Speaker Recognition From Raw Wave Form With SincNet and ResNet Fusion

Raz Rajwan rajwanraz@gmail.com, Gal Oren Gal287199@gmail.com

Electric and Electronic Engineering, Tel Aviv University

**Abstract**

# Speaker Recognition is a challenging task with essential applications such as authentication, automation, and security. It seems that DL becomes the now state-of-the-art solution for both speaker verification and identification. The standard x-vectors, additional to i-vectors [2], [7], are used as baseline in most of the novel works. The increasing amount of gathered data opens up the territory to DL, where they are the most effective

# The SincNet [3], [4] is a new deep learning based model which has produced promising results to tackle the mentioned task. In this work we try to use the advantages of the SincNet filters and ResNet [6] to produce a state-of-the-art solution to speaker identification on TIMIT data base.

# **Introduction**

Speaker Recognition and Automatic Speech Recognition (ASR) are two of the most actively researched and interesting fields in the computer science domain. Speaker recognition has applications in various fields such as biometric authentication, forensics, security and speech recognition, which has contributed to steady interest in this discipline. The conventional method of speaker identification involves classification of features extracted from speech such as the Mel Frequency Cepstral Coefficients (MFCC). With the advent of i-vectors, speaker verification has become faster and more efficient as compared to the preceding model based on the higher dimensional Gaussian Mixture Model (GMM) super vectors. i-vectors are used to generate a verification score between the embedding of two speakers. This score gives us information about whether both are the same speaker or different speakers. Previous experiments have shown that i-vectors perform better with Probabilistic Linear Discriminant Analysis (PLDA). Work has also been carried out in the field of training the i-vectors using different techniques in order to get better embedding, therefore, better results. Currently, researchers are moving towards Deep Neural Network (DNN) to obtain speaker embedding. DNN can be directly optimized to distinguish between speakers. DNN showed promising results in comparison to statistical measures such as i-vectors. X-vectors are seen as an improvement over the i-vector system (which is also a fixed-dimension vector system) because they are more robust and have yielded better results. X-vector extraction methodology employs a Time-Delayed Neural Network (TDNN) to learn features from the variable-length audio samples and converts them to fixed dimension vectors. This architecture can broadly be broken down in their order of occurrence into three units, namely, frame-level layers, statistics pooling layer and the segment level layers. X- Vectors can then be used with classifiers of any kind to carry out recognition tasks.

SincNet is a deep neural network that has embedded band pass filters for extracting features from the audio sample. The features are than fed into DNN based classifiers and in our case is fed to ResNet. We have used SincNet filters and not fed the audio waveform directly into the DNN based classifiers as the latter technique poses problems like high convergence time and less appealing results. The SincNet filters are actually band pass filters which are derived from parameterized sinc functions.

# **Related Work**

Until deep learning was introduced i-vectors [10] was the state-of-the-art architecture for speaker recognition and identification.

Recently, deep learning techniques have shown better results than the old fashion methods. For instance, CNN has proven high results on image classification and can also be used for speaker tasks by feeding the network raw audio samples. MFCC [8], [11] and Spectrogram [9] are hand crafted features which can be used as input to a CNN and improve performance significantly.

As mentioned X-Vector [7] is TDNN based method that replaces the i-vector and results high performance in speaker tasks such as speaker identification and speaker recognition.

Our architecture is mainly inspired from SincNet [1], [3], [4] and ResNet [6]. SincNet is an architecture that uses sinc-based filters to extract different frequency coefficients from the signal. The combination of these two networks results in high performance in the specified task.

# **Methods**

# **Sinc convolution Layer:**

SincNet tries to discover interpretable and meaningful filters by introducing an additional 1D convolution layer realized by sinc functions, followed by standard CNN layers. In general, it is straight forward to use rectangular (ideal) band-pass filters to decompose a signal into a number of frequency bands in the frequency domain. In fact, the frequency response of a band-pass filter can be written as the difference of two rectangular low-pass filters:

(1

Since the network input are time domain samples of the audio, inverse Fourier transform has to be done on the filter:

2sinc (2πn) - 2sinc (2πn) (2

Sinc(x) is defined as sin(x)/x.

An ideal band-pass filter (i.e., a filter where the pass-band is perfectly flat and the attenuation in the stop-band is infinite) requires an infinite number of elements L. Any truncation of g thus inevitably leads to an approximation of the ideal filter, characterized by ripples in the pass-band and limited attenuation in the stop-band. A popular solution to mitigate this issue is windowing. Windowing is performed by multiplying the truncated function g with a window function w, which aims to smooth out the abrupt discontinuities at the ends of g. we used Hann window that defined by:

(3

Eventually the sinc filters defined by:

(4

There are 2 ways to choose the cutoff frequencies (f1 and f2):

1. Linear linspace from 0 to fs/2.
2. Mel – Scale filter bank.

We choose to work with Mel – Scale filter bank since its emphasis the lower frequencies where many crucial speech information is located.

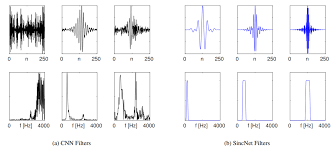


Figure 1: Comprasion between CNN learned Filters (a) and SincNet filters (b) structrue in frequency and time domain.

Figure 1illustrates the difference between a SincNet filters and CNN learned filters. Eventually the CNN filters resemble band pass filters with different cutoff frequency. The advantages of using SincNet over CNN are:

1. Less parameters to train –SincNet learns only the cut off frequencies.
2. Interpretability – after training it is clear which frequencies of the signal consists most of the information.
3. Minimum distortion to the input signal resulting better feature extraction and thus better classification results over the Corpus.

