Hybrid Network For End-To-End Text-Independent Speaker Identification

Wajdi GHEZAIEL, Luc BRUN, Olivier LÉZORAY

Normandie Univ,ENSICAEN,UNICAEN, CNRS GREYC NormaSTIC

Abstract

- We propose a speaker identification system with limited data. Short utterances with few training examples are used.
- A end-to-end hybrid architecture combining convolutional neural network (CNN) and Wavelet Scattering Transform (WST) for text-independent speaker identification is proposed. WST is used as a fixed initialization of the first layers of a CNN network.
- Results: Experiments are conducted on Timit and Librispeech datasets. Our hybrid architecture provides satisfactory results under the constraints of short and limited number of utterances.

Material and Methods

- The wavelet scattering transform (WST) [1], is a deep representation, obtained by iterative application of the wavelet transform modulus. It has been defined so as to be invariant to translations of the input signal, and stable to small deformations.

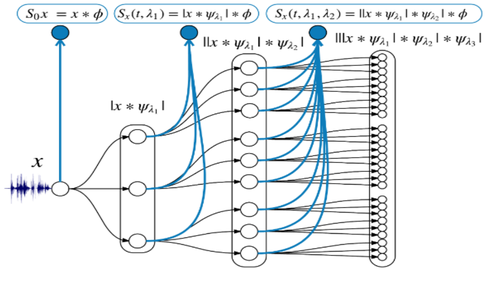

Figure 1:Hierarchical representation of wavelet scattering coefficients at multiple layers [1].

- The proposed hybrid network is composed of a WST for feature extraction and a convolution neural network CNN for classification in the back end.

Architecture

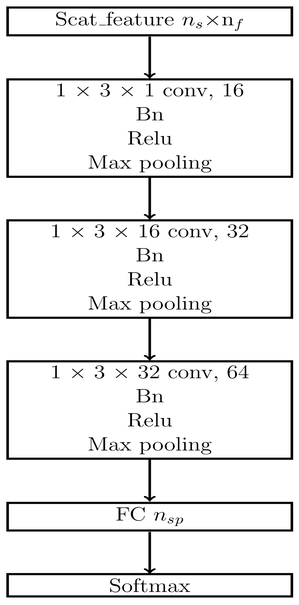

Figure 2:Proposed Hybrid Network.

Experiment & Results

- Experiments on TIMIT [2] and LibriSpeech [3].
- 462 speakers from TIMIT. 5 sentences for training (15s in total) and 3 sentences for testing.
- 2484 speakers from LibriSpeech database. 7 utterances for training (12-15s in total), and 3 utterances for testing.
- Experiments are conducted with longer and shorter raw waveforms.

- Comparaison with SincNet [4], CNN-Raw [5].

	LibriSpeech	TIMIT
CNN-raw	98.91	98.62
SincNet-raw	98.93	**99.13**
HWSTCNN	**99.28**	98.12

Table 1:Identification accuracy rate (%) of the proposed HWSTCNN and related systems trained and tested with full utterances.

- Effect of training and testing utterances duration per speaker on performances:

	Train utterance duration		
Test	**8s**	**12s**	**full**
1.5s	96.86	97.20	97.38
3s	98.76	98.93	98.97
full	99.12	99.25	99.28

Table 2:Identification accuracy rate (%) of the proposed HWSTCNN on LibriSpeech dataset trained and tested with different utterances durations.

- Effect of short utterance duration on HWSTCNN , SincNet [4] and CNN-Raw [5].

	SincNet-raw	CNN-raw	HWSTCNN
1.5s-full	91.51	94.28	**97.38**
3s-full	97.57	96.87	**98.97**

Table 3:Identification accuracy rate (%) of the proposed HWSTCNN and related systems trained on LibriSpeech dataset and tested with different utterances durations.

Conclusion & Future Work

- Effectiveness of this hybrid architecture with limited data.
- Significant improvements over SincNet, CNN-Raw.
- Ability to reduce the required depth and spatial dimension of the deep learning networks.
- Future works: Evaluate HWSTCNN on Voxceleb.

Acknowledgements

This work was supported by BPI France, project HomeKeeper: https://home-keeper.io/

References

[1] J. Andén, S. Mallat, "Deep scattering spectrum," IEEE Transactions on Signal Processing, vol. 62, number 16, pp. 4114–4128, 2014.

[2] L. Lamel, and R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proc. of DARPA Speech Recognition Work-shop, 1986.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," Proc. of ICASSP, pp. 5206–5210, 2015.

[4] M. Ravanelli and Y. Bengio, "Speaker Recognition from raw waveform with SincNet," Proc. of SLT, 2018.

[5] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs," Proc. of Interspeech, 2018.