DEEP CNN BASED FEATURE EXTRACTION FOR TEXT-PROMPTED SPEAKER RECOGNITION

Speech Technology Center

ITMO UNIVERSITY

Sergey Novoselov[1,2], Oleg Kudashev[2], Vadim Shchemelinin[1] Ivan Kremnev[3], Galina Lavrentyeva[1]

[1]ITMO University, St. Petersburg, Russia; [2]STC-innovations Ltd., St. Petersubrg, Russia; [3]STC Ltd., St. Petersubrg, Russia

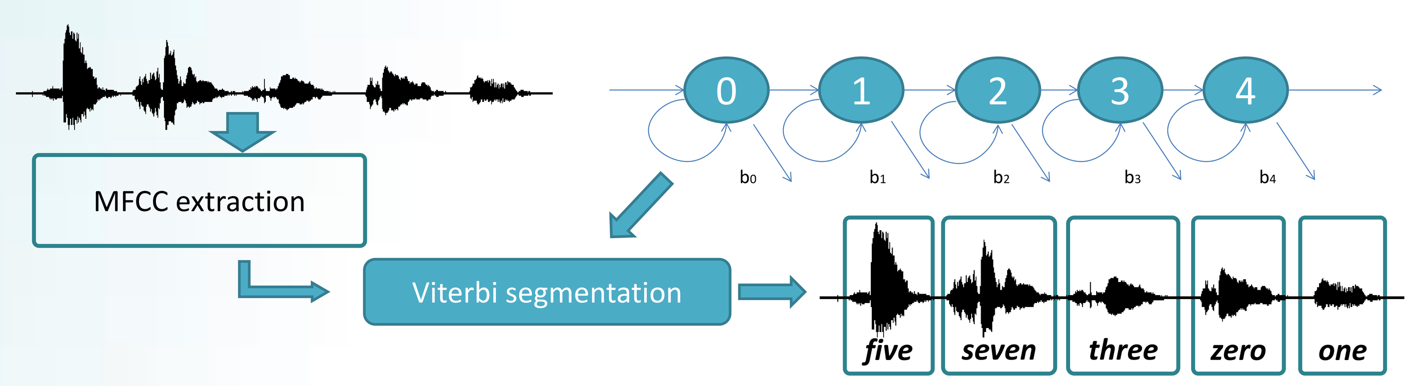
{novoselov, kudashev, shchemelinin, kremnev, lavrentyeva}@speechpro.com

1 Introduction

- **Text-dependent** speaker recognition task [1,2,3,4] is studied
- **Deep convolutional neural network** based speaker specific features extractor in the text-prompted speaker verification task **is presented**
- The prompted **passphrase is segmented into word states** —i.e. digits — to test each digit utterance separately
- A **single high-level feature extractor for all states** is used and cosine similarity metric is applied for scoring
- **Multitask learning scheme** is used to train the high-level feature extractor

