In-Ear Aid for Speech Separation

Andrea Ferlini
University of Cambridge
af679@cam.ac.uk

Cecilia Mascolo
University of Cambridge
cm542@cam.ac.uk

1. INTRODUCTION

People struggle to identify and isolate the source of a sound. This is particularly true in social situations. Crowded places, where multiple conversations, involving several speakers, happen at the same time, result in the phenomenon known as *Cocktail Party Problem* [2]. While this is already a though hearing task in general, it becomes particularly hard for people who suffer from deafness or hearing losses.

The importance of addressing the cocktail party problem grows with the portion of the world's population affected by hearing disabilities, which is steadily increasing. As in 2019, more than 466 million people, among whom around 34 children, are suffering from hearing impairments[1]. Moreover, the World Health Organization forecasts over 900 million with hearing disabilities by 2050[2].

Our work aims to shed light on this problem and the system challenges related to it. Moreover, it forms the basis to the design, and the requirements of a wearable specifically thought to address the cocktail party problem and the challenges that come with it.

2. PROBLEM STATEMENT

From a system perspective, there are multiple challenges that must be tackled to produce a device to enhance hearing capabilities. The presence of background noise and multiple sources increase the difficulty of the task. The target voice has to be separated not only from the noise but also from other voices. Hence, the need for correctly labeling and identifying the source of speech. Past works has tackled these challenges without real-time, device-constrained requirements [1][4][5]. There has been no research trying to solve the cocktail party problem with real-time constraints, on power and resource constrained wearables. The requirements for a device able to do this are several:

- it has to isolate the voice of the speaker the user wants to interact with, removing both background noise and other speakers' voice;
- it has to support real-time learning to efficiently perform speech separation and segregation;
- it has to perform most processing on-device, to preserve users' privacy;
- its battery has to last a sufficient amount of time, in order to provide real aid to the users;
- it has to be situation aware, e.g. the user should be able to "lock" and "unlock" its focus on the conversation partner;

3. PRELIMINARY DESIGN & DISCUSSION

In this work, we present the preliminary design of a new earable device that satisfies the list of requirements in Section 2. Using an earable will allow us to better capture audio and other features, being able to directly excite the user hear with our output. Our choice is supported by works like the one of Min et al. [3], which shows the enormous, and mostly unexploited, sensing potential of earable devices. To successfully separate the targeted voice from other voices and the noise, we leverage Multimodal Learning with binary weights cope with the lack of resources available. Multimodal Learning enables the leveraging of multiple inputs, such as audio, head movements, and eye movements, to better amplify and isolate the desired signal.

4. REFERENCES

[1] A. Ephrat and et al. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation.
[2] J. H. Mcdermott. The cocktail party problem.
[3] C. Min and at al. Exploring audio and kinetic sensing on earable devices.
[4] A. Owens and at al. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features.
[5] H. Zhao and et al. The Sound of Pixels.

[1]Hearing disabilities are considered so when referring to hearing loss greater than 40dB in adult subjects and 30dB in children.
[2]https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss