JOINT SPEAKER DIARIZATION AND RECOGNITION USING CONVOLUTION AND RECURRENT NEURAL NETWORKS

Zhihan Zhou, Yichi Zhang & Zhiyao Duan

Department of Electrical and Computer Engineering, University of Rochester

Introduction

Speaker Recognition -> ***absolute identity*** of a given utterance
Speaker Diarization -> ***relative identity*** and ***time boundaries*** in a conversation

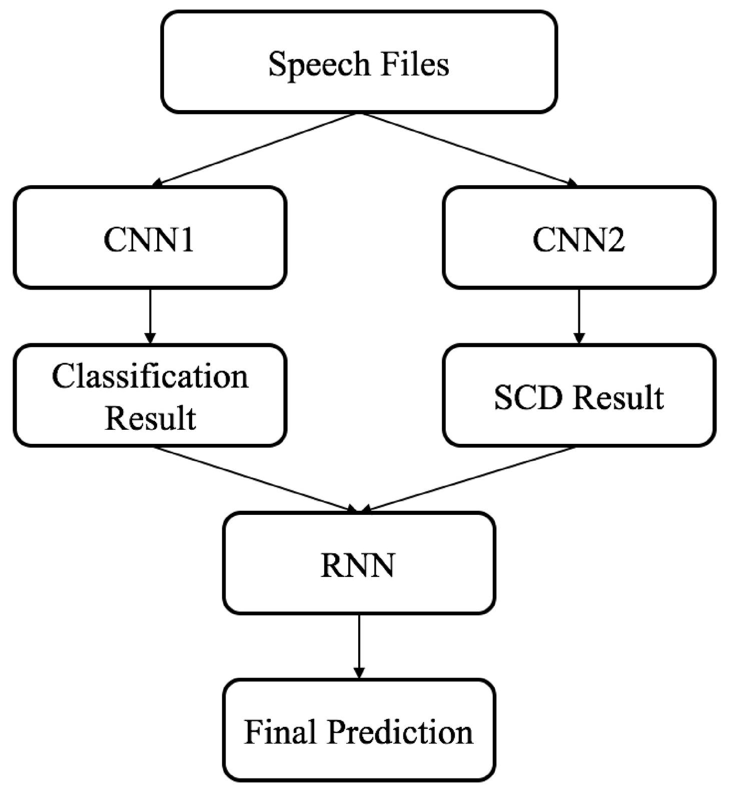
Why Joint?

- Needed In certain scenarios (e.g., call center)
- They could benefit each other:

❖ Diarization → Recognition: 1. temporal continuity (Speaker Change Detection, SCD); 2. sparsity (only a few identities exist in a conversation)

❖ Recognition → Diarization: Cross-speaker, cross-context training in speaker recognition helps capture highly discriminative features of speech.

Proposed System



➢ CNN1: Independently classifies the **absolute speaker identity** on equally spaced audio segments

➢ CNN2: Performs **Speaker Change Detection (SCD)** [1]

