Visualizing Automatic Speech Recognition

Karla Markert, Romain Parracone, Philip Sperl, Konstantin Böttinger

Fraunhofer AISEC, Garching, Germany

Attribution Methods

For this work, we have adapted two attribution methods from DeepExplain [1] for the automatic language recognition system DeepSpeech [2]:

- **Saliency mapping**,
- **Layer-wise relevance propagation (LRP)**.

We propose to visualize attributions as Mel-frequency cepstral coefficients (MFCCs).
These new visualizations can be exploited to get a better understanding of both, the automatic speech recognition system DeepSpeech and the applied attribution methods.

Normal Samples

For 360 benign speech examples, we have plotted the MFCC's attributions. The **values represent the contribution of each input feature to the predicted character** indicated below the x axis.
Here, we consider the benign audio file *"it has been mentioned but the article is not mine"*. We present both, frame-wise plots as well as averaged attribution plots.

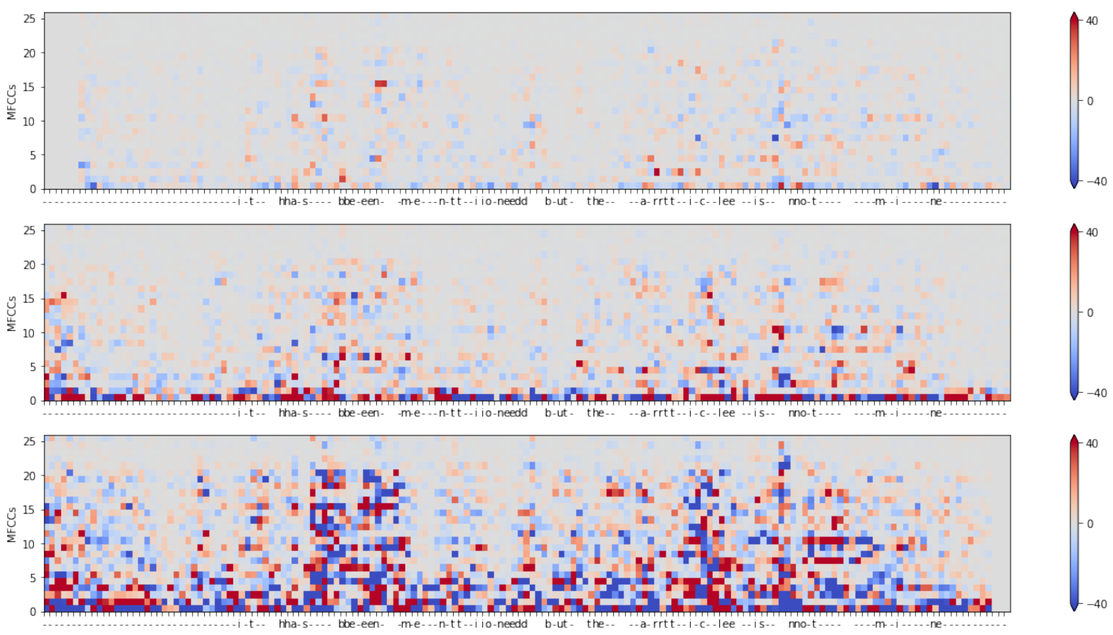

Figure: Attribution computed with LRP, plotted for frames 3, 9 and 12.

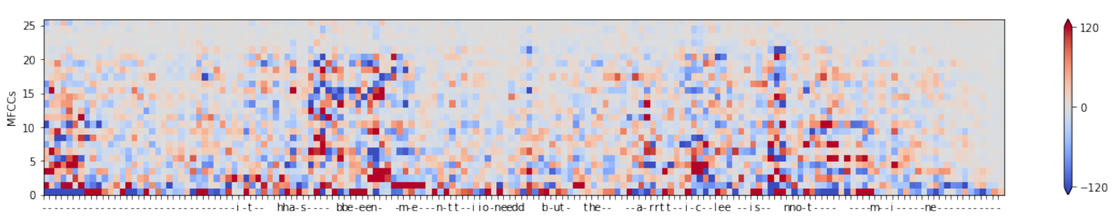

Figure: Averaged attribution computed with LRP.

