Interspeech 2024
1-5 September 2024, Kos, Greece

Detection of Background Agents Speech in Contact Centers

Abhishek Kumar, Srikanth Konjeti, Jithendra Vepa

Observe.AI, India

{abhishek.kumar, srikanth.konjeti, jithendra}@observe.ai

Abstract

In a typical contact center environment, multiple agents often handle calls simultaneously and they are frequently in close proximity to one another. Consequently, there is a possibility that conversations of nearby agents may inadvertently be recorded during calls. This represents instances of background agents speech being captured during agent-customer interactions. Such unintended background speech may not only impact the quality of conversation but may also contain some sensitive information which may pose security concerns in contact centers. Therefore, contact centers are interested in identifying such scenarios. This knowledge can assist them to implement appropriate mitigating strategies and enhance the quality of audio conversations, thereby improving the overall customer experience. In this work, we utilise the pauses and gaps in the agent speech to clearly identify the background speech. Our approach that is based on speech features is simple, tuneable, computationally efficient and cost effective.

Index Terms: background speech, contact centers, pitch, intensity, spectrogram

1. Introduction

Contact centers usually have a busy and dynamic environment characterised by frequent interactions between agents and customers. The environment can be relatively loud if a large number of agents are working simultaneously in close proximity. As a result, various background noises from the agent side may inadvertently be recorded during these interactions, including background speech, background music, ringing telephones, notification alerts, and more. Among these, background agents speech is more frequent and of greater concern. Background speech may contain sensitive information and can lead to breaches of confidentiality, data privacy issues, identity theft, compliance violations, etc. Additionally, these background noises significantly impact the overall quality of the calls by causing misunderstandings, repeated requests for clarifications, reduced agent productivity, etc.

Detecting background noise in contact center calls can help in understanding its occurrence and patterns, enabling contact centers to implement effective noise mitigation strategies. This not only enhances the call quality and addresses security risks but also ensures operational efficiency and customer satisfaction within the contact centers. Additionally, it will contribute to create a conducive environment for effective communication between agents and customers.

In a contact center call, background noise may originate from either the agent or the customer side. The noise from the customer side may vary depending on the specific environment on that end, and the contact center may have limited control over it. In this work, we will focus on identifying the presence of background speech or noise in the agent channel of a stereo call. To achieve this, we will primarily utilise speech features such as pitch (fundamental frequency) and intensity of the speech. A similar approach could also be applied to detect background noise in the customer channel.

2. Background Speech Detection Algorithm

The fundamental concept of the algorithm is to identify background speech during the pauses and gaps in the agent channel of a stereo call. While the presence of background conversations is noticeable when the agent is speaking, it becomes more distinctly identifiable during pauses in the agent's speech, particularly when the customer is speaking. These pauses create distinct segments making it easier to isolate and analyse. This analysis allows for specific targeting instances where background noise is most likely to occur, leading to more accurate detection. Therefore, in this approach, we aim to utilise the pauses of the agent for more effective and targeted approach to detect the background speech, as shown in Figure 1. Importantly, this detection occurs at the call level rather than at a segment level, as only the segments between the pauses containing background speech in the agent channel are utilised.

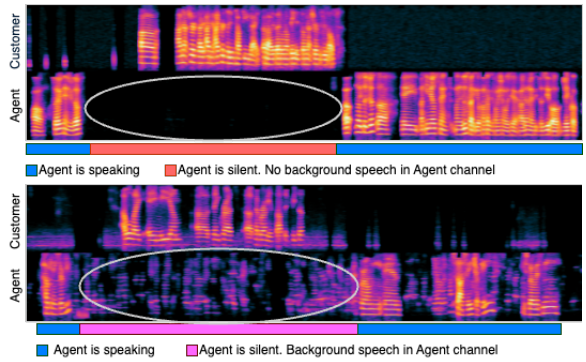

Figure 1: *Spectrograms (in oval) showing presence or absence of background speech in the Agent channel*

3. System Overview

3.1. Call Segmentation

The call transcript is generated using an Automatic Speech Recognition (ASR) system, providing word timing information for all words spoken during the call. We leverage this timing data to identify non-speech segments in the agent channel of stereo calls. If the time gap between consecutive words exceeds a threshold (2.5 seconds), these timings are categorised as segments within the non-speech category.

3.2. Feature Extraction

The Parselmouth library [1] is used to extract frame-level Fundamental Frequency, F0 [2] instants within each segment, with a frame size of 40 ms and a frame shift of 10 ms. The F0 instants is refined by iterating through it within each segment and checking for continuous sub-segment of non-zero F0 values. If the sub-segment duration is less than a threshold (100 ms), F0 of the corresponding sub-segment is assigned zero values. Finally, **voiced regions** for each segment is obtained by multiplying the number of speech frames having non-zero F0 values with frame size.

Furthermore, intensity instants (in dB scale) is extracted using the Parselmouth library for each segment and the mean statistics are computed from **non-zero intensity** instants. Only non-zero intensity instants are considered as focus is on those portions of a segment where background speech may be present.

