ENHANCED SPEECH DETECTION MODELS FOR AIR TRAVEL DISEASE RISK

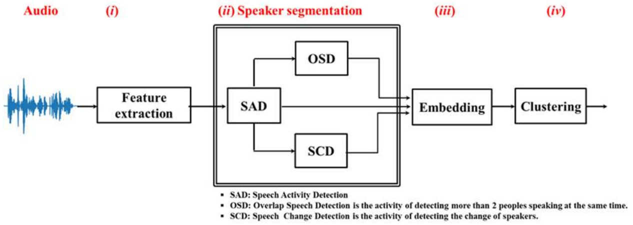
Tenon Charly Kone [*1], Sebastian Ghinet [†1], Sayed Ahmed Dana [‡1] and Anant Grewal [♦1]
[1]National Research Council Canada, Flight Research Laboratory, Ottawa, Ontario, Canada.

1 Introduction

In today's interconnected world, air travel is a vital mode of transportation for people and goods, yet it also poses risks for spreading communicable diseases due to the large number of travelers sharing the confined spaces of airport terminals and aircraft cabins. Elevated noise levels in these environments, often prompt travelers to speak louder and gather closely, inadvertently increasing the risk of spreading respiratory particles and infectious agents. Understanding this risk is crucial. This paper discusses the use of advanced signal processing techniques, specifically artificial intelligence-based speaker diarization [1], to accurately analyze speech patterns in noisy environments with multiple interlocutors. Speaker diarization involves determining "who spoke when" in a multi-speaker audio stream, with applications ranging from information retrieval to healthcare, including neuropsychology and COVID-19 analysis [2]. Despite its potential, the application of speaker diarization in noisy environments remains limited. This paper explores its relevance in such settings, particularly in air travel scenarios, aiming at the development of measures and protocols to effectively manage communicable diseases in confined spaces like airports and aircraft cabins.

2 Modeling

Speaker diarization, the process of segmenting an audio stream based on speaker identity, is supported by various open-source tools. S4D, an extension of SIDEKIT [3], is a Python package offering comprehensive speaker diarization capabilities. It covers the entire processing chain, from audio data to system performance analysis, with educational and practical aims. In contrast, Kaldi [4] provides speaker diarization recipes but is not Python environment and is mainly focused on speech and speaker recognition systems. ALIZÉ and its LIASpkSeg extension, developed in C++, lack recent deep learning approaches for speaker diarization. For broader audio signal analysis, pyAudioAnalysis, implemented in Python, is available and adaptable for speaker diarization. Alternatively, pyannote.audio [5], a Python-based open-source toolkit utilizing the PyTorch machine learning framework, offers end-to-end neural building blocks for speaker diarization pipelines with pre-trained models. In this paper, pyannote.audio software was chosen for its user-friendly interface and a wider range of pre-trained models. Its operational principle involves five steps (Fig. 1): (*i*) Feature Extraction, (*ii*) Speaker Segmentation, (*iii*) Speaker Embedding, Clustering of Speaker Embeddings, and (*iv*) Speaker Diarization. These steps encompass crucial tasks like extracting relevant information from raw audio data, segmenting conversations into distinct speaker turns, generating concise representations of a speaker's voice, clustering similar speaker embeddings, and ultimately identifying different speakers in the audio stream.



- SAD: Speech Activity Detection
- OSD: Overlap Speech Detection is the activity of detecting more than 2 peoples speaking at the same time.
- SCD: Speech Change Detection is the activity of detecting the change of speakers.

Figure 1: Speaker diarization process diagram.

3 Speaker segment characterization

Characterizing speaker segments involves evaluating the overall sound pressure level associated with an individual during a conversation. The process begins by identifying "who speaks and when?" Utilizing established diarization tools like Pyannote, coupled with a Python script, allows the calculation of the Overall Sound Pressure Level (OASPL) for each speech segment resulting from the diarization.

3.1 Audio track

To validate the approach, three audio samples were compiled in-house, each containing the same 20 sentences delivered by four different speakers. The first audio track (Audio 0) served as the baseline configuration, with no changes made to the peak Sound Pressure Level (SPL) between consecutive speakers. In the second track (Audio 1), the SPL peak difference between consecutive speakers was set to 8 dB, resulting in peak level adjustments of +0 dB and +8 dB between speaker segments. In the third track (Audio 2), the SPL peak difference between consecutive speakers was increased to 16 dB. These compilations constituted a validation dataset for assessing the efficacy of the developed speaker segment characterization tool. The difference in SPL can be observed in the waveform and frequency spectrum of each audio track, allowing for an easy detection of alternating acoustic levels between consecutive speakers (Fig. 2).