➢ RNN: Integrates predicted results from CNN1 and CNN2

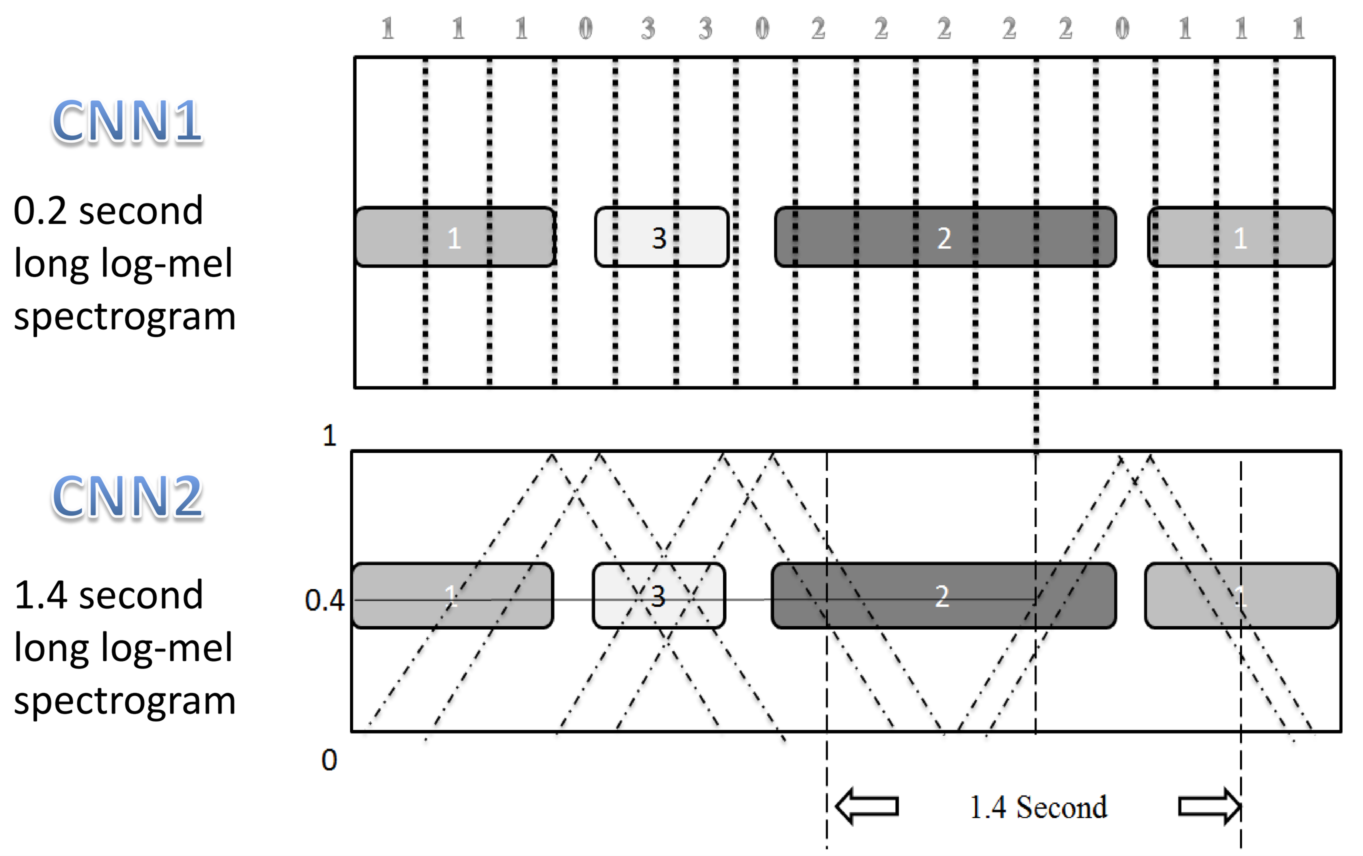
sparsity

CNN1 $\quad loss = y_{true} \times \log(y_{pred}) + \sqrt{y_{pred}}$

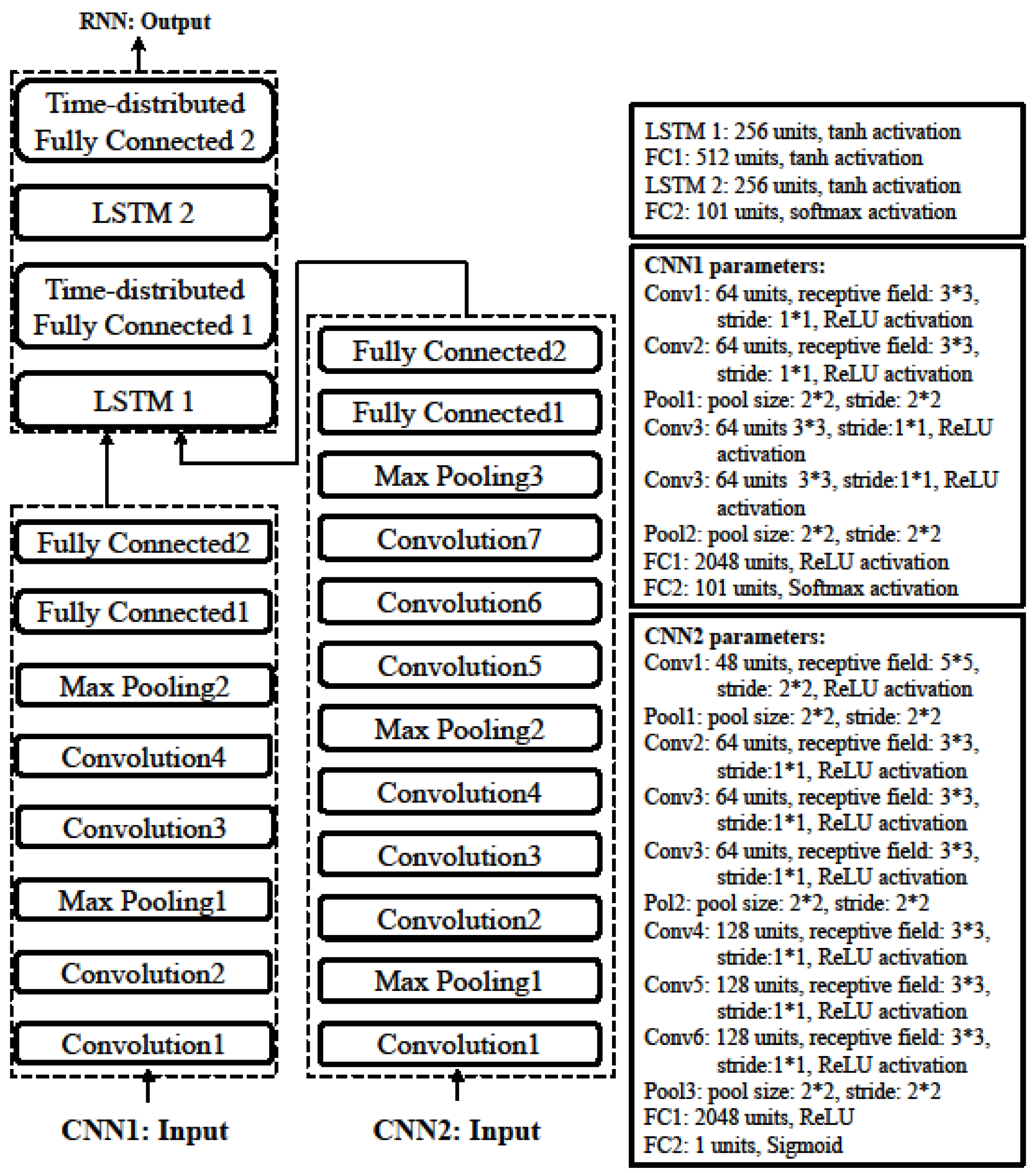
CNN2 $\quad loss = (0.1 + y_{true}) \times (y_{true} - y_{pred})^2$

