FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2018

# Detection of Overlapping Speech using a Convolutional Neural Network: First Experiments

Marie Kunešová[1]

## 1 Introduction

Many speech processing applications, such as speaker diarization and speech recognition, have problems with overlapping speech, i.e. intervals in which multiple speakers are talking simultaneously. This happens particularly often in spontaneous conversations, where speakers may regularly interrupt each other or interject short utterances while the original speaker keeps talking. Detecting such occurrences can help improve the performance of the impacted systems. However, this is still an actively researched task, which has not yet been fully solved.

In this work, I describe my initial experiments in using a convolutional neural network (CNN) to detect overlapping speech in an artificial dataset created for the purpose.

## 2 The Convolutional Neural Network

The main idea of this work was inspired by Hrúz and Zajíc (2017), who used a convolutional neural network for the detection of speaker changes. My initial experiments use the same network architecture and general approach as described in the referenced paper.

The input of the network is a spectrogram of a short window of the acoustic signal. The output of the last layer is a value between 0 and 1, indicating the probability of overlapping speech in the middle of the window. Training references use a fuzzy labeling function, with a linear slope at the boundaries between overlap and non-overlap (see Figure 1 for an example).

## 3 Data

Training a neural network for overlap detection requires a large amount of well-annotated data with frequent overlaps. Unfortunately, there do not appear to be any publicly available datasets made for this purpose, and other corpora often lack sufficiently precise labels.

Instead, I have created artificial data from the LibriSpeech corpus (Panayotov et al.; 2015), which is a large database of single-speaker utterances sourced from audiobooks.

The artificial data were generated in the following way: First, individual short utterances (each about 10-15 seconds long) from a single speaker are joined into a single long file, with random 5-10 s pauses in-between (e.g. the last two plots in Figure 1). Two such files are then merged together at different volumes and with added background noise. Reference labels were obtained by using a voice activity detector on the original files.

This results in data with a large percentage of both overlapped speech and single speaker regions: overall, the recordings contain approximately 50% single speaker speech, 25% overlapped speech and 25% silence.

[1] student of the doctoral study programme Applied Sciences and Computer Engineering, field Cybernetics, e-mail: mkunes@kky.zcu.cz

For the training of the CNN, I used 100 such generated files, each 10 minutes long. Test data were obtained in the same manner, from speakers not present in the training set.

## 4 Results

First results on test data appear very promising - for most audio files, the output of the CNN corresponds very well with the actual overlaps, as illustrated in Figure 1. Unfortunately, performance deteriorates on recordings of lower quality, particularly in the presence of reverberation (which gives the false impression of a second speaker where only one is present).

Further work now focuses on improving the performance on lower-quality speech by generating more varied training data, and on applying the CNN to real conversations.

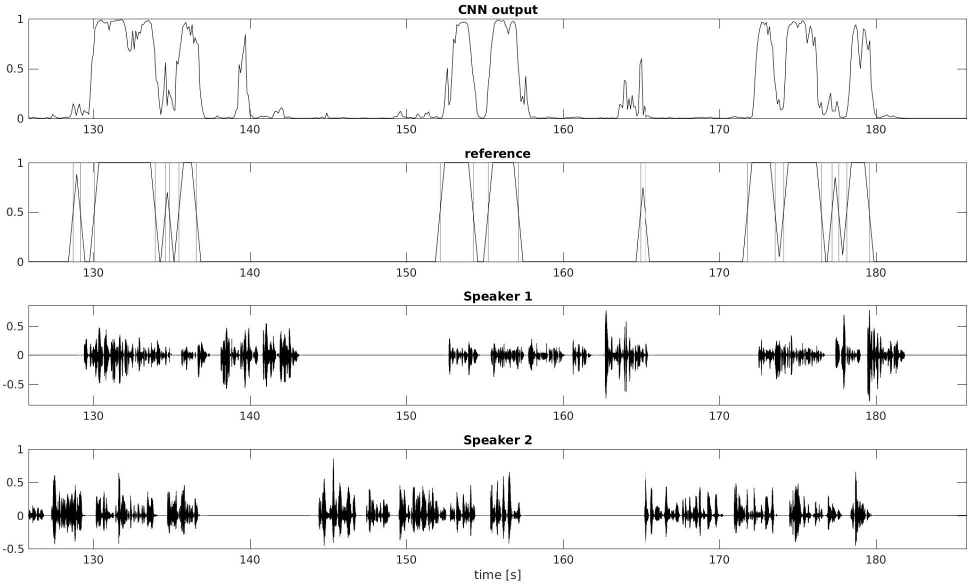

**Figure 1:** Example of the CNN's output for unseen speakers (top), the corresponding reference labels and each speaker's soundwave.

### Acknowledgements

This work is supported by grant of the University of West Bohemia, project No. SGS2016-039. I would like to thank M. Hrúz for sharing his original scripts for the CNN used in the experiments.

## References

Hrúz, M. and Zajíc, Z. (2017). Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949.

Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015). LibriSpeech: An ASR Corpus Based on Public Domain Audio Books, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210.