Demo: ClearBuds - Wireless Binaural Earbuds for Learning-Based Speech Enhancement

Ishan Chatterjee*, Maruchi Kim*, Vivek Jayaram*
Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, Steven M. Seitz
{ichat,mkimhj,vjayaram,gshyam,kemelmi,shwetak,seitz}@cs.washington.edu
*Co-primary student authors
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA

ABSTRACT

We present ClearBuds, the first end-to-end hardware and software system that utilizes a neural network to enhance speech streamed from two wireless earbuds. Real-time speech enhancement for wireless earbuds requires high-quality sound separation and background cancellation, operating in real-time and on a mobile phone. ClearBuds bridges state-of-the-art deep learning for blind audio source separation and in-ear mobile systems by making two key technical contributions: 1) a new wireless earbud design capable of operating as a synchronized, binaural microphone array, and 2) a lightweight dual-channel speech enhancement neural network that runs on a mobile device. Our demo will allow MobiSys attendees wear our earbuds, and experience noise suppression as they talk in a noisy environment. Companion video can be accessed using the link below:

https://youtu.be/vzHHZnlfTe8

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Audio source separation, earable computing, noise cancellation, cascaded neural networks, audio and speech processing, real-time mobile deep learning, binaural earbuds

ACM Reference Format:
Ishan Chatterjee*, Maruchi Kim*, Vivek Jayaram* and Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, Steven M. Seitz. 2022. Demo: ClearBuds - Wireless Binaural Earbuds for Learning-Based Speech Enhancement. In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25–July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3498361.3538654

MobiSys '22, June 25–July 1, 2022, Portland, OR, USA

ACM ISBN 978-1-4503-9185-6/22/06.
https://doi.org/10.1145/3498361.3538654

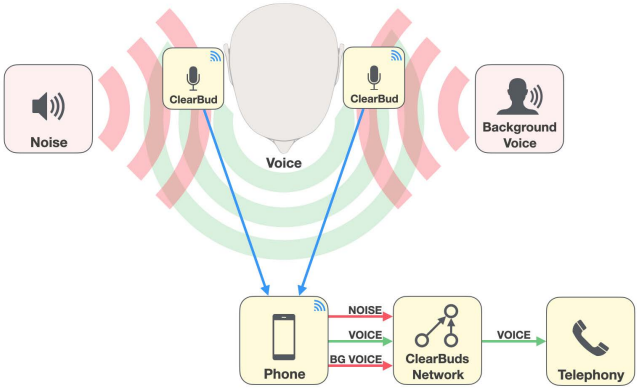

Figure 1: ClearBuds Application. Our goal is to isolate a user's voice from background noise by performing source separation using a pair of wireless, synchronized earbuds.

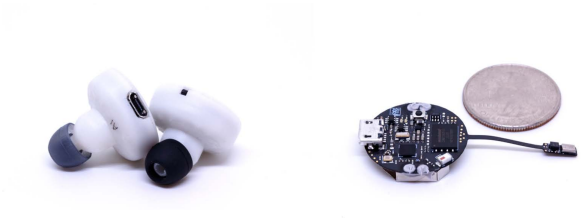

Figure 2: ClearBuds hardware. Images show ClearBuds inside 3D-printed enclosure and when placed beside a quarter.

1 INTRODUCTION

With the rapid proliferation of wireless earbuds, more people than ever are taking calls on-the-go. While these systems offer unprecedented convenience, many kinds of environmental noise (e.g. street sounds, home appliances, people talking) can interfere and reduce the quality of calls. We therefore seek to enhance the speaker's voice, and suppress background sounds using speech captured across the two earbuds.

