Conv-SINet

sebastien.warichet

February 2020

1 Abstract

The aim of my internship is to design and develop a Speaker Identification Network for Alcatel-Lucent Enterprise. We propose a fully-convolutional time domain identification solution.

2 Functional components

Speaker recognition systems are comprised of two functional components, enrollment and recognition.

2.1 Enrollment

The enrollment component creates reference models for each speaker from the corresponding speaker's utterance.

2.2 Recognition

The identification component matches the sequence of feature vectors of the speaker to be identified with those reference models in the system.

3 Variants of speaker recognition

3.1 Text-Dependent

If the text must be the same for enrollment and verification this is called text-dependent recognition.

3.2 Text-Independent

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. This is the variant we are targeting.
The text independent property could be reach with an appropriate training task.

To train the network a Siamese network should be used.
Minimize anchor vs positive: Two different words with the same speaker. With that we separate the content (the word) and the modelization of the vocal apparatus. Maximize anchor vs negative: At the start of learning, we train the model using two different words with two different speakers. After this first part of training, we train using same words with two different speakers.

4 Architecture

4.1 Flowcharts

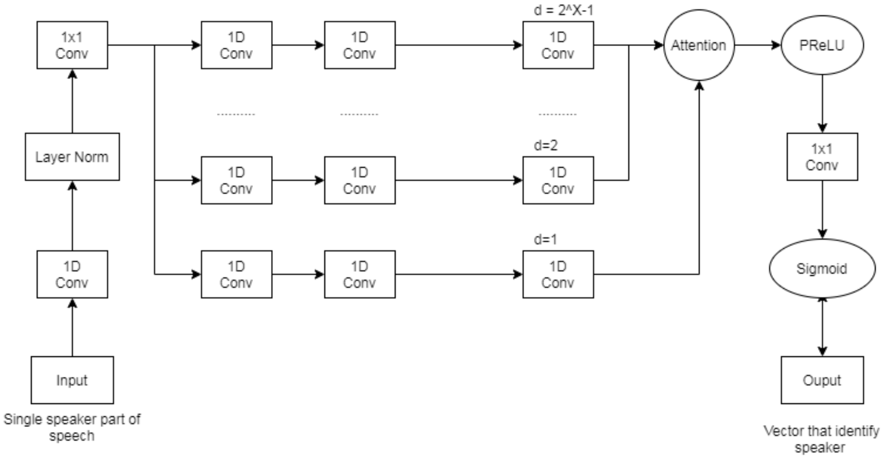

Figure 1: 1D conv block

4.2 1-D convolution block

The 1D conv block is basic brick of this neural network. This block is issued of Conv-TasNet research [2].
It has two outputs, one corresponding to the residual path and one corresponding to the skip-connection path: the residual path of a block serves as the input to the next block, and the skip-connection paths of all blocks are summed up and used as the output of the TCN.
The output is an addition of the input and the last 1x1-conv(on the upper right of the schema). The goal of this addition is to stabilize the system in two ways. The first stabilization is due to that output is forced to be dependent of the input whatever happens in the convolution.
The second stabilization is linked to the features, with this addition if the system have learned something, we can make this hypothesis input feature are compliant with last 1x1-conv (There is no chance for that: addition between potatoes and carrots could learn something). Trust in deep learning.

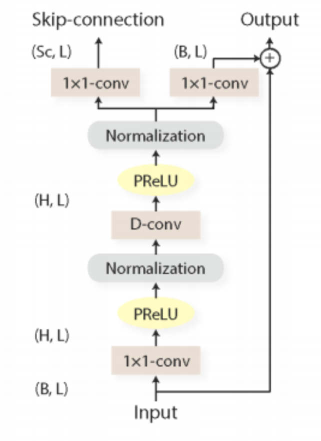

Figure 2: 1D conv block

4.3 Attention

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others. In machine learning the principle is very similar. For each features, attention mechanism produces a weight which indicate to the system how the feature should be taken into account on the result.
These weight could be used to try to understand how decision is making by our network. In our case attention weights will indicate which time range is important to resolve our problem.
To be relevant attention attention must be use-full, in other words we must have a better result using attention.

5 Experiment

5.1 Data set

5.2 Training objective

5.3 Evaluation metrics

5.4 Results

6 Discussion

References

[1] A. Nagrani, J. S. Chung, A. Zisserman VoxCeleb: a large-scale speaker identification dataset INTERSPEECH, 2017

[2] Yi Luo, Nima Mesgarani Conv-TasNet: Surpassing Ideal Time-FrequencyMagnitude Masking for Speech Separation 2019