Optimizing Video Call Experiences Through Direction of Voice Based Noise Filtration

Abiramy Kuganesan
akuganes@cs.ubc.ca

Emmanuel Sales
emsal@cs.ubc.ca

1 Motivation

As open concept offices and integrated work-home contexts become more of a norm, noise filtration during video calls is becoming increasingly important. Multiple verbal interactions may take place during a video call, some of which are not directed towards the call participants. The intention of this project is to be able to filter the audio input that is processed from an environment to the video call recipients depending on the direction of voice. More specifically, the direction of voice is defined as the direction along which a voice is projected. This is different from direction of arrival which aims to compute the direction from which the voice originated [1]. Figure 1 from [1] intends to illustrate this distinction.

2 Data Acquisition Strategy

The primary source of data for this project will be the Direction of Voice dataset released by the authors of Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems. This dataset made available by the FIGLAB contains utterances from 10 participants across a bin of 8 speaking angles (0, +45, -45, +90, -90, +135, -135 and 180°) at varying distances (1m, 3m and 5m) during 2 sessions in 2 different rooms. The test apparatus used to acquire this data is a ReSpeaker USB 4-channel microphone.

3 Proposed Approach

For the first stage of this project, we aim to recreate the method for detecting direction of voice from [1]. Their approach involved featurizing the signals by taking a measure known as the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT [4]). To recreate this functionality, we will utilize their dataset, perform computation of the GCC-PHAT as a featurization procedure, and classify the direction-of-voice based on those features using a similar decision-tree based classifier.

For the second stage, we aim to explore the case where there are multiple simultaneous signals with different directions of voice within the environment. In order to do this, we will augment the dataset by creating audio files that consist of two of the audio files from the original sample overlaid together. For this more complex task of speaker localization, the GCC-PHAT can also be used to learn other aspects of the signal, such as direction of arrival, angle of arrival, time difference of arrival, and frequency difference of arrival. We will take the GCC-PHAT of the mixed signals and use these features, combined with our model from the first stage and potentially other dependent features, to create a model that can identify both directions based on this compound training data. Localization of multiple sources of audio has been explored in works such as as [5], [3], [6].

Further exploration will involve attempting to devise methods for (a) identifying which voice direction contains the more dominant or otherwise "most relevant" signal, and (b) reconstructing a filtered version of the signal where the signal components corresponding to the less dominant voice direction is expunged. To achieve this, we surmise that a model would need to learn a mapping between the direction of voice and Fourier and/or cepstrum coefficients of the input signal.

4 Evaluation Strategy

To evaluate the accuracy of the general direction of voice estimation task, the evaluation strategy described in [Ahuja, Kong, Goel, and Harrison] will be used. The second session will be utilized as the test set with the F1 score being computed for the binary "Facing" or "Not Facing" classification task.

For the overlapping voice segmentation task, the intention is to augment a subset of this dataset that will be set aside and carefully curated for testing to overlay utterances projected along varying speaking angles. Combinations of utterances will be mixed to produce the input audio. Session, test room and participant variance will be accounted for in curating the test and train splits to ensure that the splits are as independent as possible and that difference in the distributions of the data is maximized for minimal bias. After processing, the similarity between the original utterance along the forward facing directed angle and the recovered audio will be compared.

One means of performing this comparison would be to generate the fingerprints of both audio segments using the Chromaprint algorithm and then compute a condensed similarity metric between the two fingerprints by computing the correlation between the two fingerprint vectors. The chromaprint algorithm leverages the log magnitude spectrograms of the two audio files, converts the frequencies to notes using chroma bands and then computes a similarity matrix by using the correlation between the two vectors [2]. The intention is to maximize the correlation between the two vectors and a match would be deemed as a similarity score above some threshold. The underlying assumption is that there is no time lag between the two signals being compared. Taking the cross-correlation of these vectors instead may accommodate for such lag.

References

[1] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1121–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[2] M. A. Bartsch and G. H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[3] Dirk Bechler and Kristian Kroschel. Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array. *Proc. of IWAENC, 2003*, 01 2003.

[4] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[5] B. Kwon, Y. Park, and Y. Park. Multiple sound sources localization using the spatially mapped gcc functions. In *2009 ICCAS-SICE*, pages 1773–1776, 2009.

[6] Despoina Pavlidi, Matthieu Puigt, Anthony Griffin, and Athanasios Mouchtaris. Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 03 2012.