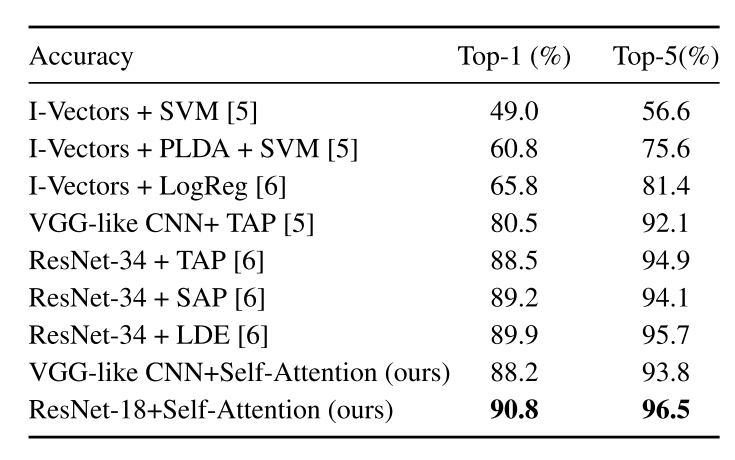
4 Previous Results

In the essay, researchers conduct experiments focusing on various elements of the speaker identification system. They quantify the testing result in the form of top-1 and top-5 accuracy, which suggests the probability of the right speaker existing in the most or 5 most possible labels. The experimental results show the superiority of self-attention mechanism, and lead us to focus on FBank acoustic feature and average pooling layer.

**4.1 Results for speaker identification**

From the perspective of methods, researchers compare CNN with self-attention mechanism, other CNN baselines and traditional i-vector-based methods on VoxCeleb database. While i-vector-based methods can only achieve a maximum top-1 accuracy of 65.8%, VGG-like CNN can get a 80.6% accuracy and ResNet-34 get a 89.9%. Our proposed systems reach 88.2% and 90.8% of the top-1 accuracy respectively and 93.8% and 96.5% of the top-5 accuracy. Generally speaking, traditional i-vector-based methods have the worst performance and CNN with self-attention mechanism has the best.

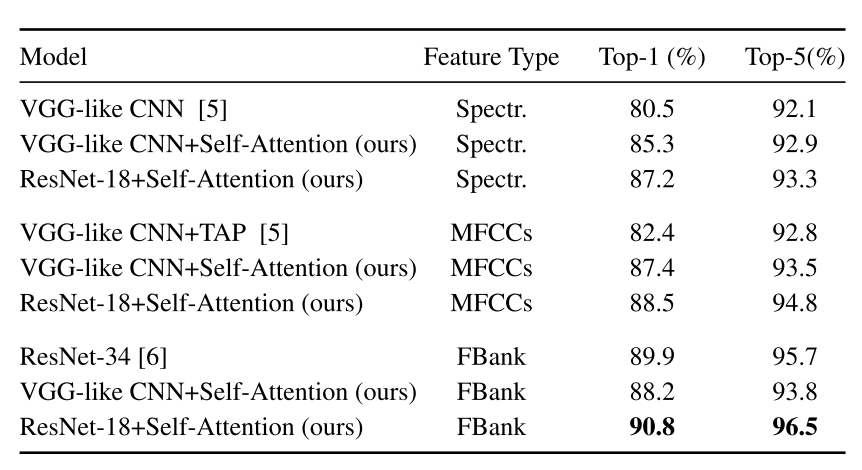
It’s obvious that ResNets outperforms VGG in speaker identification task, thus suits our future experiments better. ResNet-18 with self-attention mechanism outperforms others significantly, reaching 90.8% top-1 accuracy. The outstanding performance also suggests the importance of self-attention mechanism in improving identification accuracy.



**4.2 Results for different acoustic features**

Compared with spectrograms and MFCCs, FBank features lead to the highest accuracies. Using the FBank features, our proposed architecture reaches a top-1 accuracy as high as 90.8% and top-5 accuracy as high as 96.5%. Among all the three feature types, self-attention mechanism improves the performance obviously, especially for the top-1 accuracy.

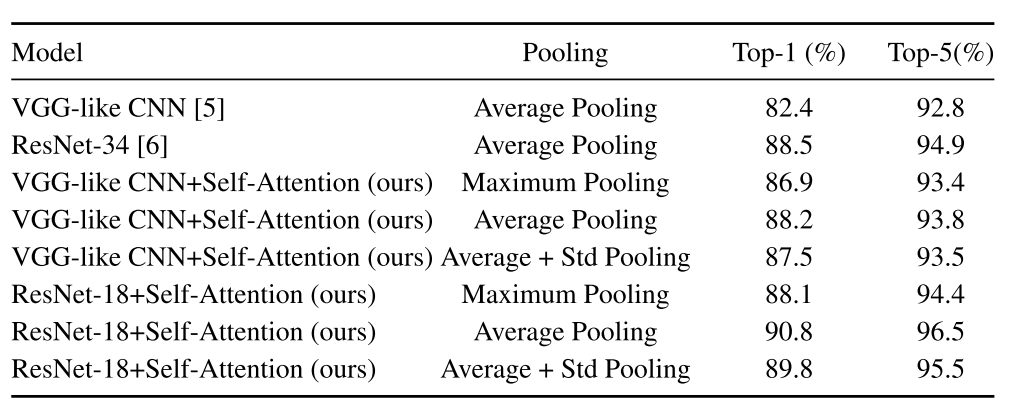
These findings prompt us to adopt FBank features in the following experiments. FBank features can be generated after the pre-processing stage.



**4.3 Results for temporal pooling layers**

In our proposed architecture, a temporal pooling layer is inserted between the self-attention layer and the last fully connected layer. There are three common types of temporal pooling layers, i.e., average pooling, maximum pooling and combination of average and standard deviation pooling. Average pooling always gets the best result regardless of the CNN type, reaching 90.8% top-1 accuracy and 96.5% top-5 accuracy for ResNet-18 variant.

These findings prompt us to adopt average pooling in the following experiments.



5 Research question

Our preliminary research question is adding new speakers to the trained network for identification. This project aims at not only the training and the identification on basic dataset, but also on additional speaker utterance. We will try to add additional speakers to the trained CNN model, adjusting parameters accordingly to achieve the lowest loss and highest identification accuracy. To be more specific, we would like to quantify the loss incurred by the newly added data and the effect on final result. If all goes well, we can compare the performance of the proposed CNN variants with other CNN baselines and traditional i-vector-based methods.

The ability to learn and identify new utterance will be critical in practical applications like smart homes and smart vehicles. That’s why we try to focus our project on it.

It is worth noting that it is possible to cluster an unknown speaker through the activations of an upper (dense or softmax) layer of a pre-trained identification CNN as a feature vector. Randomly chosen unknown speakers from VCTK database can produce a clear separation for the resulting vectors of last layers. This provides the basis for our research question.