Crowded Room Data Findings

Musy Ayoub

June 10th 2022

1 Goal

Given the ConfLab Dataset, see if I can quantify the intelligibility of the actual conversation, which has been modified to be low-frequency audio (1250 Hz), and if the audio still preserves the privacy of the speakers in the audio, given the low recording frequency.

2 Data Assumptions

The Conflab Dataset contains: the audio of 48 participants, with lower pitched voices and higher pitches voices, sampled at a frequency that keeps the conversation content ambiguous, their overhead picture with their respective number corresponding to their audio number, elevated video containing audio of the whole room and their body and faces visible at different angles.

I used headphones that allow for listening to frequencies ranging from 10 Hz - 40000 Hz. The audio has not been tampered with in any way.

The analysis procedure is as follows: I considered 6 participants, 3 with lower pitched voices and 3 with higher pitched voices. I then looked at specific metrics that are explained below and why they were chosen.

3 Metrics

I have decided to use five different metrics.

1. **Words Perceived in 3 minutes.** I defined a perceived word to mean a sound that could be considered a separate word, the word itself does not need to be actually recognized in English.

   Words perceived in 3 minutes has been chosen because trying to estimate how many words or sounds that sound like the beginning and end of a word does give us a good insight of whether this audio is too deprecated or still can be used to detect if multiple words are being said.

2. **Words recognized in 3 minutes.** I defined words recognized by words that can be heard from the audio and also distinguished what they actually mean in the English language.

   Words recognized in 3 minutes has been chosen because if actual words can be recognized from the audio, that should be okay, but if too many words get recognized in general, that could pose a problem with the original thought of not wanting the audio to be easily understandable by just listening.

3. **Longest chain of words recognized in 3 minutes.** I defined longest chain of words recognized by observing what the longest chain of recognized words are in a row, if the chain breaks, that means a new chain starts and the longest chain would be chosen for this metric.

   Longest chain of words recognized in 3 minutes has been chosen because just recognizing random words did not seem always very useful, given they can be taken out of context easily

and still not be quite clear what the actual context was. But if there are a lot of words in a row that are recognized, that would mean that the context can be more easily constructed from them.

4. **Amount of audio recognized in 3 minutes.** Amount of audio recognized in 3 minutes is dependent on $\frac{\text{Words recognized in 3 minutes}}{\text{Words perceived in 3 minutes}}$. With 1 being the maximum and 0 the minimum.

   Amount of audio recognized in 3 minutes was chosen because it would put in perspective how many words are actually recognized from all the words that were said in the audio by the person, giving more context how much of the audio snippet can be recognized.

5. **How many speakers were participating in the conversation.** This metric explains how many speakers could be heard from the audio.

   How many speakers were participating in the conversation was chosen because it was an interesting way to see how accurate the audio snippets are to distinguish different voices that also got caught in the participants audio recording device.

All these evaluations will be done with and without video. This will also test how useful it is to have video with it and how much more accurate these metrics will be. These 3 minute snippets of the conversation are all different, where the speaker did actively participate in the conversation.

4 Predictions

Given that the audio was quite deprecated, I can assume that sounds can be recognized as words, but the actual content of the words most probably will be difficult to decipher, therefore a low amount of words recognized and an even lower amount of words recognized in a row will most likely be the result, therefore having the Amount of audio recognized be low. I can also say that the results with video will be more accurate then without, given mouth and arm movements can be seen, which makes it easier also to distinguish who is talking and when they are actually talking.

5 Results

I get the following results (Without video / with video):

- Participant 36 (higher pitched voice)

  Words perceived in 3 minutes: 98 / 126
  - Words recognized in 3 minutes: 2 / 2
  - Longest chain of words recognized in 3 minutes: 2 / 2
  - Amount of audio recognized in 3 minutes: $\frac{2}{98}$ / $\frac{2}{126}$
  - How many speakers were participating in the conversation: 3 / 4
- Participant 22 (lower pitched voice)

  Words perceived in 3 minutes: 184 / 85
  - Words recognized in 3 minutes: 0 / 3
  - Longest chain of words recognized in 3 minutes: 0 / 3
  - Amount of audio recognized in 3 minutes: $\frac{0}{184}$ / $\frac{3}{85}$
  - How many speakers were participating in the conversation: 2 / 2