Using the **voiced regions** as well as **mean intensity** for each non-speech segment of a call, a threshold based decision is made for the background speech.

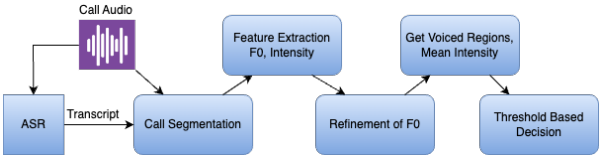

Figure 2: *Background Speech Detection System*

4. Experimental Results and Application

Datasets: We used an internal dataset consisting of 249 annotated samples, wherein 118 samples had some background speech in the Agent channel, while the remaining 131 samples did not. With this dataset, we employed voiced regions and mean intensity and conducted experiments with decision tree-based classifiers to determine the threshold values for the following parameters:

%call_voiced_duration: Total duration of voiced regions (in seconds) as a percentage of the total duration of non-speech segments (in seconds) within a call

%seg_voiced_gte_20pc: Percentage of non-speech segments where a segment has voiced regions greater than 20% of the segment duration.

%seg_intensity_gte_30: Percentage of non-speech segments where a segment has mean intensity greater than or equal to 30 dB.

Table 1: *Precision, Recall and F1-score using different parameters and corresponding thresholds for background speech*

Parameters & Thresholds	Precision	Recall	F1-score
%call_voiced_duration > 4.15	**0.95**	0.983	**0.966**
%seg_voiced_gte_20pc > 4.88	0.853	0.838	0.844
%seg_intensity_gte_30 > 4.88	0.729	0.822	0.772
%call_voiced_duration > 4.15 & %seg_voiced_gte_20pc > 4.88	0.847	**0.991**	0.913
%call_voiced_duration > 4.15 & %seg_intensity_gte_30 > 4.88	0.738	0.983	0.843
%call_voiced_duration > 4.15 & %seg_voiced_gte_20pc > 4.88 & %seg_intensity_gte_30 > 4.88	0.717	**0.991**	0.832

The results for various combinations of these parameters are tabulated in Table 1. Based on these, we observe that **%call_voiced_duration** is **most effective for background speech detection**, yielding higher F-1 score.

While voiced regions effectively identify background noise, integrating intensity is sometimes necessary to address scenarios leading to false positives. One such instance involves music as an IVR (Interactive Voice Response) in the agent channel when the agent places the customer on hold or wait. In such cases, voiced regions tend to have higher values, resulting in %call_voiced_duration exceeding the threshold. Music as IVR typically exhibits very high mean intensity, often surpassing 70 dB. Conversely, nearly all mean intensity values for non-speech segments, even those with background noise present, remain below 60 dB. Hence, we filter out such music segments using mean intensity from the decision-making process.

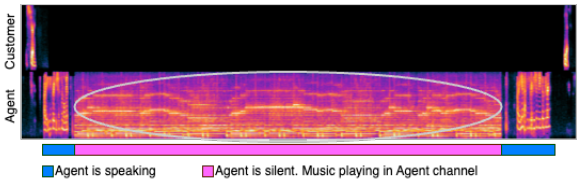

Figure 3: *Case of Music as IVR in Agent Channel*

The Background Speech Detection algorithm has been implemented on a contact center call and background noise evidences have been shown in the Figure 4. The dashboard displays transcriptions of both the agent and the customer at a turn level. First, background noise is identified at the call level, then the corresponding evidence is presented for agent turns where the agent pauses and sufficient voiced regions are detected.

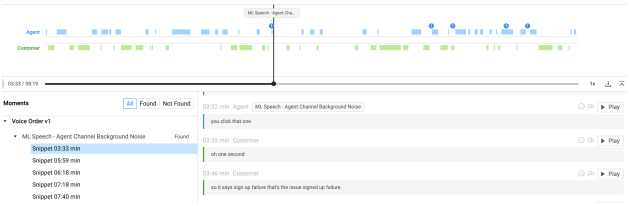

Figure 4: *Dashboard showing Agent Channel Background Noise*

This aids contact centers in identifying calls with significant background noise. By examining the background noise evidences linked to these calls, specific reasons can be found by listening to snippets. Finally, contact centers can work on the solutions to mitigate them effectively.

5. Conclusion

In this work, an algorithm has been proposed to detect background speech in the agent channel of contact center conversations and the results have shown that the proposed algorithm works effectively for this scenario. This algorithm uses speech features based on pitch (fundamental frequency) and intensity and makes the decision using thresholds. Detecting background agents speech can significantly help the contact centers in effectively addressing this issue , thereby enhancing the quality of conversations and mitigating security risks.

6. References

[1] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[2] P. S. Rathore and R. B. Pachori, "Instantaneous fundamental frequency estimation of speech signals using desa in low-frequency region," in *2013 International Conference on Signal Processing and Communication (ICSC)*, pp. 470–473, IEEE, 2013.