In this demo, we present the first system that uses neural networks to achieve real-time speech enhancement from binaural wireless earbuds. Our key insight is to treat wireless earbuds as a binaural microphone array, and exploit the specific geometry – two well-separated microphones behind a proximal source – to devise a specialized neural network for high quality speaker separation. In

(a) (b) (c) (d)

Figure 3: Scenarios where ClearBuds can be utilized on-the-go (crowded cafe, busy intersection, outdoor plaza, classroom).

contrast to using multiple microphones on the same earbud to perform beamforming, as is common in Apple AirPods [1] and other hearing aids, we use microphones across the left and right earbuds, increasing the distance between the two microphones and thus the spatial resolution. Achieving this goal is challenging for three key reasons. **First**, today's earbuds are not capable of operating in this manner; AirPods and similar devices upload microphone output from only a single earbud at a time. To achieve binaural speaker separation, we need to design and build novel earbud hardware that can synchronously transmit audio data from both the earbuds. **Second**, binaural speech enhancement networks are not lightweight and have not been demonstrated with wireless earbuds. Reducing the network size naively often leads to unpleasant artifacts. Thus, we also need to optimize the neural networks to run in real-time on smart devices that have a limited computational capability compared to cloud GPUs. **Finally**, we need to meet the end-to-end latency requirements for telephony applications and ensure that the resulting audio output has a high quality from a user experience perspective. A full description of our system is described in [2].

In our demo, we will present an end-to-end system which is built off of a pair of custom wireless earbuds and a new hybrid neural network architecture. Our system is capable of (1) source separation for the intended speaker in noisy environments, (2) attenuation and/or elimination of both background noises and external human voices, and (3) real-time, on-device processing on a commodity mobile phone paired to the two earbuds. To achieve this system, we make two contributions spanning earable hardware and neural networks.

2 CLEARBUDS HARDWARE

We designed a binaural wireless earbud system (Fig. 2) capable of streaming two time-synchronized microphone audio streams to a mobile device. The hardware schematic and layout for ClearBuds was designed using the open source eCAD tool KiCad. A 2-layer flexible printed circuit was fabricated and assembled by PCBWay ($2K for 50 units). The 3D printed enclosures were designed using AutoDesk Fusion 360 and printed with a Phrozen Sonic Mini using a liquid resin fabrication process. The MEMS microphone sits behind the lid on the earbud's outer surface, and a single button on the enclosure provides access to turn on and off the earbuds. Our custom hardware design contains a pulse-density modulated (PDM) microphone (Invensense ICS-41350) and a Bluetooth Low Energy (BLE) microcontroller (Nordic nRF52840). Our binaural wireless earbuds can stream audio to a phone with a synchronization error less than 64μs and operate continuously on a coin cell battery for 40 hours.

3 CLEARBUDS NETWORK

Our network needs to perform in real-time on a mobile device with minimal latency. A main challenge is that the processing device has a much lower compute capacity, especially compared to cloud GPUs. Additionally, the network should separate non-speech noises as well as unwanted speech. To do this, it must learn spatial cues and human voice characteristics. Finally, the resulting output should maximize the quality from a human experience perspective while minimizing any artifacts the network might introduce. We introduce a light-weight neural network that utilizes binaural input from wearable earbuds to isolate the target speaker.

To achieve real-time operation, we start with the Conv-TasNet source separation network [3] and redesign the network to achieve a 90% re-use of the computed network activations from the previous time step for each new audio segment. While these optimizations make this network real-time, they also introduce artifacts in the audio output. To address this, we combine our mobile temporal model with a real-time spectrogram-based frequency masking neural network. We show that by combining the two networks and creating a light-weight cascaded network, we can reduce artifacts and improve the audio quality further.

Our network operates on packets of 22.4ms. We evaluate the runtime on several different mobile devices, and find that an iPhone 12 Pro processes these packets in 21.4ms. This means that it can keep up with real-time with a total processing delay of < 50ms.

4 DEMO SETUP

Our demo setup at MobiSys will consist of a pair of ClearBuds and an iPhone. We will provide the earbuds to participants who can wear them and participants will speak into them in the presence of other noises that we might play from a laptop or that may be present in the environment. The binaural audio will be streamed to the iPhone where participants can visualize the raw data, as well as the noise suppressed output of our network in real-time. Once a participant is done speaking, they will be able to hear the clean speech output of our system from the iPhone using a pair of headphones.

REFERENCES

[1] Apple airpods. https://www.apple.com/airpods/.
[2] Jayaram et al Chatterjee, Kim. Clearbuds wireless binaural earbuds for learning-based speech enhancement. 2022.
[3] 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.