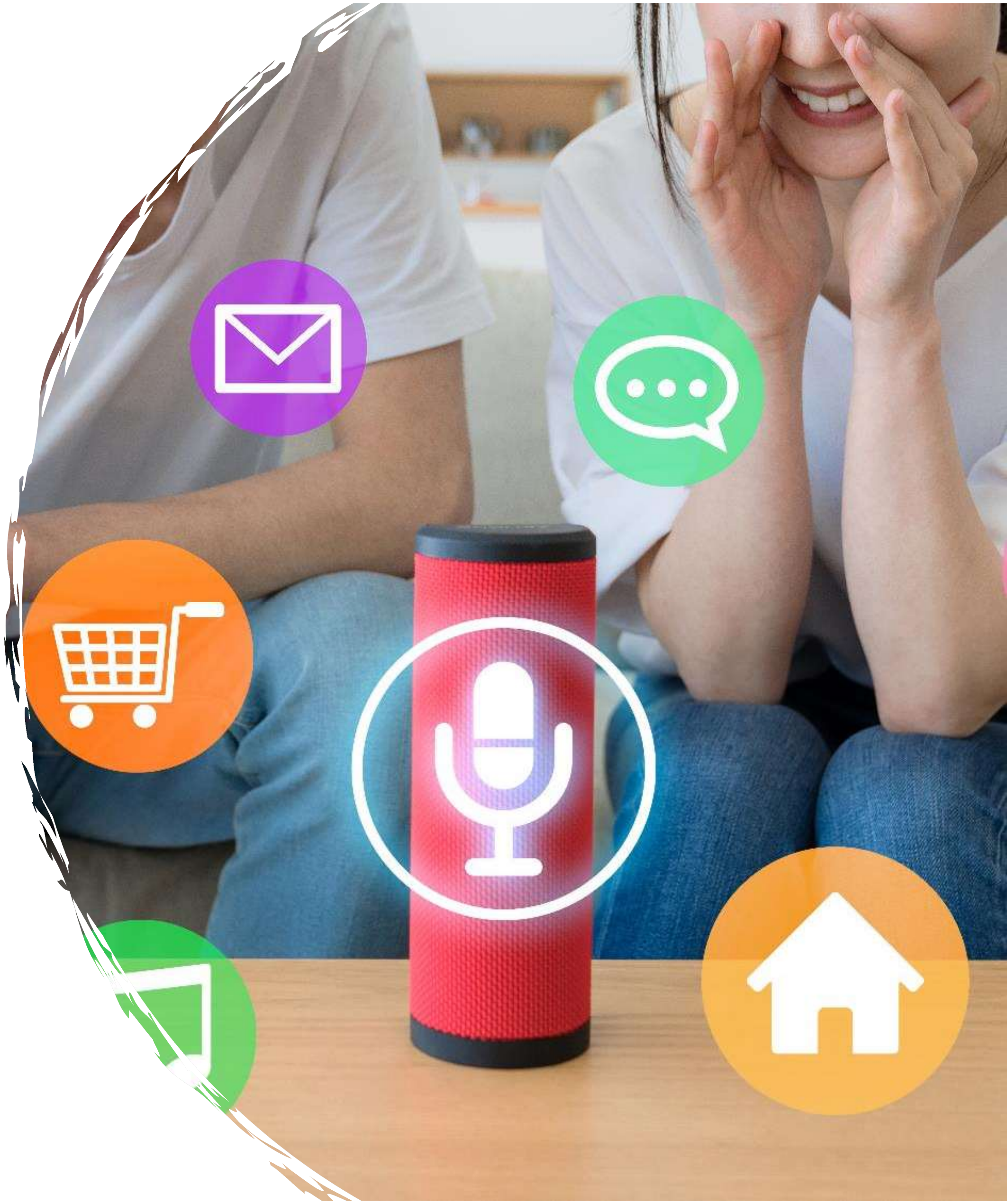
STQ Workshop

Dependence of EER and MOS in the context of automatic speaker recognition and adverse conditions

Maroš Jakubec

22/11/2022



Content

- Introduction
- Speaker embeddings
- Proposed DNN architecture and experiment setup
- Results
- Conclusion

Introduction

Speaker Recognition:

- Humans have the innate ability to recognize familiar voices within seconds of hearing a person speak.
  - How do we teach a machine to do the same?

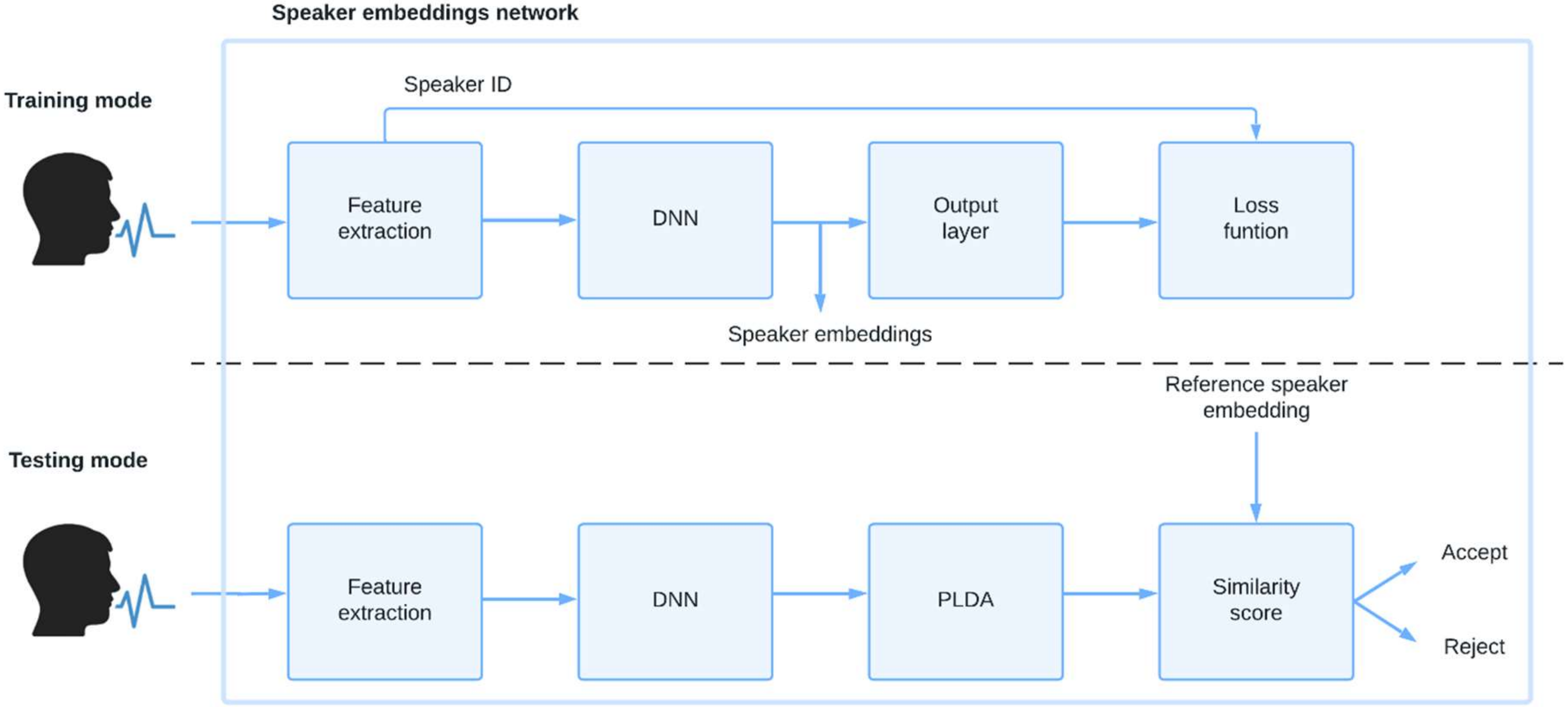
Biometrics based on voice recognition:

- Harder to fake than other forms of authentication
- Contactless login
- Accessible and convenient on a variety of devices

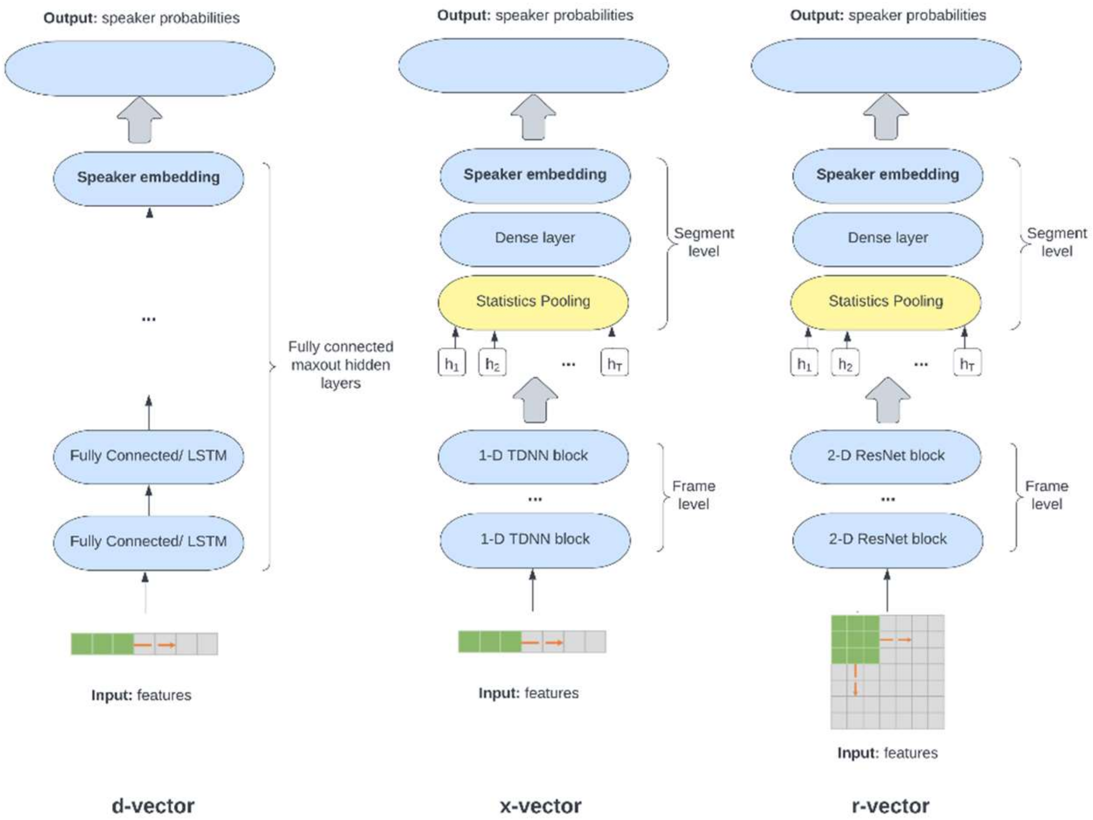
- Tremendous application spike in the field of DNN, including increasing interest in the development of speaker recognition systems.

- DNN-based speaker embeddings, such as x-vectors or d-vectors, have begun to replace standard i-vectors based on factor analysis.

Overview of speaker embedding-based speaker verification system



Comparison of DNN architectures based on speaker embeddings



Proposed speaker embedding

each utterance or recording is compressed into a unique **embedding** or a **„voiceprint"** of the same length. This „voiceprint" becomes a high-level feature for further classification.

Our proposal is based on the basic **x-vector embedding**

- Time-Delayed Neural Network (TDNN)
- fixed-length embeddings or features are extracted from the layers located after the pooling layer.