Biases towards false positives of SCD

Pre-Processing



The Model



Dataset & Evaluation Measure

Dataset 1 – Call_Home Dataset:

➢ 50 two-person conversations (in total 100 speakers)

➢ Speaker labels are from "1" to "100". Silence is labeled "0".

➢ Use ***Classification Accuracy*** to evaluate on all of the 100 speakers

Dataset 2 – Prisoner Dataset from *Voice Biometrics Group*:

➢ 100 two-person conversations between a prisoner and an external partner (in total 10 prisoners)

➢ Speaker labels are from "1" to "10" for the prisoners. Silence is labeled "0". All the external partners are labeled "11".

➢ Use ***Precision*** and ***Recall*** to evaluate on only the prisoners.

$$\textbf{\textit{Precision}} = \frac{\#\ correctly\ detected\ segments\ of\ the\ prisoner}{\#\ total\ predicted\ segments\ of\ the\ prisoner}$$

$$\textbf{\textit{Recall}} = \frac{\#\ correctly\ detected\ segments\ of\ the\ prisoner}{\#\ ground-truth\ segments\ of\ the\ prisoner}$$

Experimental Results

Table 1. Predicted accuracy (mean ± std) comparisons.

Method	Acc.
(1) CNN1 w/ cross-entropy loss	0.711 ± 0.019
(2) CNN1 w/ sparsity constraint loss	0.741 ± 0.009
(3) CNN1 in (2) + all zeros SCD	0.743 ± 0.008
(4) CNN1 in (2) + predicted SCD	0.829 ± 0.004
(5) CNN1 in (2) + GT SCD	0.867 ± 0.003
(6) CNN1 restricted to GT identities	0.847 ± 0.007

Proposed

Unpractical baseline

- Comparing (1) and (2) shows the effectiveness of sparsity team in loss function of CNN1.
- Comparing (3), (4) and (5) shows the important role of SCD.
- Comparing (4) and (6) shows satisfying performance of the proposed method.

Table 2. Precision and recall for 10 prisoners.

ID	Pre.	Rec.	ID	Pre.	Rec.
1	0.921	0.776	6	0.933	0.832
2	0.767	0.836	7	0.235	0.006
3	0.796	0.837	8	0.941	0.753
4	0.786	0.838	9	0.743	0.777
5	0.899	0.830	10	0.370	0.607

Conclusions

➢ Proposed a system using two CNNs and an RNN to perform joint speaker diarization and recognition.

➢ Experiments show that our approach achieves satisfying results compared to a baseline that uses unpractical side information: *(6) CNN1 restricted to ground-truth identities*.

➢ Speaker Change Detection (SCD) plays an important role in the Final RNN prediction.

References

➢[1] Marek Hruz and Zbynek Zajic, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.