Optimizing Video Call Experiences Through Direction of Voice Based Noise Filtration

ABIRAMY KUGANESAN* and EMMANUEL SALES*, The University of British Columbia, Canada

Noise filtration has become a prominent application area as reliance on video communication to facilitate interactions with coworkers, peers, and family members increases. The transition towards remote living, working, and studying experiences place a higher demand on video conferencing services now more than ever. Direction of voice (DoV) estimation is a recent development in audio signal processing which can be leveraged to enhance video call experiences such that they more closely resemble natural audio perception. Using some of the methods from recent research in this area, we provide a filtering-based application of DoV estimation to facilitate the transmission of relevant conversation to video call recipients. An extra-tress classifier determines which segments of audio are targeted towards the recording device. A filter blocks audio segments that are projected along irrelevant axes from being transmitted to the wrong audience. Empirically, the proposed work validates this filtering functionality using existing datasets and recordings that were acquired to emulate video call and streaming experiences.

CCS Concepts: • **Human-centered computing** → **Sound-based input / output**; **Auditory feedback**.

Additional Key Words and Phrases: directionally selective auditory attention, direction of voice, noise filtration, conversational cues

ACM Reference Format:
Abiramy Kuganesan and Emmanuel Sales. 2020. Optimizing Video Call Experiences Through Direction of Voice Based Noise Filtration. 1, 1 (December 2020), 9 pages.

1 INTRODUCTION

As remote working environments and integrated work-home contexts become more of a norm, noise filtration during virtual interactions is becoming increasingly important. While numerous verbal interactions may occur during a video call, some of these interactions may not be directed toward the call participants. The ability to discern the directional nature of sound is a fundamental human characteristic that facilitates targeted human to human interactions. This conversational cue provides an understanding of who or what a person wishes to address as well as whether or not conversation is directed toward an individual. Many times, the brain inherently suppresses irrelevant speech as determined by spatial and directional perception of these auditory signals [1]. Although spatially selective auditory attention can be accomplished by computing systems through direction of arrival and speaker localization [2][3], directionally selective auditory attention has typically been imitated through inferred gaze direction [4].

In this work, we explore a method for filtering speech input depending on the direction of voice. This is illustrated through a technique for filtering the audio that is processed from an environment such that only conversation directed toward video call participants is transmitted. This work is an extension of the methods presented in *Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems* presented by Ahuja, et al. at UIST '20 [5]. The significance of this work is the ability to show how computing systems can imitate the directionally selective auditory attention that is intrinsically performed by humans.

*Both authors contributed equally to this research.

Authors' address: Abiramy Kuganesan, akuganes@cs.ubc.ca; Emmanuel Sales, emsal@cs.ubc.ca, The University of British Columbia, 2329 West Mall, Vancouver, British Columbia, Canada, V6T 1Z4.

2 RELEVANT WORK

2.1 Direction of Arrival Estimation

In order to understand the significance of this work, it is important to comprehend the distinction between direction of arrival (DoA) and direction of voice (DoV). Direction of arrival estimation aims to compute the direction from which the voice originated. In contrast, direction of voice is the axis along which the voice is projected by the speech emitting source [5]. Figure 1 aims to illustrate this difference.

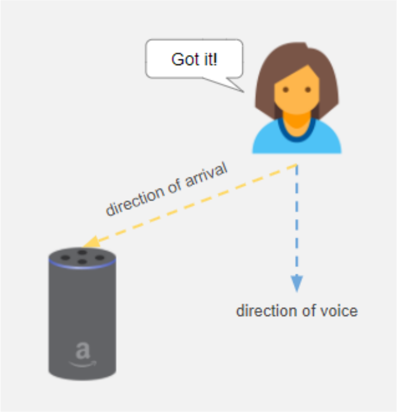

Fig. 1. Distinction between direction of voice (blue) and direction of arrival (yellow)

DoA estimation has a wide range of applications ranging from enabling social robots to provide targeted responses [6] to focusing rotating cameras on the active speaker during meetings [7]. A relevant application to our work is the use of DoA estimation for speech enhancement in noisy environments. One particular technique uses the phase difference between dual-microphone signals to obtain an estimate of the signal-to-noise ratio of the targeted speech to non-targeted speech. The estimate for the DoA based signal-to-noise ratio is fed into a Wiener filter to product a spectral-gain attenuator [8]. One limitation of this work is that it models the target speech with the assumption that it contains no room reverberation effects. Depending on the material composition of components within the room, humidity and general room structure, this may not always be the case.