We modified the system topology by including components of the popular ResNet architecture (denoted as **r-vectors**)

- Res2Net, a novel building component for CNNs that seeks to enhance multi-scale representation by expanding the number of possible receptive fields.
- Squeeze excitation (SE) block

The configurations of the proposed network

- Our design is fully implemented in Pytorch

TABLE SPEAKER EMBEDDING ARCHITECTURE BASED ON RES2NET

Layer Name	Module	Output Size
Input	—	$80 \times T \times 1$
Conv2D-1	$(3 \times 3, 2)$	$80 \times T/2 \times 64$
SE-Res2NetBlock-1	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 2$	$40 \times T/2 \times 64$
SE-Res2NetBlock-2	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 2$	$40 \times T/4 \times 128$
SE-Res2NetBlock-3	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 2$	$20 \times T/4 \times 128$
SE-Res2NetBlock-4	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 2$	$10 \times T/8 \times 256$
MHA Pooling Layer	—	1×256
Dense-ReLU (r-vector)	—	256
AM-Softmax	—	N

- Multi-head Attention (MHA)
- Additive Marginal Softmax (AM-Softmax)

Experimental setup

Environments – Python:

- Librabry: Pytorch, Librosa

Datasets:

- experiments were performed on the VoxCeleb1
- consisting of short videos extracted from videos uploaded to YouTube.

Working environment:

- Ubuntu 20.04 LTS
- Pycharm Community

PC:

- Intel® Core ™ i9-7900X
- NVIDIA GeForce 980ti

TABLE I. VOXCELEB1 DATASETS DETAILS

Dataset	#	*Dev*	*Test*	*Total*
VoxCeleb1	*POIs*	1 211	40	1 251
	utterances	148 642	4 874	153 516
	hours	~ 335h	~ 17h	352

Experiment setup

Acoustic Features:

- 80-dimensional FB (logarithm of the signal energies in the frequency sub-bands)
- 25ms duration and 10ms shift.
- mean normalization

DNN Setup:

- trained on 20 epochs with a batch size of 128.
- SGD optimizer together with $\beta 1 = 0.9$, $\beta 2 = 0.98$, $\varepsilon = 10-9$.
- learning rate 0.01

Evaluation Metrics:

- EER (Equal Error Rate)
- MinDCF (Minimum Normalized Decision Cost Function)

Experiment results

- We investigated the effectiveness of the proposed system based on the speaker embeddings extracted using deep CNN.
- Aside from the accuracy of the speaker embedding system, we took into account the computing needs as well.

TABLE COMPARISON OF SPEAKER EMBEDDING RESULTS WITH THE PROPOSED MODEL.

	#Parameters	Training Time	EER (%)	DCF10^{-2}
x-vector (baseline)	8.5M	~ 30h	4.22	0.4011
ResNet-34	9M	~ 32h	3.18	0.2768
SE-Res2Net	8.5M	~ 35h	2.71	0.2482

NISQA: Speech Quality and Naturalness Assessment

- is a deep learning model/framework for speech quality prediction
  - focused on distortions occurring in communication networks.

- besides overall speech quality, NISQA also provides predictions for the quality dimensions **Noisiness, Coloration, Discontinuity, and Loudness.**

- The NISQA Corpus includes **more than 14,000 speech samples** with simulated (e.g. codecs, packet-loss, background noise) and live (e.g. mobile phone, Zoom, Skype, WhatsApp) conditions.