- Participant 35 (lower pitched voice)

  Words perceived in 3 minutes: 153 / 105
  - Words recognized in 3 minutes: 0 / 0
  - Longest chain of words recognized in 3 minutes: 0 / 0
  - Amount of audio recognized in 3 minutes: $\frac{0}{153}$ / $\frac{0}{105}$
  - How many speakers were participating in the conversation: 2 / 2
- Participant 43 (higher pitched voice)

  Words perceived in 3 minutes: 138 / 130
  - Words recognized in 3 minutes: 0 / 0
  - Longest chain of words recognized in 3 minutes: 0 / 0
  - Amount of audio recognized in 3 minutes: $\frac{0}{138}$ / $\frac{0}{130}$
  - How many speakers were participating in the conversation: 4 / 3
- Participant 45 (higher pitched voice)

  Words perceived in 3 minutes: 162 / 135
  - Words recognized in 3 minutes: 3 / 3
  - Longest chain of words recognized in 3 minutes: 2 / 2
  - Amount of audio recognized in 3 minutes: $\frac{3}{162}$ / $\frac{3}{135}$
  - How many speakers were participating in the conversation: 3 / 4
- Participant 12 (lower pitched voice)

  Words perceived in 3 minutes: 170 / 151
  - Words recognized in 3 minutes: 7 / 11
  - Longest chain of words recognized in 3 minutes: 3 / 3
  - Amount of audio recognized in 3 minutes: $\frac{7}{170}$ / $\frac{11}{151}$
  - How many speakers were participating in the conversation: 2 / 3

6 Conclusion

Here I are going to discuss the different metrics and how they performed.

- **Words perceived in 3 minutes**: The sounds that I could distinguish as words were quite easier to count and differentiate than thought, the numbers that were acquired, with and without video, show it is possible to dissect around how many words they are spoken in a certain time frame. Even though, as predicted, having audio and video will benefit in the accuracy of counting the words, given body language and lip movements are shown and when the person in question is actually speaking or not, or just some other speaker who else got caught by the device, but was loud enough or sounded similar to the speaker, this was mostly a problem with the participants with a lower pitched voice. Participants with higher pitched voices were slightly harder to follow the word count from than participants with lower pitched voices.

- **Words recognized in 3 minutes**: As expected, there were not many words actually recognized from the counted words. The audio is mostly too deprecated for this, unless someone was quite loud and said very simple terms like "exactly" or "yeah, but...". The most context I have got was from participant when I heard the words "The problem is...", which can only be said that there was a negative connotation, but I have no idea what the actual context was and what the sentence afterwards portrayed. This is a good way to show that the frequency that the audio was sampled as is good enough to count the words, but not good enough to actually recognize a lot of words and get actual context on what the conversations were about. Video did make some words easier to recognize, but barely helped when looking at the differences of actual words recognized. The audio here would be the most important deciding factor for recognizing actual words.

- **Longest chain of words recognized in 3 minutes**: Most of what was said before at **Words recognized in 3 minutes** can be applied here as well, it was expected that recognizing a chain of words is quite harder than recognizing separate words. Video here did not help in this process again. It is true that the most context I have gotten was still not nearly enough to know what the actual conversation was about.

- **Amount of audio recognized in 3 minutes**: Looking at the Amount of audio recognized in 3 minutes, all of them are quite low, given that the numbers don't even pass 0.1. This would imply that the conversations are barely recognizable from the audio snippets I have listened to. Video often did increase the Amount of audio recognized in 3 minutes, but did not give a substantial improvement to the understanding of the context of the conversations.

- **How many speakers were participating in the conversation**: Because of the people in the background, it was often hard to assume how many people were actively participating in a conversation. Sometimes it seemed like there were more people joining in on an conversation than there actual were. And sometimes there were less people than expected, because their voice was not registered by the recording device. Having less voices on a track did make it quite easier to actually distinguish the words the actual participant was speaking. Sometimes the participant sounded quite similar to another participant, which made us think there were only 2 participants, when there were actually 3. Which made some of the **Words perceived in 3 minutes** findings without video more than the ones observed with video. Video did help quite a lot here of course, given you can directly see how many people are interacting, which is quite hard to do with only audio.