2.2 Room Reverberation Effects

A method to mitigate reverberation effects in multi-microphone settings is through the use of the Generalized Cross-Correlation function with Phase Transform weighting (GCC-PHAT). When it was first introduced, the intention of the generalized cross-correlation (GCC) was to determine the time delay between signals which were received by two spatially separated sensors in the presence of uncorrelated noise [9]. Subsequent developments illustrated that the GCC-PHAT was robust to room reverberation effects when compared to other weighting alternatives of the GCC [10] and could produce reliable DoA angle estimates [11].

3 METHOD

The approach taken to solve this problem stems largely from the work presented in *Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems*.

3.1 Dataset

The primary source of data for this work is the Direction of Voice dataset [5]. This dataset contains approximately 11.5k audio recordings from 10 participants. The participants recite "Hey assistant" and "The quick brown fox jumped over the lazy sheep" along 8 different angles with 45° increments from 0° to 360°. Each of these statements are recorded from 9 different polar positions in 2 room setups with varying reverberation properties. Figure 2 illustrates the experimental layout through which these recordings were collected. These recordings were recorded by a Seeed ReSpeaker USB 4-channel microphone which was also purchased for experimentation. Additional data for testing and demonstration purposes was acquired using this microphone. The raw data from the microphone is fed through a featurization task, classification task and finally a filtration process.

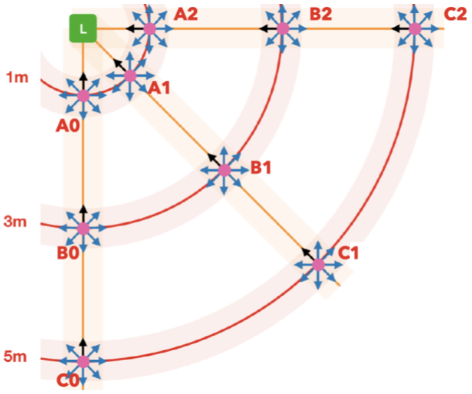

Fig. 2. Speaker positions and voice projection axes for DoV dataset acquisition. The green square denotes the microphone array's position. The participants recited each of the two phrases from the 9 positions outlined by A0 through C2 along the 8 directions denoted by the blue and black arrows. The black arrow is the 0° reference direction for a voice that is directed toward the microphone array [5].

3.2 Featurization

Each raw input data sample contains 4 audio recordings from each of the 4 channels of the recording apparatus. The GCC-PHAT cross-correlations between pairs of the 4 channels are taken. Due to the distance between orthogonal microphones, a maximum theoretical delay between the recordings can be established. The 6 raw outputs of the resulting GCC-PHAT correlations are cropped by this maximum theoretical delay of 0.236ms around the maximum valued peak. The standard deviation, mean and range of the cropped GCC-PHAT outputs, the value of the max peak of each correlation, the delay between channels, and area under of the curve of each correlation serve as features for the classification task [5].

3.3 Classification

The classification task is performed by an Extra-Trees classifier. This is an ensemble decision-tree based approach which sub-samples the training data randomly and fits randomized decision trees to these sub-samples without replacement. The cut points used to split the nodes of the trees are selected at random. Averaging is then used to finalize the weights in such a way that the predictive accuracy is maximized while over-fitting is minimized [12]. The Extra-Trees classifier if trained on the features described in Section 3.2. Targeted audio samples are classified as either *facing* or *not-facing* as determined by a predefined range.

3.4 Target Audio Filtration

The intention of the filtering application is to segment the audio sequence by varying window sizes ℓ to permit certain windows of data to pass granted that they contain *facing* audio. The trained classifier will be used to discern segment of audio with facing content from segments of audio with not-facing content as defined by the threshold that classifier was trained to discern. Hann windows of varying sizes, ℓ, will be applied to the input audio prior to pre-processing for classification. If the audio is determined to be *facing*, the audio is concatenated to the output signal; if not, then a zero-signal of the same length as the audio segment is concatenated instead.