2 Features

Viterbi segmentation to word states



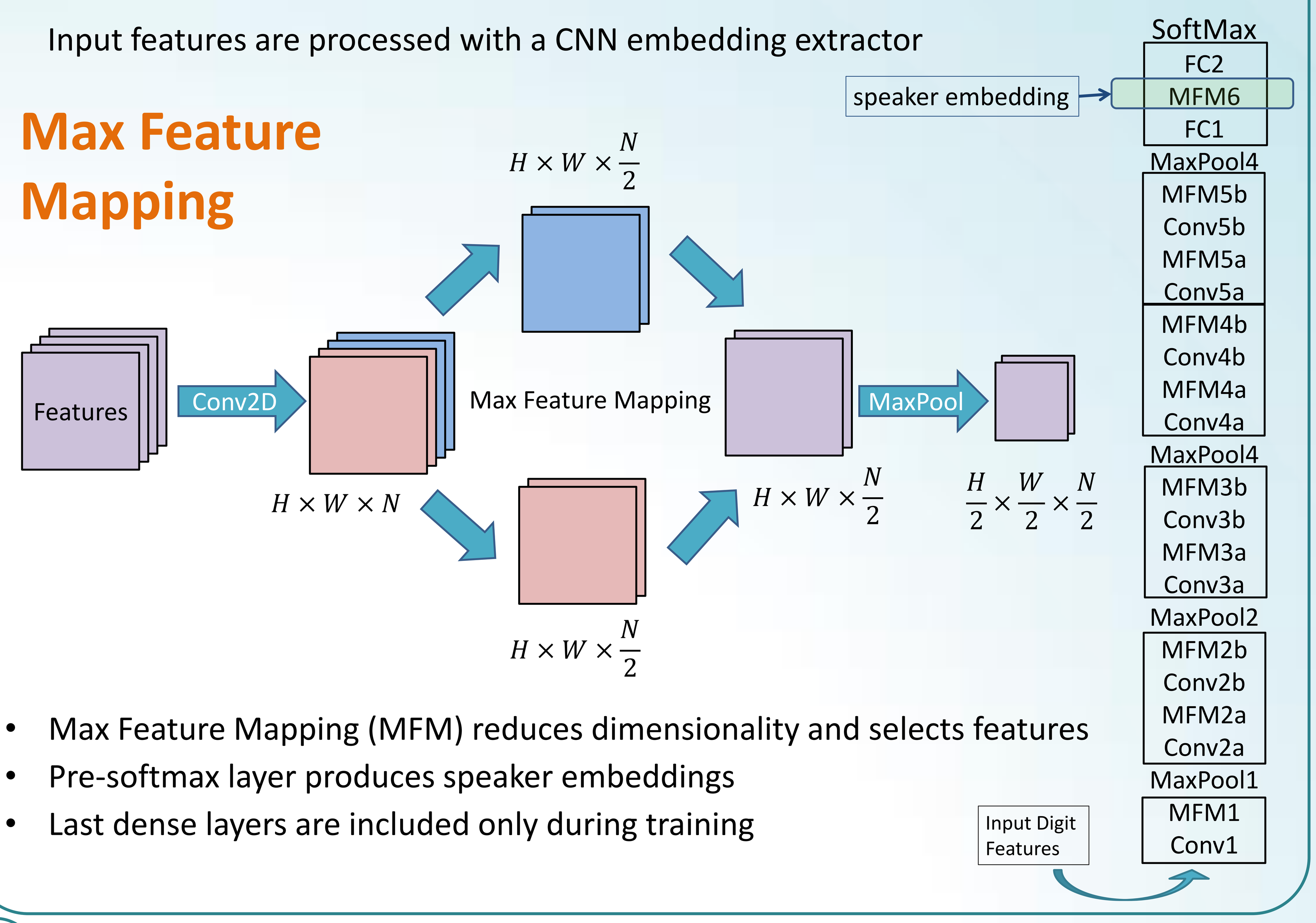
Input features for the CNN are 64×96 log mel power spectra:

- 64 frequency bands
- 96 frames (longest single digit utterance)
- Voice activity detector removes non-speech frames

3 Convolutional Neural Network

Input features are processed with a CNN embedding extractor

Max Feature Mapping

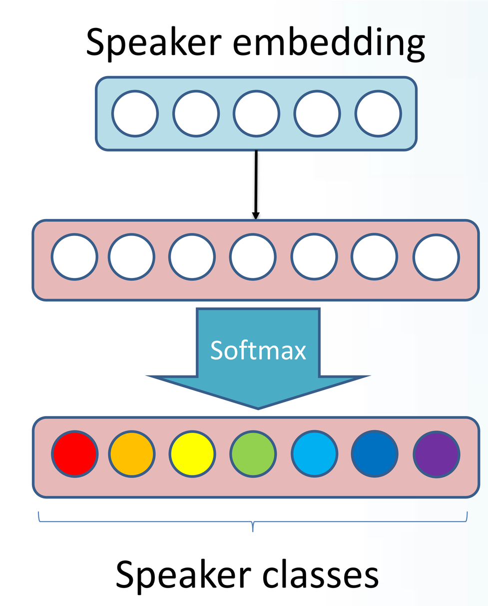


- Max Feature Mapping (MFM) reduces dimensionality and selects features
- Pre-softmax layer produces speaker embeddings
- Last dense layers are included only during training