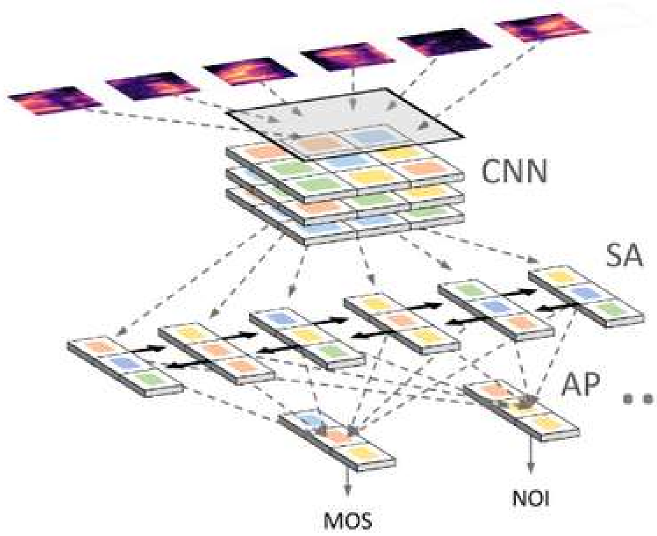
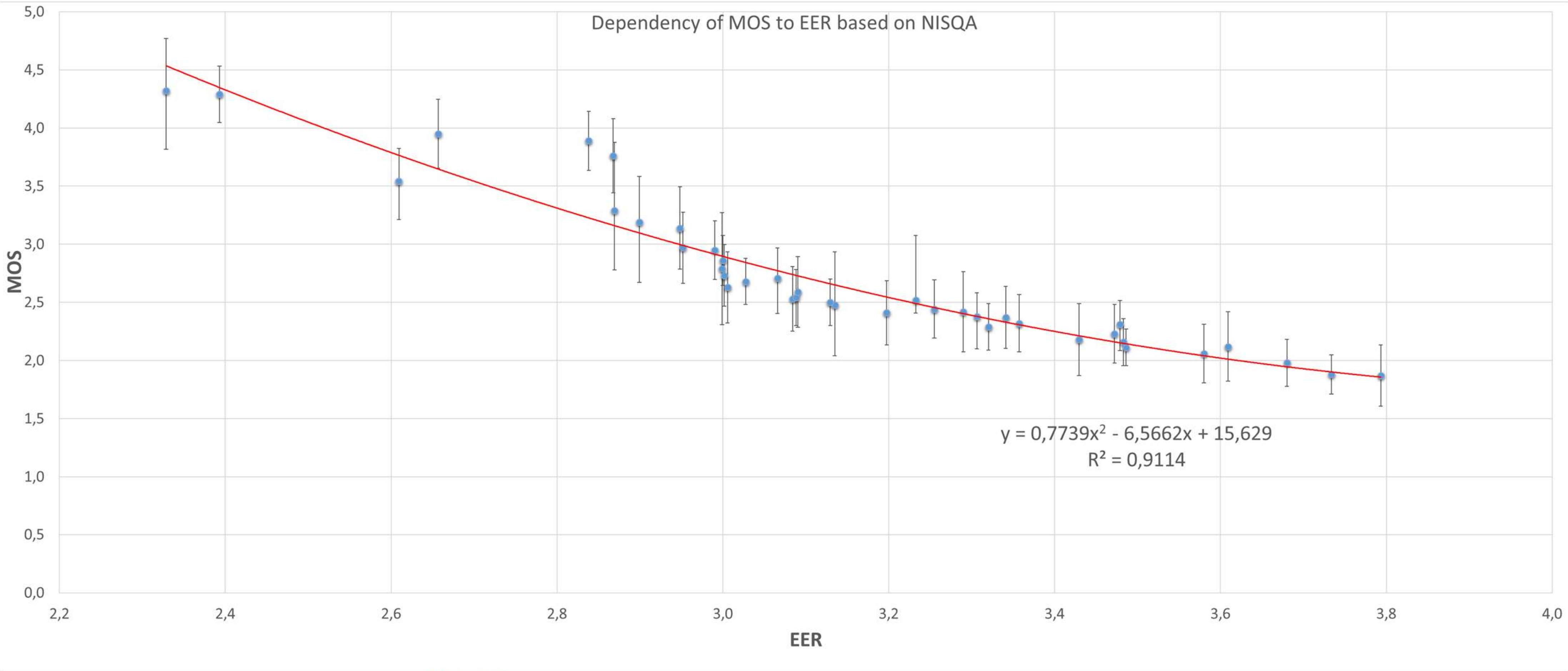


Figure : *NISQA neural network architecture.*

Dependency between the accuracy of the designed speaker voiceprints and the quality of the speech recordings.



Conclusion

- In this work, we investigated the effectiveness of systems based on the r-vector embedding of the speaker's voice, which was obtained using CNN.
- The deep residual based CNNs represent the best option for speaker embeddings learning and can provide robustness in SV tasks.
- The NISQA model can be used to tune the transmission system
  - With the help of quality assessment, we retrospectively determined how the speaker recognition system behaved in real conditions.
  - The system can be tuned and optimized to cope with real/adverse conditions according to obtained characteristics.

Thank you for your attention!