4 IMPLEMENTATION

4.1 Cross-Utterance Train-Test Split

Initially, a cross-utterance train-test split was devised for this work. Of the two utterance types contained in the dataset ("Hey assistant" and “The quick brown fox jumped over the lazy sheep"), all of the participant’s utterances of “The quick brown fox jumped over the lazy sheep" were held out from training. This subset of the dataset was used for testing purposes. An Extra-Trees classifier with 1000 estimators was trained to classify ±45° angles as *facing* and all other angles as *not-facing*. The results obtained with this train-test split are detailed in Section 5.1.

After careful consideration, it was determined that the room reverberation properties and participant voices could bias the classifier. For such a computational system to be generalizable, it is necessary that it is robust to participant voices and the environment in which it is employed. A suitable train-test split that was representative of this requirement was then devised.

4.2 Cross-Participant and Cross-Room Train-Test Split

The purpose of this train-test split is to achieve cross-speaker and cross-room robustness. Video calls may be conducted in rooms with drastically different reverberation properties and must be robust enough to handle new speakers and thus this train-test split encourages results that are capable of generalizing along these axes.

All utterances from two speakers, notably participant 2 and participant 9, representing a male and female voice profile, were omitted from the training set. The utterances recorded for these participants in the upstairs room setup were reserved for testing. Both utterance types from the remainder of the speakers from the downstairs room setup were used during training.

4.3 Classification Range

Classification ranges for voices projected along ±45° and ±90° axes were used to train two respective classifiers for discerning between *facing* and *not-facing* voices on the cross-participant and cross-room train-test split. Voices that were projected along an axis within each of the two ranges would be classified as *facing*. The model weights were saved and loaded for the filtration task. Any further mention of the ±45° or ±90° classifier references these models.

4.4 Curated Test Dataset For Filtration Task

Apart from being used to evaluate the performance of the classifiers, the held-out test set was also used to curate a test dataset for the filtration task. One *facing* audio segment was paired with another *not-facing* segment as determined by the direction of voice range in the specific use case (±45° or ±90° classification). The order in which the audio segments were concatenated was randomized. Some of these curated audio recordings are compounds of files from the test portion of the dataset. Others were acquired using the ReSpeaker 4-mic array to verify the robustness of the classifier.

4.5 Filtering Application

The filtering application accepts .wav audio files as input using librosa and removes the portions of the audio that were recorded when the speaker was not facing the microphone. Trained classifiers for both the ±45° and ±90° ranges were used to establish segments of audio with facing content. The complete implementation of this application is available on Github at https://github.com/emsal1863/554xproj. During implementation, the window size, (ℓ), was found to have a direct effect on the quality of filtering. Experiments were conducted to examine the consequences of changing the window size. This will be discussed further in Section 5.

5 EVALUATION

5.1 Classifier Performance

With 1000 estimators, a test accuracy of 82.8% accuracy on the cross-utterance test split was achieved for the ±45° classifier. This is approximately 5% less than the accuracy achieved by the DoV paper [5]. For the novel cross-speaker and cross-room test split, accuracies of 79.7% for the ±45° classifier and 81.9% for the ±90° classifier were attained.

5.2 Filter Performance

Window sizes of 200, 400, 600, 800, and 1000 ms were tested. The filter application outputs the *facing* or *not-facing* classification for each window of the audio signal. This was compared against the ground truth from the audio file. From this, the false positive and false negative classification rates were computed for each audio clip. The complete data is available in Table 1. For the shorter clips where a large enough window size would cause overlap between the facing and non-facing segments of the audio, the false positive and negative rates are listed as N/A. On the files that are compounds of files from the test dataset, the filtering application performed unilaterally well, usually getting a 100% classification rate. As for the recordings on the ReSpeaker array, there were some audio recordings on which the classifier performed well and some on which the classifier introduced many false negatives. Across the board, the classifier did always yield a low false positive rate, with the highest rate that could not be explained by an edge case being 2/11.