4 Learning mode

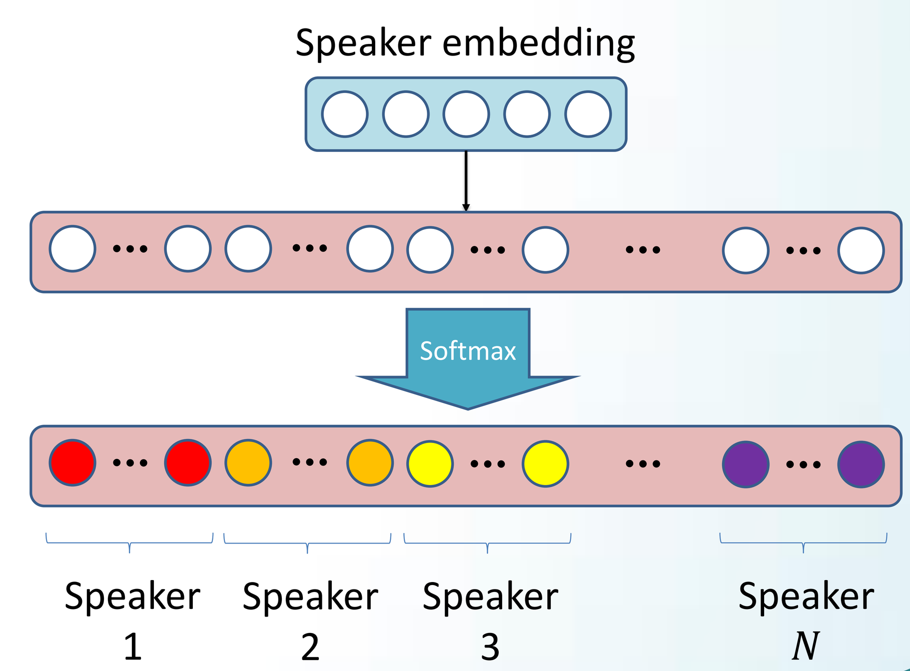
Single-task

- Extractor is trained to discriminate speakers
- $N_{speakers}$ neurons at softmax layer



Multi-task

- Extractor is trained to discriminate speakers and word states
- $N_{speakers} \times N_{digits}$ neurons at softmax layer



5 Experiments

We explored 5-digit password verification scenario when the speaker pronounces the correct passphrase. Training/evaluation bases consist of short digit passphrases

Training Datasets:

- **RSR2015** [1] **Part 3 train set** : 194 speakers (94 Female + 100 Male) - $RSR2015_{tr}$
- **Wells Fargo Bank set**: 300 speakers (150 Female + 150 Male) – WF
- **STC-Russian-digits train set**: 786 speakers (263 Female + 523 Male) – $STCRus_{tr}$

Evaluation Datasets:

- **RSR2015 Part 3 eval set** : 106 speakers (49 Female + 57 Male) - $RSR2015_{ev}$
- **STC-Russian-digits eval set**: 92 speakers (42 Female + 50 Male) - $STCRus_{ev}$

Results

Table 1. EER [%] and minDCF ($C_{miss} = 10, C_{fa} = 1, P_{tar} = 10^{-2}$) for 5-digit password verification

System	Multi-Task mode	Training data	Evaluation data	EER (%)	Min DCF
Baseline State-GMM-SVM[2]	None	$RSR2015_{tr}$ +WF	$RSR2015_{ev}$	3.11	0.14
State-CNN	None	$RSR2015_{tr}$		7.83	0.39
	Speaker & Digits			5.12	0.25
		$RSR2015_{tr}$ +WF		4.27	0.2
		$STCRus_{tr}$	$STCRus_{ev}$	5.86	0.29
	Speaker & Digits & Language	$RSR2015_{tr}$ +WF+$STCRus_{tr}$	$RSR2015_{ev}$	**2.85**	**0.13**
			$STCRus_{ev}$	**4.24**	**20.45**

Fusion results

Systems description:

State-GMM-SVM [2]:
Viterbi segmentation, state supervector extraction, state SVM based scoring, S-norm

State-GMM-PLDA [3]:
Viterbi segmentation, state supervector extraction, state TV space transform, state PLDA scoring

State-CNN:
Viterbi segmentation, state CNN deep speaker embedding extraction, cosine based scoring

Table 2. Fusion. EER [%] and minDCF for 5-digit password verification

System	EER (%)	Min DCF
State-CNN + StatePLDA	2.09	0.1
State-CNN + State-GMM-SVM	1.63	0.07
State-CNN + State-GMM-SVM	1.57	0.08
All	1.43	0.07

6 Conclusions

- A deep CNN based speaker feature extractor for speech digits is presented
- Multitask learning mode allows to train effective high-level speaker embeddings extractor for all states (digits)
- Discriminatively trained deep CNN based solution is able to surpass the classic baseline systems in terms of quality
- No complex trainable backend is needed for scoring. Speaker embeddings can be compared simply with cosine similarity metric
- CNN-based method fuses well with our previous methods [2,3]

7 References

1. A. Larcher, Kong A. Lee, B. Ma, and H.Li, "RSR2015: Database for text-dependent speaker verification using multiple pass-phrases," in *Thirteenth Annual Conference of the ISCA*, 2012
2. S. Novoselov, T. Pekhovsky, A. Shulipa, and A. Sholokhov, "Text-dependent GMM-JFA system for password based speaker verification," in *2014 IEEE ICASSP*. IEEE, 2014, pp. 729–737.
3. S. Novoselov, T. Pekhovsky, A. Shulipa, and O. Kudashev, "PLDA-based system for text-prompted password speaker verification," in AVSS, 2015 12th IEEE International Conference on, pp. 1–5.
4. H. Zeinali, L.Burget, H. Sameti, O. Glembek, and O. Plchot, "Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification," in Odyssey-2016, pp. 24–30.

8 Acknowledgements

This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.578.21.0189 (ID RFMEFI57816X0189).