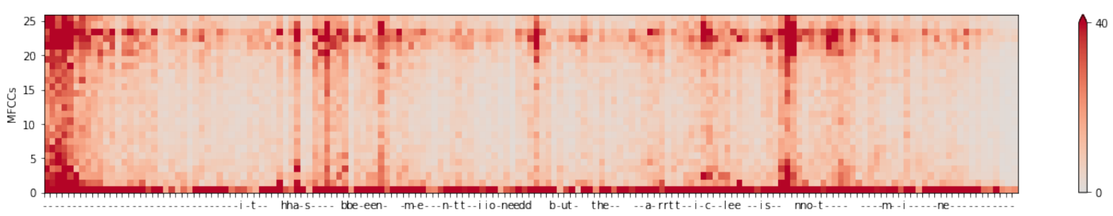

Figure: Averaged attribution computed with saliency maps.

Deepspeech Overview

DeepSpeech is an **open-source speech-to-text engine** developed by Mozilla, based on Baidu's Deep Speech research paper [2] and using Google's TensorFlow. The engine is composed of three stages: one feature extraction stage that computes MFCCs; **a Long Short Term Memory (LSTM) neural network that takes MFCCs as input and outputs character probabilities**; and a language model that turns the NN output into a properly formatted text. In this work, we use DeepSpeech v0.4.1 and analyse attacks against the LSTM.

Adversarial Samples

The attributions were computed using LRP and averaged frames. The adversarial example has stronger attribution values.

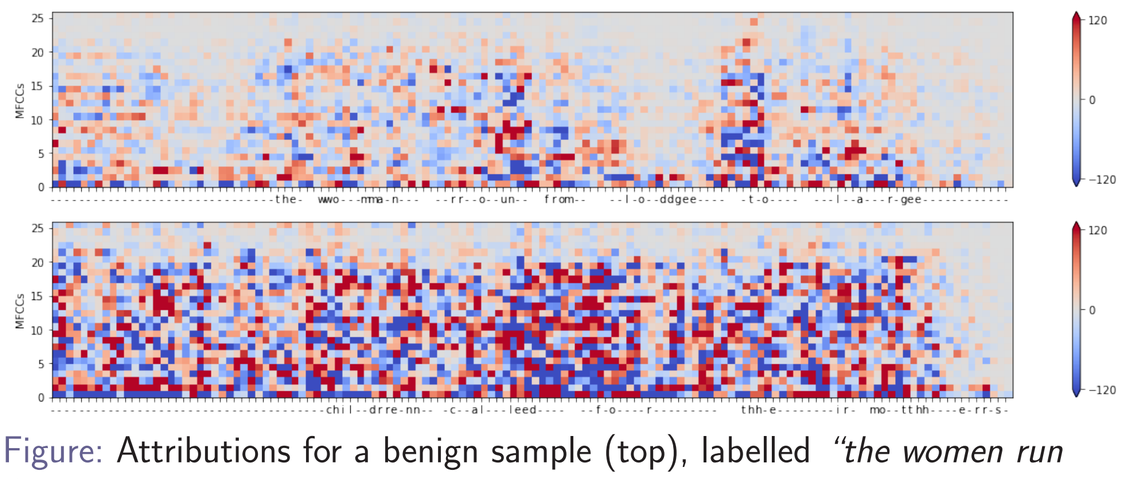

Figure: Attributions for a benign sample (top), labelled *"the women run from lodge to lodge"*, and an adversarial example based on the same benign input (bottom), wrongly labelled *"children called for their mothers"*.

Results

First, in experiments covering 360 benign speech examples, we observed that the same pattern repeats: the magnitude of the attribution in frame 12 and frame 9 is higher than in frame 3. Thus, **frames from the future have a higher impact on the final classification** than frames from the past.
Second, we observe that **the attribution is also unevenly distributed among the different MFCC bins**.
Third, **the visualization of benign samples and adversarial examples shows a striking difference**. Generally speaking, adversarial examples tend to activate the NN to a greater extent:

- The **attribution values are more extreme**, and they are at a high level most of the time.
- Based on our tests on various adversarial examples, we conclude that adversarial examples have a stronger relative effect on parts with lower values (with less activation) in benign samples, namely the first frames and the highest MFCCs.
- This implies that **little changes in certain frequencies have a strong influence on the transcribed text**.

Sources

[1] Marco Ancona et al. "Towards better understanding of gradient-based attribution methods for deep neural networks". In: *arXiv preprint arXiv:1711.06104* (2017).

[2] Awni Hannun et al. "Deep speech: Scaling up end-to-end speech recognition". In: *arXiv preprint arXiv:1412.5567* (2014).