Proceedings of ACM MobiSys*, 2019.

[37] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of ACM MobiCom*, 2018.

[38] Ye-Yi Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.