Zurich University
of Applied Sciences

School of Engineering

Speaker identification and clustering using convolutional neural networks [1]

Yanick Lukic, Carlo Vogt, Oliver Dürr, and Thilo Stadelmann
ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland

Problem Statement

- Speaker recognition performance by machines << by humans [2]
- Evidence: Clustering performance orders of magnitude lower than identification performance

➔Improve core speaker recognition performance on both tasks by
…using **learned features** instead of handcrafted
…an approach capable of **capturing sequence information**
…free **from complicating side effects** of application scenarios

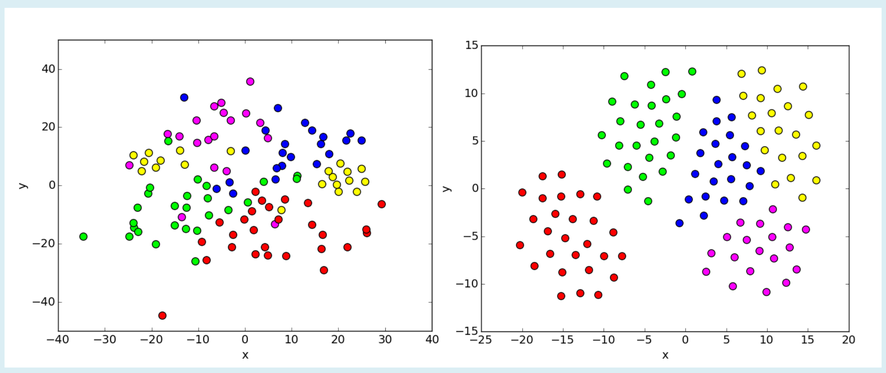
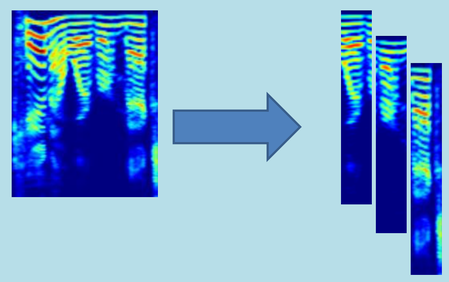


Figure 1: t-SNE plots based on the output vectors of the softmax layer L8 (left) and the first dense layer L5 (right). Different colors correspond to different speakers.

Approach

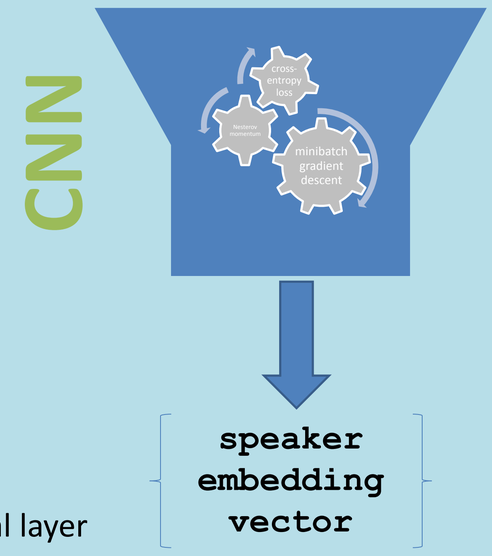
Feature Learning (identification training)
Form mini batches by taking
- …128 random snippets among all training utterances
- …1 second long, from spectrograms (no overlap)
- …train to classify by speaker



Application (clustering test)
First
- train identification CNN as above…
- for a large number of speakers (unrelated to the ones to cluster)

Second for each unknown utterance…
- chop into 1s non-overlapping segments
- put through CNN
- take output of specific post-convolutional layer (see below) as a “speaker embedding”
- Average all embeddings per utterance
  ➔ speaker-specific feature vector per utterance

Third perform agglomerative hierarchical clustering on embedding vectors

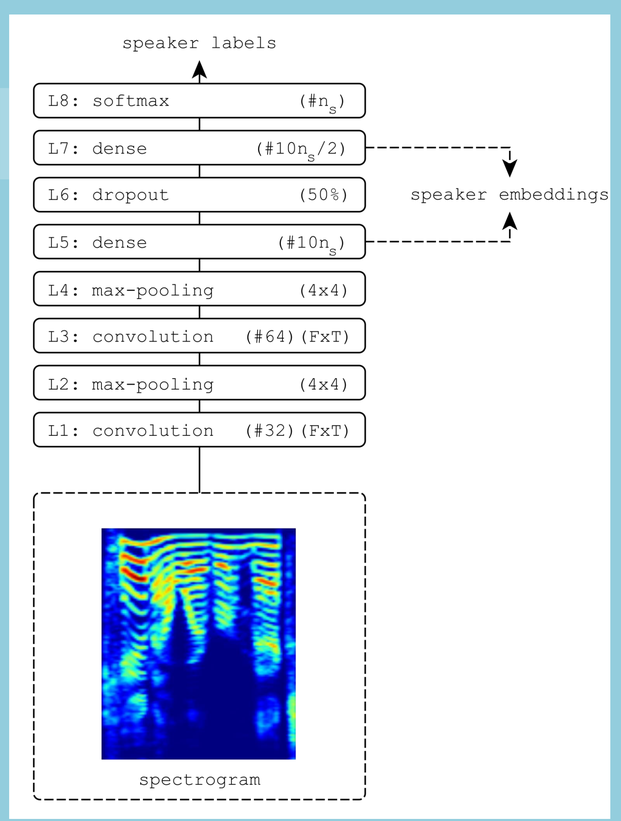


Experimental Setup

- Closely follows [2] on TIMIT
- Architecture of CNN from [3]
- Evaluation by miss-classification rate MR

- Implementation in python using `lasagne` and `librosa` for spectrograms

- **Pre-trained CNN for clustering hasn’t seen any utterance to cluster during training!**



Results

- Identification accuracy on TIMIT test set (630 speakers): 97.0%
- Clustering MR on 40 speakers [2]: 0.05 (see Tab. 1)
- Embeddings on the lower-level post-convolutional layers work considerably better (see Fig. 1 above)

	MR 100	MR 590
L5: dense	0.300	0.125
L7: dense	0.325	**0.050**
L8: softmax	0.700	0.450

Table 1: MR for clustering 40 speakers from TIMIT test using embeddings from different post-convolutional layers of the identification CNN (having been trained with either 100 or 590 different speakers).

➔ **Clustering compares favorably with SotA** (MR 0.065) and GMM-MFCC approach (MR 0.125) **without any task-specific pre-processing or model**

References

[1] Y. Lukic, C. Vogt, O. Duerr, and T. Stadelmann, *Speaker Identification and Clustering using Convolutional Neural Networks*. In Proceedings of IEEE MLSP 2016.
[2] T. Stadelmann and B. Freisleben, *Unfolding speaker clustering potential: a biomimetic approach.* In Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009, pp. 185–194.
[3] S. Dielemann and B. Schrauwen, End-to-end learning for music audio. In Proceedings of ICASSP 2014, pp. 6964-6968.

This work has been supported by the Bachelor’s programme IT, the Institute of Applied Information Technology InIT, and the Institute of Data Analysis and Process Design IDP at ZHAW. Thank you!