Laughter detection in privacy-sensitive audio

Matteo Fregonara
Supervisors: Hayley Hung and Jose Vargas Quiros

TU Delft

1. Background

Privacy is a major concern when studying audio recordings of conversations in the field of social signal processing.

Decimation, the process of lowering the sample frequency and reducing the bandwidth of an audio signal, can be used to theoretically make speech unidentifiable [1] while still preserving data about social cues such as laughter. However, this has not been experimentally verified.

Laughter detection models can be used for automatic laughter recognition. Background noise in an important factor to take into account when choosing what model to use.

2. Research Question

How does the reduction in sample frequency hinder the detection of laughter?

References

[1] J. Shen, O. Lederman, J. Cao, F. Berg, S. Tang, and A. Pentland, "Gina: Group gender identification using privacy-sensitive audio data," 2018 IEEE International Conference on Data Mining (ICDM), 2018.

[2] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust laughter detection in noisy environments," Interspeech 2021, 2021.

3. Method

Sub-question: How does the performance of laughter detection models get affected by downsampling?

Model [2]:

- ResNet (Residual Network)
- mel spectrograms features
- trained with sample frequency of 8kHz
- good performance with background noise
- hyper-parameter: **threshold**
  - can be used to recover higher percentage of laughs at the cost of more potential FP

LaRed dataset:

- dataset of recorded conversations
- recorded at 44.1 kHz
- downsample audio manually when needed for evaluation
- annotations of laughter with timestamps

Evaluation:

- evaluate performance over a range of sample frequencies: 300Hz - 44.1Hz (using log scale)
- use threshold values: 0.5 - 0.8

Simulate decimation:

- model's ResNet fitted to audio with sample frequency of 8kHz
  - therefore cannot directly decimate audio to desider frequency and feed it to model for evalation
- Instead, simulate decimation through low-pass filtering
  - recreates the bandwith reduction which occurs during decimation
  - Eg: decimate from 44100Hz to 800Hz => Low-pass filter with cutoff at 1600Hz

4. Results

Figure 1: F1-score comparison

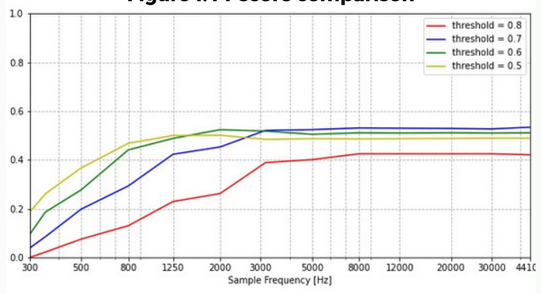


Figure 2: precision comparison

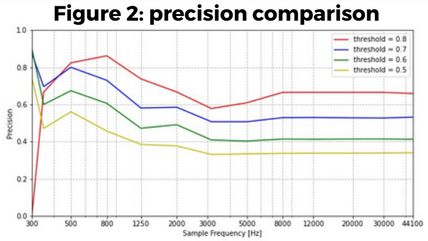
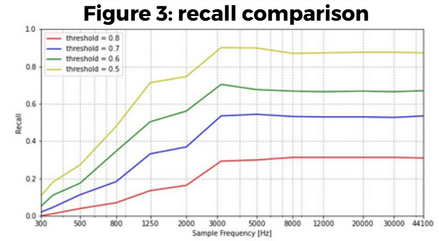


Figure 3: recall comparison



5. Discussion and Conclusions

Sample frequency and performance

- Relatively linear drop in performance from sample frequency **f_d**
- As sample frequency decreases: precision rises and recall decreases:
  - a lower sample frequency results in less laughs being detected, but they are more likely to be TP

Threshold hyper-parameter

- threshold influences:
  - **f_d** (positive relation)
  - precision (positive relation)
  - recall (negative relation)
- A lower threshold result in:
  - higher performance with sample frequencies < **f_d**
  - lower performance with sample frequencies > **f_d**
- Too high a threshold result in overall loss of performance (eg: 0.8)