Poster: VoCopilot: Enabling Voice-Activated Tracking for Everyday Interactions

Sheen An Goh
National University of Singapore
e0926997@u.nus.edu

Ambuj Varshney
National University of Singapore
ambujv@nus.edu.sg

ABSTRACT

Voice plays a crucial role in our daily lives, enabling communication, conveying emotions, and indicating our health. As a result, tracking vocal interactions can provide valuable insights into various aspects of our lives. This poster presents our preliminary work for a novel voice tracker (VoCopilot) that effectively tracks various vocal interactions. For example, the VoCopilot tracker can help document meetings and generate notes, even when participants speak different languages. Additionally, it can serve as a life-logger, monitoring daily conversations and extracting key points to summarize their content. Central to VoCopilot's design is an energy-efficient, co-developed acoustic hardware and firmware combined with a comprehensive integration of VoCopilot tracker with advanced machine learning systems. This harmonious integration ensures precise voice transcription, summarization, and analysis. We present our early thoughts on VoCopilot hardware design and share early results of utilizing Whisper for efficient multilingual transcribing. We acknowledge VoCopilot may raise privacy issues; therefore, we provide early thoughts to address these concerns.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Speech-to-Text, Wearable, AI, Acoustic, Embedded Systems

ACM Reference Format:
Sheen An Goh and Ambuj Varshney. 2023. Poster: VoCopilot: Enabling Voice-Activated Tracking for Everyday Interactions. In *The 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23), June 18–22, 2023, Helsinki, Finland.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3581791.3597375

1 INTRODUCTION

Every day, we use our voices to communicate with each other. Vocal patterns can also provide insight into our emotional states and overall well-being [1, 2, 4]. We can therefore gain valuable information about our lives by tracking our vocal interactions.

Despite the potential benefits of continuous vocal interaction tracking, current smartphones and wearable devices do not actively track our conversations. Although capable of monitoring vocal interactions, smartphones are power-intensive for this task, thus reducing battery life. Voice recorders lack integration with machine learning systems. There have been efforts in using earphones as sensing platforms [3]. However, they are designed for communication, information consumption, and entertainment rather than vocal tracking. Consequently, they offer limited battery life, lack capacity for onboard storage of vocal interactions, and lack integration with advanced machine learning systems.

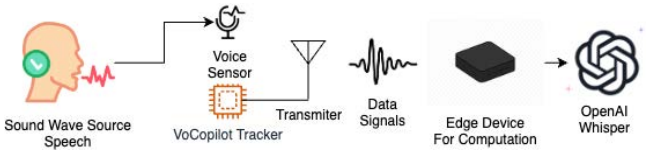

Figure 1: *System overview.* Voice is received by the microphone and pre-processed on the tracker before it is wirelessly communicated to the edge device. The edge device runs an advanced machine learning system, OpenAI Whisper, and performs transcribing. Finally, the transcript can be analysed and summarised using a large language model.

To address this unexplored application space, we introduce our preliminary work on developing a novel class of wearable devices specifically designed to monitor our vocal interactions continuously. These devices could record multiple days of vocal interactions before needing battery replacement or recharging, or they might even function without batteries by harvesting ambient energy. Next, these conversations could be transcribed and analysed using advanced machine learning systems. Thus, through this study, we envision the development of a wearable device to effortlessly record daily conversations and offer valuable insight into your life.

We anticipate challenges in developing this system that needs to be addressed: *Firstly*, from the device perspective, it should be able to record multiple days' worth of conversations. Furthermore, supporting several days of operation on standard batteries should be possible. *Secondly*, the challenge involves efficiently transcribing audio into text, which can then be used for generating summaries and extracting additional insights. *Finally*, our vocal interactions can contain very sensitive information; hence, we need to ensure the security and privacy of the collected information.

This preliminary study provides suggestions and considerations for designing an energy-efficient tracking device. However, our primary focus is to address the second challenge: transcribing audio and extracting insights from the transcriptions. We utilize the OpenAI Whisper system for audio transcription and assess its performance. Our initial findings, presented in this work, showcase the capability of OpenAI Whisper to transcribe text on an edge device within a reasonable processing time.

MobiSys '23, June 18–22, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0110-8/23/06.
https://doi.org/10.1145/3581791.3597375

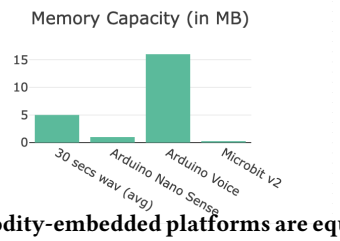

Figure 2: Commodity-embedded platforms are equipped with limited memory. VoCopilot has to store locally vocal interactions before transmitting them to an edge device. This can exceed the available memory capacity on these platforms.

