Analysis of Speech Datasets for Communication Scenarios for Hearing Aid Users

Robert Sutherland, Stefan Goetze, and Jon Barker

Department of Computer Science, The University of Sheffield, UK

1. Abstract

Despite recent advances in technology, modern hearing aids still struggle with a wide variety of real-world conversational situations. One major gap in the domain of hearing aids is the lack of publicly available, realistic datasets for typical, day-to-day conversation scenarios. This work aims to study some of the existing speech datasets, with a view to proposing an idea for scenarios to be recorded in future datasets specifically for hearing aids.

One key point to consider with these datasets is how well they capture natural behaviour. While simulated datasets, such as the Clarity Enhancement Challenge 2 (CEC2) dataset, are able to realistically simulate the acoustic environment, the artificial mixture of pre-recorded speech does not fully reflect natural behaviour present in human conversations.

An interesting behaviour not covered by any speech datasets is the head movement of the listener. Recent research has shown that there is a link between head motion and turn-taking during a conversation [1]. In the dataset analysed in [1], three participants sat in an equilateral triangle and held a conversation on various topics, and the data recorded for each participant included their x and y coordinates in the room, the *yaw* of their heads, and a binary classification of whether they were speaking, all sampled at 100 Hz.

Further analysis is shown as Figure 1; for each sample, the analysis detects who the active speaker is (ignoring any samples where nobody is speaking), calculates the angle between the non-speaking participants and the active speaker and then finds the difference between this angle and the respective yaw measurements. This is then displayed in a histogram. From Figure 1 it is clear to see that in general, listeners will face towards the active speaker, though not directly at them.

This information has potential applications for detecting which speech sources are relevant to the listener. While there are sophisticated source separation algorithms, for hearing aids it is important to know which sources the listener wants to attend to, and which sources are interfering. Figure 1 gives a clear indication that the head movement could be a valuable feature to use in determining which sources are useful to the listener, but as of yet there is no dataset to evaluate this.

While this dataset on head orientations was able to capture realistic behaviour of individuals, it only records a very specific arrangement of participants so it is not sufficient to generalise to other scenarios. The CHiME-5 dataset [2] may have more success in this area. It recorded a dinner party between four familiar participants in three phases: preparing the meal in the kitchen area, eating the meal in the dining area, and an after-dinner period in a living room area. This indicates there will be a more diverse range of participant arrangements, leading to a more realistic representation of everyday scenarios. The COSINE corpus [3] provides even more variety, with conversations containing from two to seven participants, recorded in real, noisy environments.

This work highlights a gap in the existing speech datasets, and aims to devise new scenarios which may be useful in the recording of future datasets for hearing aids. Key areas to consider include the presence of competing speech, the variety of participant arrangements, and the recording of head movements.

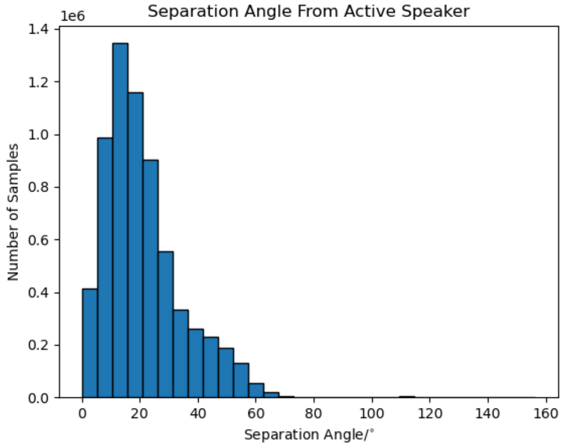

Figure 1: *A histogram displaying the separation angle between a listener's head orientation and the direction of arrival of the speech signal, from the dataset analysed in [1].*

2. References

[1] L. V. Hadley and J. F. Culling, "Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions," *Frontiers in Psychology*, vol. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1061582

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," Mar. 2018, arXiv:1803.10609 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1803.10609

[3] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - A corpus of multi-party COnversational Speech In Noisy Environments," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4153–4156, iSSN: 2379-190X.