3.2 Results

The OASPL is a crucial metric when assessing the impact of noise, as it provides a comprehensive measure of the overall intensity or loudness of a sound across different frequencies within a specified frequency range. Its expression is given by

$$OASPL = 10\log_{10}(\textstyle\sum_{N_{freq.}} V), \quad (1)$$

[*] TenonCharly.Kone@nrc-cnrc.gc.ca
[†] Sebastian.Ghinet@nrc-cnrc.gc.ca
[‡] Dana.SayedAhmed@nrc-cnrc.gc.ca
[♦] Annant.Grewal@nrc-cnrc.gc.ca

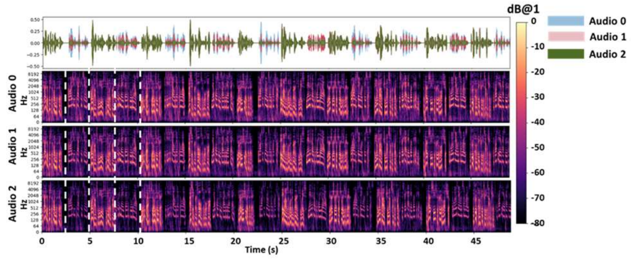


Figure 2: Audio tracks: Waveforms (top figure) and spectrum (3 bottom figures).

where $V = PSD/p_{ref}^2$ with PSD is Power Spectral Density of the signal with a reference to unit.

The characterization of speaker segments for the three aforementioned audio tracks is detailed in Fig. 3. For each audio track, the diarization model successfully identified four speakers along with the corresponding timestamps for each speaker. The Fig. 3 showcases the speech segments of each speaker, where each segment aligns with a sentence attributed to a specific speaker. The OASPL for each segment, measured in dB, is indicated above or below the corresponding speaking duration segment. All values have been rounded to the nearest integer and calculated using a reference value of $p_{ref} = 1$.

In the case of Audio 0, the baseline audio, all speakers exhibit an OASPL of -82 dB. Transitioning to Audio 1, the developed tool predicts an average decrease of 5 dB for speaker 02, resulting in an average value of -87 dB, and a decrease of 2 dB for speaker 03. However, the OASPL for the other speakers, 00 and 01, remains unchanged at -82 dB. Moving on to audio 2, a notable average decrease of -12.5 dB for speaker 02 and -10 dB for speaker 03 is predicted, while the OASPL for speakers 00 and 01 remains constant at -82 dB. These reductions affirm the efficacy of the developed algorithms in capturing variations in sound pressure levels across different speakers and audio tracks.

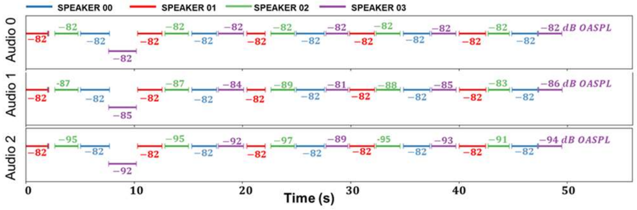


Figure 3: Speaker segment characterization prediction.

To summarize, these fluctuations demonstrate that the developed algorithms, accurately capture the variations in sound pressure levels across diverse speakers and audio tracks.

4 Conclusion

In conclusion, this study highlights the importance of accurately understanding the communicable diseases spreading risks due to speaking in noisy environments, such as airports or aircraft cabins, where elevated noise levels prevail. The involuntary behavior of speaking loudly in these settings inadvertently increases the dispersion of respiratory particles, potentially carrying infectious agents. To address this challenge, advanced signal processing strategies, particularly AI-based speaker diarization approaches, were implemented to accurately determine speech patterns, including duration and sound pressure level. Implementing and adapting these algorithms offer promising tools aiming at the development of measures and protocols to manage communicable diseases spread in air travel settings.

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, Speaker diarization: A review of recent research, IEEE Trans. Audio Speech Lang. Process., vol. 20, no. 2, pp. 356-370, 2012.

[2] B.W. Schuller BW, Schuller DM, Qian K, Liu J, Zheng H, Li X. COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis. Front Digit Health. 2021 Mar 29;3:564906.

[3] A. Larcher, K. A. Lee and S. Meignier, "An extensible speaker identification sidekit in Python," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5095-5099.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, et al., "The kaldi speech recognition toolkit", IEEE 2011 Workshop on Automatic Speech Recognition and Understanding., Dec. 2011.

[5] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W.Bouaziz, and M-P. Gill Pyannote. audio: Neural building blocks for speaker diarization. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 7124–7128 (IEEE, 2020).