How Does Reduction in Sample Frequency Hinder the Detection of Words?

Lucia Alonso Arenaza - supervised by Hayley Hung and Jose Vargas Quiros

1 Introduction

Language recognition software is present in personal devices that are used daily such as mobiles and tablets, or electronic home devices such as Alexa to ease and improve the usability of technology.

However, this technology can also be used spitefully [1]. This fact raised concerns about privacy. Even then, there is not much research done in this aspect of language recognition.

To address this knowledge gap, this research focuses on the analysis of how the reduction in sample frequency hinders the detection of words and affects the privacy of the speaker.

2 Background and Related Works

Automatic Speech Recognition (ASR) software, as a part of language recognition, processes the voice's audio signal in various pipelined steps to obtain a text representation of the speech.

There are several commercial ASR systems such as those by Google, Apple, etc., and some open-source systems; CMU Sphinx and Kaldi [2].

Multiple initiatives such as, RhythmBadge developed by MIT in 2018 and ConfLab developed in TU Delft, devoted at studying social interactions without violating the privacy of the participants [3,4].

3 Methodology

A The audio and dataset

March15LaRedBirthdayParty contains 16 audio files. Each audio file corresponds to one speaker, there were 16 people wearing a microphone. The audio was recorded at a standard sample rate of 44.1kHz.

B Processing the Audio

From each audio, one minute was selected. Background noise and the vocal foreground was separated using Librosa's documentation [5]. Files were down-sampled using a low-pass filter.

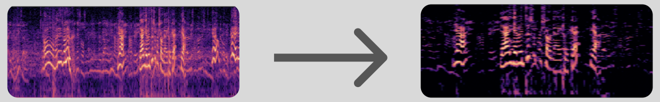

Figure 1: An example of how the vocal foreground is separated from the background noise.

C ASR Methods

Google Speech Recognition software which is a library offered by Google that supports multiple languages including Dutch.

Kaldi-NL, an existing model trained for Dutch language.

D Evaluation

The metric used in this research is **Word Error Rate (WER)** and it's calculated with the following formula:

$$WER = (S + D + I)/N$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the total number of words.

4 Results

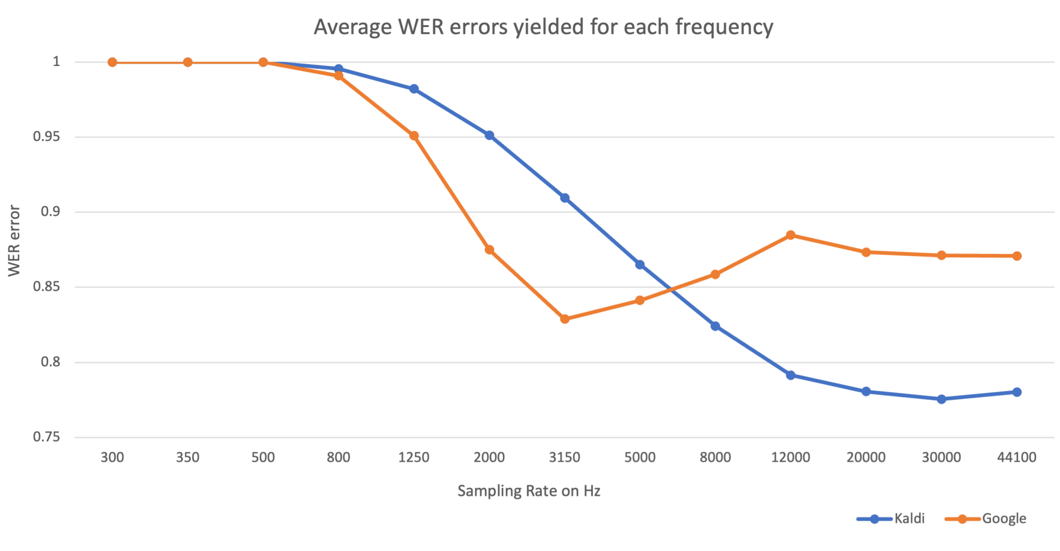

Figure 2: A visual representation of the average WER errors yielded for each frequency using Kaldi and Google Speech Recognition soft- ware.

We have choosen to set the threshold of unintelligibility when the WER is 1, that is, the transcription software returns no text. Otherwise, even if a few words are picked up by the software we can't assure that there is no critical information about the speaker.

We can conclude that for both Google Speech Recognizer and Kaldi-NL, the audio files are unintelligible at 500Hz and below.

5 Discussion & Future Work

What effects do accents have in lower sample frequencies?

- Early versions of Siri would not recognize audio from people with different English accents [6].

Would the results obtained in this research be applicable to not widely used languages?

- Not so widely spoken languages, have fewer data to train models with and results are usually less accurate [6].

There are also people with speech impairments, will these results be accurate in these cases?

- We haven't worked with audio from people with speech impediments so we were not able to determine how this condition impacts the accuracy of the results.

What would the results be when there is more than one language in the same conversation?

- Right now, the ASR used in this research needs one predefined language from the start.

[1] J. Y. Hui and D. Leong, "The Era of Ubiquitous Listening: Living in a World of Speech-Activated Devices," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 3021623, Aug. 2017. Accessed: Jun. 17, 2022. [Online]. Available: https://papers.ssrn.com/abstract=3021623

[2] G. Bohouta and V. Këpuska, "Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)," Int. Journal of Engineering Research and Application, vol. 2248-9622, pp. 20–24, Mar. 2017, doi: 10.9790/9622-0703022024.

[3] O. Lederman, "Rhythm Badge," MIT Media Lab. https://www.media.mit.edu/posts/rhythm-badge/ (accessed Jun. 18, 2022).

[4] C. Team, "The Socially Perceptive Computing Lab," 2019. [Online]. Available: https://conflab.ewi.tudelft.nl

[5] Librosa Development Team. Vocal separation – librosa-gallery 0.1.0 documentation. URL: https : / / librosa.org/librosa gallery/auto examples/plot vocal separation.html (visited on May 7, 2022).

[6] A. Koenecke et al., "Racial disparities in automated speech recognition," Proceedings of the National Academy of Sciences, vol. 117, no. 14, pp. 7684–7689, Apr. 2020, doi: 10.1073/pnas.1915768117.

L.AlonsoArenaza@student.tudelft.nl

21/06/2022