Revealing Hidden Conversations in Privacy-Sensitive Audio Using Neural Networks

Background

Privacy

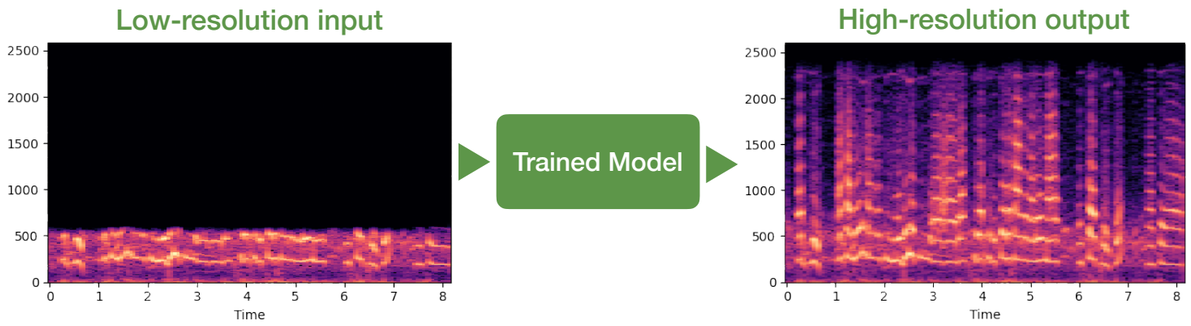
With widespread use of advanced technology for the recording, storing and sharing of social interactions, protecting privacy of people has been a growing concern. This research zooms in on the collection of spoken audio with regard for the privacy of recorded individuals.

Rhythm and MINGLE

Rhythm[1] and MINGLE[2] are two experiments in the area of social signal processing. For both experiments a badge was designed to collect audio — among other types of data — in a privacy aware manner. This was achieved by recording the audio in a low sampling frequency (700Hz and 1.25kHz respectively), such that conversations are obfuscated. The resulting recorded audio is also known as **privacy-sensitive** audio.

Audio Super-Resolution

It still needs to be verified whether the down-sampling technique is actually sound when it comes to hiding conversations in privacy-sensitive audio. To investigate this I have applied a technique called **audio super-resolution** — also known as artificial bandwidth extension — to reconstruct higher frequency components. In theory these reconstructed higher frequency components should increase the intelligibility of the audio and could possibly reveal distinct words in the audio. A visualisation of super-resolution is shown in the figure below.



Research Question

Can existing super-resolution techniques be used to reveal hidden conversations in privacy-sensitive audio?

Method

To verify the main research question I have used a model designed by Kuleshov et al.[3] based on a neural network with an auto-encoder-like architecture. The model was trained on the **LaRed** dataset and verified using the remaining part of this dataset as well as the **ConfLab** dataset. The verification done using a range of objective metrics Signal-to-Noise Ratio (SNR) Log Spectral Distance (LSD) to compare models based on the time and frequency domain audio features respectively. More importantly, to verify speech quality the PESQ and Word Error Rate (WER) metrics were used.

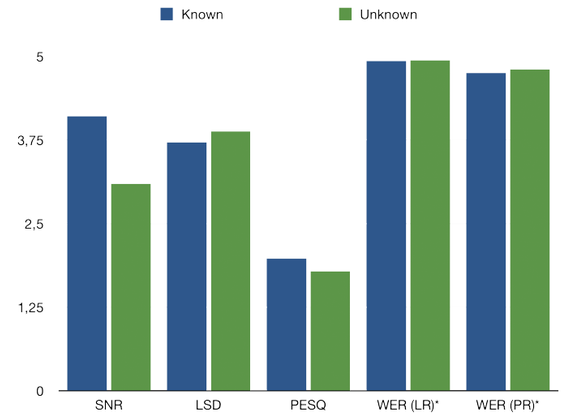
The PESQ measure emulates a subjective Mean Opinion Score (MOS) using proprietary software[4], where a 1 corresponds with poor audio quality and 5 with good quality. WER is used to compare transcriptions generated using the predicted output with a reference transcription generated on higher resolution 8kHz versions of the same audio. This gives an indication of speech intelligibility.

The **LaRed** dataset contains high-resolution (44.1kHz) audio and its audio was used to source 5 different sets (300, 550, 800, 1250 and 2000Hz sample rate) for training and validating the neural network. The low-resolution (1.25kHz) **ConfLab** dataset was used for a final informal validation of the model.

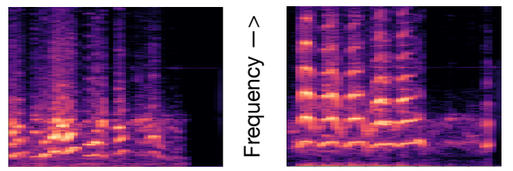
Results

Sample rate	PESQ	WER (LR)	WER (PR)
300Hz	1.56	0.999	0.993
550Hz	1.59	0.999	0.972
800Hz	1.73	0.995	0.962
1250Hz	2.33	0.972	0.929
2000Hz	2.67	0.967	0.896

Sample rate	PESQ	WER (LR)	WER (PR)
300Hz	1.30	1.00	0.989
550Hz	1.25	0.999	0.976
800Hz	1.51	0.998	0.981
1250Hz	2.13	0.978	0.943
2000Hz	2.73	0.967	0.917



The **top left** table displays validation results based on audio from speakers that were present in the training data. The **top right** table displays validation results based on audio from speakers not present in the training data. For SNR and PESQ, higher is better. For LSD and WER, lower is better. Below is a comparison between predicted ConfLab (left) and LaRed (right) audio.



Conclusion

Based on the WER results it can be concluded that an improvement in intelligibility has been achieved on audio similar to the training audio. Similar, though slightly more modest improvements are seen on data from an unknown speaker. Though these results are promising, not enough words are revealed to reveal significant parts of conversations. Based on the ConfLab results, the model seems to perform poorly on out-of-distribution audio.

Recommendations

Better results can likely be achieved by training on more data and audio with more varying noise profiles and with speech from different languages. There are also more recent super-resolution models.

[1] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, “Rhythm: A unified measurement platform for human organizations”, IEEE MultiMedia, vol. 25, no. 1, pp. 26–38, Jan. 2018.
[2] “ConfLab - ACM MM 2019”. (2019), Available: https://conflab.ewi.tudelft.nl/ (visited on 04/26/2022).
[3] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks”, Aug. 2, 2017.
[4] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”, in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), vol. 2, Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

Pepijn Vunderink — CSE 3000 - Research Project — Supervised by H. Hung and J. Vargas Quiros