When using the ±90° classifier, however, the classifier generally introduced a much higher false positive rate on clips that were recorded outside of the original test dataset. It performed reasonably well on clips from the original dataset. On some of the newly recorded clips, the ±90° classifier unilaterally designated all windows as facing for a 100% false positive rate on all window sizes.

6 LIMITATIONS AND FUTURE WORK

6.1 False Negative Rate on Out-of-Dataset Clips

The cause for the potentially high false negative rate on the out-of-test-data samples is still subject to speculation. We suspect that some underlying difference separates the new recordings from the dataset recordings. Any of the following factors may contribute to the discrepancy, including:

- Volume discrepancy caused by shorter distance in the new clips compared to the dataset clips (the new clips had a higher average volume)
- Differences between room setups used in the original experiments, including placement of the microphone array
- Differences in the profile of the voices used in the dataset's recordings and the new recordings

In future work, when these causes are identified more precisely, it would be possible for the filtering application to transform new audio recordings in order to make them more similar to the audio recordings from the original dataset. The effectiveness of this transformation functionality would need to be experimentally validated.

6.2 Alignment problems

It is possible for a single window to contain portions of audio that are facing and portions that are non-facing. The classifier's behavior in these scenarios is still undetermined. In these cases, shifting the start of the windows such that the windows only contain one type of audio could end up yielding a more accurate prediction.

In such cases of mixed directions of voice in a single audio window, it would be beneficial to work towards automatically shifting the window when a new direction is detected to make the predictions more accurate. A finer sampling size to determine the precise transition between the recorded directions of voice may also assist in improving filter accuracy.

6.3 Classification time

During the course of the experiments, the script was also benchmarked to evaluate the amount of time that it took to classify each window. Throughout the course of all the experiments each window took between 650 and 750 ms to classify regardless of the window size.

This implies that there is an inherent trade-off between choosing smaller window sizes for better accuracy and the amount of time it takes to filter the entirety of the audio clip. This would only become more significant when attempting to filter longer audio files.

Furthermore, this classification time is also prohibitive to potential applications of the filtering application to low-latency live streaming, including video calls. (We note that it may be possible to apply this to live streaming if there were enough of a constant delay, such as the 10-second delay on popular live streaming website Twitch.tv). Efforts to lower the classification time will be central to accomplishing the eventual goal of applying the filtering live.

Possible ways that this reduction could be achieved is through optimizing or changing the model from the Extra-Trees classifier with the parameters that were used throughout the course of the project. Lowering the complexity (in this case, number of classifiers used in Extra-Trees) of the model could introduce a reduction in the quality of classification and filtering, so experiments or other validation would need to be done in order to find a threshold at which the quality of the classifications is satisfactory and the operation is fast enough for livestreaming.

It may be possible to use models other than Extra-Trees or use a variant of the decision tree-based classifier that is more suitable to parallel architectures such as those present on GPUs. It is worth noting that other types of classifiers did not perform as well as Extra-Trees on the original dataset as noted in [5]. Parallelizable variants of decision tree classifiers have been developed [13], which led to significant speed-ups over CPU-based classifiers of the same type, so it may be possible to use such an algorithm in this application.

6.4 Noise reduction and multi-speaker cases

Neither the original dataset or the newly recorded audio clips were created in environments with the presence of a significant amount of background noise from other speakers, other sound emitting sources or environmental noise. Although the ReSpeaker 4-Mic array automatically denoises audio, this was not tested to its limits.

Previous work [11] has used diffusion masks for better-quality direction-of-arrival estimation using GCC-PHAT. A potential extension of the filtering application is to attempt to use this augmented featurization in order to make the model more robust to diffuse noise.

7 CONCLUSION

In the course of undertaking this work, the featurization in [5] was deemed to be effective in predicting the direction of voice of audio signals. In fact, the featurization was so effective that a relatively simple machine learning method such as an extra-trees classifier was able to achieve high prediction accuracies. This even proved to be the case for a train-test split which was not originally explored by Ahuja et al. in the DoV work. Joint generalizability with respect to the room setup and speaker voice axes was tested and achieved by this method. This also extended to audio recordings using the same device in a drastically different and unconstrained room setting.