2 DESIGN

Figure 1 demonstrates a high-level overview of VoCopilot. It operates as follows: The tracker monitors vocal conversations and stores all or a portion of them on its onboard storage. Subsequently, these conversations are transmitted to an edge device. Next, the edge device employs a machine learning system, such as OpenAI Whisper, to transcribe these conversations. Finally, the transcribed conversations are processed using a large language model (LLM) system. As a result, our system is divided into three main components.

VoCopilot Tracker. Trackers are wireless embedded devices with microphones. It records vocal interactions, performs pre-processing, and performs computations to ensure that the size of the audio recordings remains manageable for onboard storage. As a preliminary step, we have explored whether commodity platforms are suitable for implementing the tracker device. Figure 2 shows the available memory on some commonly used embedded platforms, which we found limited to a few to tens of megabytes. This presents a challenge for our system, as storing hours of vocal interactions requires memory capacities orders of magnitude higher.

As a result, our first insight is that custom hardware development is necessary for implementing the tracker device. In this respect, we are considering using Syntiant NDP120 to enable selective recording of conversations to maintain manageable storage and power consumption. However, as we develop the custom hardware, we utilise a powerful Raspberry Pi Zero W device with external memory (MicroSD) and microphone as the VoCopilot tracker.

Edge Device. Due to embedded devices' limited computational, storage, and networking capabilities, we employ an edge device for the computationally intensive transcription process. We implemented the edge device using a Mac Mini M2 (16GB), chosen for its superior processing capabilities. For audio transcription, we utilized OpenAI Whisper, motivated by several reasons: Firstly, Whisper is open-source and supported by an active community of technologists driving its development. Secondly, Whisper runs locally, ensuring secure transcription and translation without cloud service transmission, thus preserving privacy.

Large Language Model. We intend to utilize a LLM to extract insights from the transcribed conversations. However, sending private information to a cloud-based LLM may not be feasible. Therefore, we plan to employ a locally running LLM on the edge device.

Privacy. Privacy is a cornerstone of our design to protect the system's users. *Firstly*, our system architecture ensures that the data is only consumed within the locally running AI models [5] without sending the data to the cloud. *Secondly*, we propose to protect the data via both encryption-at-rest and encryption-in-transit. *Lastly*, to prevent recording voices without consent, we may consider using NDP120 to record conversations upon selective keywords. [6]

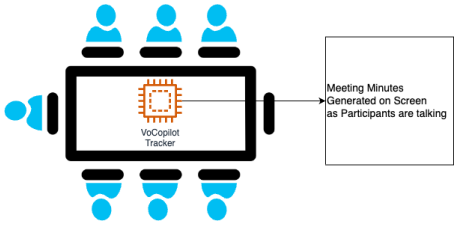

Figure 3: VoCopilot can transcribe meetings, generate meeting notes, and give a summary of the discussed points

3 PRELIMINARY RESULTS

We conducted early experiments to evaluate transcribing ability of our system. We used datasets of different time durations and two languages. Different models of OpenAI Whisper: "base", "base.en", "medium", and "medium.en" is evaluated to demonstrate (table below) that transcribing can be achieved in a reasonable time duration.

Models	base	medium
English 30s	2.20s	16.20s
English 3min	10.08s	79.18s
English 22min	69.60s	537.57s
Mandarin 30s	2.19s	19.52s
Mandarin 3min	16.92s	154.28s
Mandarin 22min	113.17s	858.54s

We also observed Whisper is able to generate highly accurate transcripts across different runs(above 90%) which confirms results of [5] that Whisper can produce state-of-the-art WER of voice datasets.

4 CONCLUSION

Our preliminary work introduces the system architecture for a novel class of wearable devices designed to track vocal interactions. We envision various use cases for our system, such as enabling the transcription of university lectures or meetings with automatic distribution to all participants, as shown in Figure 3. Another potential application is for journaling. It benefits mental health and alleviates stress. Despite these benefits, many perceive it as time-consuming. By offering a speech-to-text journaling service, we aim to lower the barrier to entry, enabling users to effortlessly journal on the go by tracking their voice interactions throughout the day.

Acknowledgement. This work is partially funded through a startup grant from the National University of Singapore (ODPRT).

REFERENCES

[1] Kayla-Jade et al. Butkow. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *IEEE PerCom 2023*.
[2] Jing et al. Han. 2022. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine* 5, 1 (2022).
[3] Fahim et al. Kawsar. 2018. ESense: Open Earable Platform for Human Sensing. In *ACM SenSys 2018*.
[4] Daryush D. et al. Mehta. 2012. Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform. *IEEE Trans. on Bio. Eng.* (2012).
[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).
[6] Lu Zeng, Sree Hari Krishnan Parthasarathi, Yuzong Liu, Alex Escott, Santosh Cheekatmalla, Nikko Strom, and Shiv Vitaladevuni. 2022. Sub 8-Bit Quantization of Streaming Keyword Spotting Models for Embedded Chipsets. In *TSD (Lecture Notes in Computer Science, Vol. 13502)*. Springer, 364–376.