Evaluation of data augmentation for speaker recognition with convolutional neural network

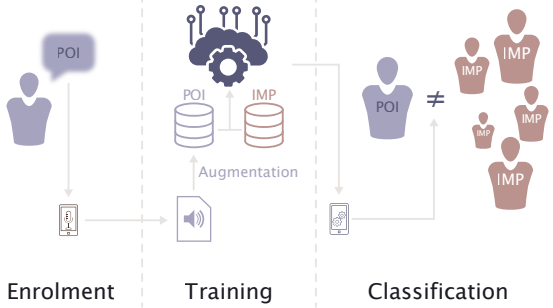
Degree programme: MAS Data Science

Performance of automatic speaker recognition mainly relies on the amount and variability of training data. Some applications might be impaired by the limited amount of training data when confronted with text-independent speech utterances in various acoustical environments. Speaker verification, based on different deep learning approaches, was evaluated for this project with data augmentation applied to a single speech utterance used for the training phase.

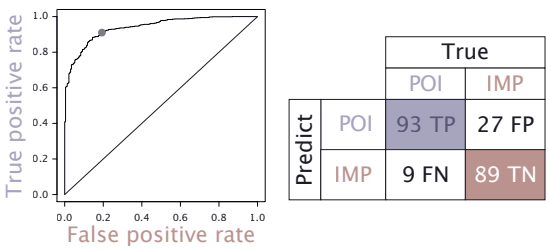
Anybody can easily recognize a given speaker. This human trait is responsive to changes in speech production, sound reproduction systems, and listening environments. Speaker recognition applications are designed to automatically reproduce this trait. While it has made significant progress with accessible deep learning techniques, it is easily challenged by text-independent applications, degraded acoustical conditions, and limited training data.
The defined use case is to train a classifier with a single speech utterance from the target speaker. A training dataset was generated with data augmentation to cover a wide range of potential application conditions. Various network architectures were trained and finally evaluated with a ROC curve and Matthews correlation coefficient. The single fully connected layer network, one of the simplest tested architectures, was finally selected.

Beside data augmentation, many factors were identified to influence the results, like the dataset quality and size, the extracted features, the decision threshold optimization, and the evaluation measures.
These aspects were discussed in order to explain the achieved results and to evaluate the generalization potential. Interpreting each effect independently is challenging as many interactions are linked to the variation of a single parameter.
These first results show the potential of data augmentation for a speaker recognition model starting with a single speech utterance in the training set. It reflects the power of deep learning associated with data augmentation to learn complex patterns in the available data. The discussion highlights the susceptibility to add variations in the results with minor changes in the available data, during the feature extraction, or in the selected network.

Christophe Lesimple



Process for text-independent discrimination between the person of interest (POI) and different impostors (IMP)



		True	
		POI	IMP
Predict	POI	93 TP	27 FP
	IMP	9 FN	89 TN

Confusion matrix after decision threshold optimization with person of interest (POI) and impostors (IMP)