The work presented in this paper proposed a method for filtering non-facing audio segments from a continuous audio stream based on the direction of voice. The filtering application discussed in this work successfully segmented audio signals into their facing and non-facing segments. This was shown on data from the original DoV dataset, specially curated audio samples as well as realistic recordings with the ReSpeaker 4-Mic array. While progress is still required before deployment of this filtering application for live audio feed filtration as in a video call experience, promising leads have been presented in this paper and further developments in this area seem promising.

REFERENCES

[1] Lia M. Bonacci, Lengshi Dai, and Barbara G. Shinn-Cunningham. Weak neural signatures of spatial selective auditory attention in hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 146(4), October 2019.

[2] Nilanjan Dey and Amira S. Ashour. *Direction of Arrival Estimation and Localization of Multi-Speech Sources.* Springer Publishing Company, Incorporated, 1st edition, 2017.

[3] Nilanjan Dey and Amira S. Ashour. *Challenges and future perspectives in speech-sources direction of arrival estimation and localization.* Springer Publishing Company, Incorporated, 1st edition, 2018.

[4] Richard A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 3rd. annual workshop on Librarians and Computers*, page 262–270, New York, NY, USA, 1980. Association for Computing Machinery.

[5] 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.

[6] Ivan Meza, Caleb Rascon, Gibran Fuentes, and Luis A Pineda. On indexicality, direction of arrival of sound sources, and human-robot interaction. *Journal of robotics*, 2016, 2016.

[7] Cha Zhang, Dinei Florencio, Demba E. Ba, and Zhengyou Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Transactions on Multimedia*, 10(3):538–548, 2008.

[8] Seon Man Kim and Hong Kook Kim. Direction-of-arrival based snr estimation for dual-microphone speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2207–2217, 2014.

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

[10] Michael S. Brandstein and Harvey F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 375–378 vol.1, 1997.

[11] Ran Lee, Min-Seok Kang, Bo-Hyun Kim, Kang-Ho Park, Sung Q Lee, and Hyung-Min Park. Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments. *IEEE Access*, 8:7373–7382, 2020.

[12] Pierre Geurts and Damien Ernst. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[13] Toby Sharp. Implementing decision trees and forests on a gpu. In *European conference on computer vision*, pages 595–608. Springer, 2008.

A DATA

	200ms	400ms	600ms	800ms	1000ms
"Streamer clip": False negative rate	0.5333333	0.4666667	0.5	0.625	1
"Streamer clip": False positive rate	0.07692307692	0	0	0	0
"Podcast clip": false negative rate	0.7352941176	0.8823529412	0.8181818182	1	1
"Podcast clip": false positive rate	0.1481481481	0	0.2222222222	0	0
Dataset file 1: s9 no-wall 0 deg + s2 no-wall 180 deg: false negative rate	0	0	0	0	0
Dataset file 1: s9 no-wall 0 deg + s2 no-wall 180 deg: false positive rate	0	0	0	0	1
Dataset file 2: s2 nowall 225 deg vs s9 nowall 0 deg: false negative rate	0	0	0	0	0
Dataset file 2: s2 nowall 225 deg vs s9 nowall 0 deg: false positives	0.07692307692	0	0	0	0
90-90 model: "Streamer clip":False negative rate	0	0	0	0	0
90-90 model: "Streamer clip": False positive rate	1	1	1	1	1
90-90 model: "Podcast clip": false negative rate	0	0	0	0	0
90-90 model:"Podcast clip": false positive rate	1	1	1	1	1
90-90 model: Dataset file 1: s9 no-wall 0 deg + s2 no-wall 180 deg: false negative rate	0.1428571429	0	0	N/A	N/A
90-90 model: Dataset file 1: s9 no-wall 0 deg + s2 no-wall 180 deg: false positive rate	0.2857142857	0	0	N/A	N/A
90-90 model: Dataset file 2: s2 nowall 225 deg vs s9 nowall 0 deg: false negative rate	0	0	0	0	0
90-90 model: Dataset file 2: s2 nowall 225 deg vs s9 nowall 0 deg: false positives	0.4615384615	0.6666666667	0.25	0	0.5

Table 1. False positive and negative rates from varying window size classifications of audio clips