The effects on speech detection of low sample frequency audio data

Taichi Uno (5056763, t.uno@student.tudelft.nl)

1. Introduction and Background

ConfLab[1]

- A social experimental event that collects data of participants.
- Collects the audio data at a low sample frequency (1250Hz)

Background information

- Aliasing and loss of data due to downsampling (Nyquist frequency)[2]
- VAD (Voice Activity Detector) : Technology to detect if someone is talking or not. Many approaches exists including supervised and unsupervised. [3]
- “Rhythm” by MIT (Similar to ConfLab but uses 700Hz sample frequency)[4]

2. Research Question

Main Question :

“How does the reduction in sample frequency hinder the detection of speaking?”

Subquestions :

- How does performance of VAD change over different sample frequencies?
- Is there any difference between different methods of speech detections?
- Is there a difference between human ears and a machine in terms of the detection of speaking in low and high sample frequency data?

3. Methodology

These two state-of-art VADs are used to compare its performance.

rVAD (Unsupervised model)[5]

- Robust to both stationary and burst-like noise
- Pitch mode detects speech based on a posteriori SNR weighted energy difference.
- Flatness mode relies on a simple spectral flatness based detector.

Pyannote (Supervised model)[6]

- RNN for classification of feature vectors
- Trained with 100 hours of meeting recordings sampled at 16000Hz.

4. Experiment Setup

Data set used : March15LaRedBirthdayParty (Contains : chatting, background music and noises, silence). 1-3 hours long audio with 12 different speakers

Experiment frequencies : 300, 350, 500, 800, 1250, 2000, 3150, 5000, 8000, 12000, 20000, 30000, 44100Hz

Ground Truth : rVAD pitch mode at 44100Hz

Metrics : False Alarm Rate (FAR), False Rejection Rate (FRR), False Error Rate (FER)

5. Result

Fig1. FAR over different sample frequencies

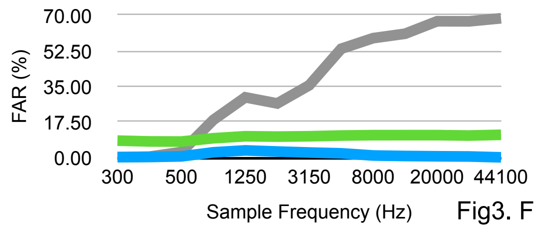


Fig2. FRR over different sample frequencies

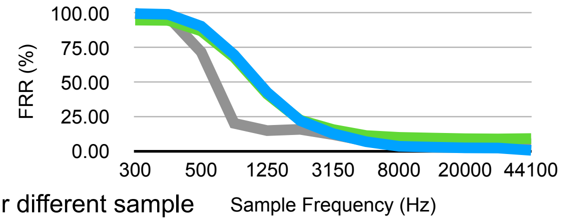
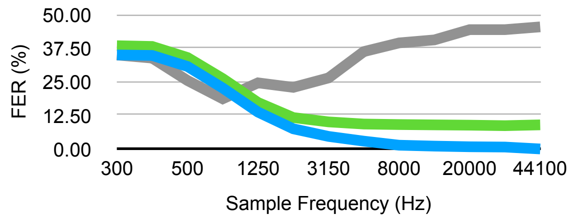


Fig3. FER over different sample frequencies



6. Discussion and Conclusion

- For the unsupervised methods, higher performance for higher sample frequency.
- The unsupervised outperformed the supervised.
- rVAD pitch mode works as good as the state-of-art supervised model at 8000Hz or higher, as the unsupervised one at 2000Hz or higher.
- Unexpected result for pyannote (supervised) (Higher FAR for higher sample rates)
- Human ears have better detection ability (The content partially is recognisable at 2000Hz)
- Not possible to use downsample audio to detect speech while preserve privacy.

7. Future Work

- Train the supervised model with more similar data and use different sample frequencies.
- Use other types of VAD to be able to generalise more.
- Scientific human experiments for speech detection and compare with computers.

8. Reference

[1] D. U. o. T. The Socially Perceptive Computing Lab. (2019) Conflab - acm mm 2019. [Online]. Available: https://conflab.ewi.tudelft.nl

[2] R. E.Isufi, “Cse2220 signal processing sampling iir filters,” 2020.

[3] S. S. Meduri and R. Ananth, “A survey and evaluation of voice activity detection algorithms,” 2012.

[4] O.Lederman, A. Mohan, D. Calacci, and A. S. Pentland, “Rhythm: A unified measure- ment platform for human organizations,” IEEE MultiMedia, vol. 25, no. 1, pp. 26–38, 2018.

[5] Z.-H.Tan, A. kr. Sarkar, and N. Dehak, “rvad: An unsupervised segment-based robust voice activity detection method,” Computer Speech Language, vol. 59, pp. 1–21, 2020. [Online].Available:https://www.sciencedirect.com/science/article/pii/ S0885230819300920

[6] H. B. et al, “pyannote.audio: neural building blocks for speaker diarization,” 2019. [Online]. Available: https://arxiv.org/abs/1911.01255