# **ResNet architecture**

ResNet is one of the most well-known backbone networks in deep neural networks. Comparing to prior network architectures, ResNet introduces a shortcut connection to address the problem of vanishing gradient, and further extracts abundant semantics from the input data to build a robust classifier. In this paper, ResNet-18 pre trained using the ImageNet dataset is selected as our 2D kernel learning backbone network. The input to ResNet-18 is the two-channel 2D representations obtained from the 1D sinc convolution stage, **and the output of the last convolution layer (i.e., conv5\_2) is a 512-channel 5 × 32 feature volume from which discriminative features will be extracted.**

# **Our Architecture**

Our architecture is described in figure 2**.**

Given the raw waveform, ResincNet which consists of three sets of SincNet filters of different lengths, is first designed to learn 2D representations. The outputs of each set of SincNet filters are concatenated to form a 2D representation, and then all 2D representations are stacked and fed into the ResNet for extracting embeddings. Finally, an adaptive pooling module is applied to summarize the information across time and frequency aspects to obtain compact features for the classification task. In this study, the end-to-end training strategy is used to fine-tune the parameters in the proposed ResincNet. The parameters of the Sinc filters are initialized by Mel-scale cut-off frequencies, and the parameters of ResNet are initialized by pre trained model on ImageNet dataset.

# **Data**

In our experiments we used TIMIT Corpus. TIMIT is a corpus of phonemically and lexically transcribed speech of [American English](https://en.wikipedia.org/wiki/American_English) speakers of different sexes and dialects. Each transcribed element has been delineated in time. TIMIT contains high quality recordings of 630 individuals/speakers with 8 different American English dialects, with each individual reading up to 10 phonetically rich sentences. TIMIT Corpus documentation suggests training (≈ 70%) and test sets. The training set contains 4620 utterances, but usually only SI and SX sentences are used, resulting in 3696 sentences from 462 speakers. The test set contains 1344 utterances from 168 speakers. The core test set, which is the abridged version of the complete testing set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). With the exception of SA sentences which are usually excluded from tests, the training and test sets do not overlap.

# **Experiments**

# **Data Preparation and net set up:**

The waveform (sampled in 16 kHz) of each speech sentence was split into chunks of 200 ms (3200 samples) with 40 ms overlap which were fed into our described ResincNet architecture. The first layer performs three sinc-based convolutions in parallel as described **in section 3.1** using three different length filters of L=251, L=501 and L=1001 taps. All of the Sinc layers have **40** different mel-frequencies filters which emphasis the lower part of the spectrum. Afterwards batch normalization and Relu was used on the output of the convolution layers. Subsequently, a ResNet-18 network was used and afterward a Soft Max layer and Cross Entropy loss.

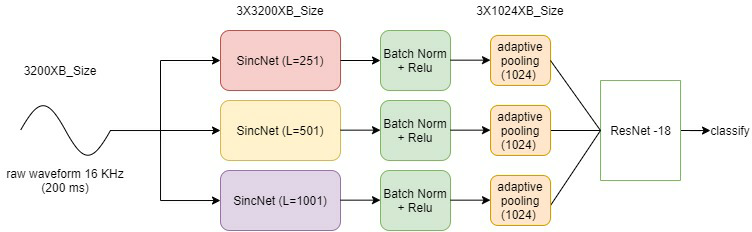
In the training we used learning rate of 0.001 and mini batch of 64 chunks for **2000** epochs on NVIDIA RTX 2070. To optimize the net parameters, we used RMSProp.

Figure 2: Blocks of our Res SincNet architecture.

|  |  |
| --- | --- |
| Architecture | TIMIT CER(%) |
| CNN | 4.19 |
| MFCC-CNN | 1.6 |
| SincNet | 1.52 |
| Res SincNet | 0.44 |

Table 1: Comprasion between the performance achieved on speaker identification (Classification Error Rates - CER%). The table shows that our architecture outperforms other methods on TIMIT data set.

# **Results**

We compared 4 different Architectures on TIMIT data set:

* CNN – feeding CNN raw samples.
* MFCC-CNN – feeding to a CNN Mel-frequency cepstral coefficients [11].
* SincNet –our implementation to SincNet described in 3.1.
* Res SincNet – our architecture described in 3.3.

Table 1 emphasizes the contribution of adding a pre trained ResNet layer after the traditional SincNet. Clearly, our architecture surpasses other standard methods performance.

Figure 3 shows the Classification error rate as a function of the epoch in the training process. We can see that all of the models converge after 1000 epochs with this set up. Res SincNet takes more epochs to converge due to is higher number of parameters.

# **Appendix**

In this work we Re implemented Sinc filters and designed Res Sinc ResNet architecture. In addition we implemented CNN and MFCC architectures for comparison. Our code is available in github: <https://github.com/galoren287199/finalProjectSincNet>.

Our project was mainly inspired from SincNet[4] and ResNet [6] papers.

# **Conclusions**

In this study, we proposed an end-to-end CNN architecture, called ResincNet, which can jointly learn 1D kernels and 2D kernels, for speaker identification task. For 1D kernel learning, we use SincNet filters to obtain variant 2D representations from raw audio waveform rather than pre-computed hand-crafted features such as MFCC [11] and spectrogram coefficients. Then, 2D kernel learning using ResNet-18 is used to extract embeddings from these learned 2D representations. The spatial average pooling module is used to get the compact features from the output of the last convolution layer of ResNet-18. In the experiments, the proposed architecture achieves classification accuracy 99.56% on TIMIT dataset, which outperforms the traditional methods.

We can see that the combination of the classification results of different features often yields better performance than each individual feature.

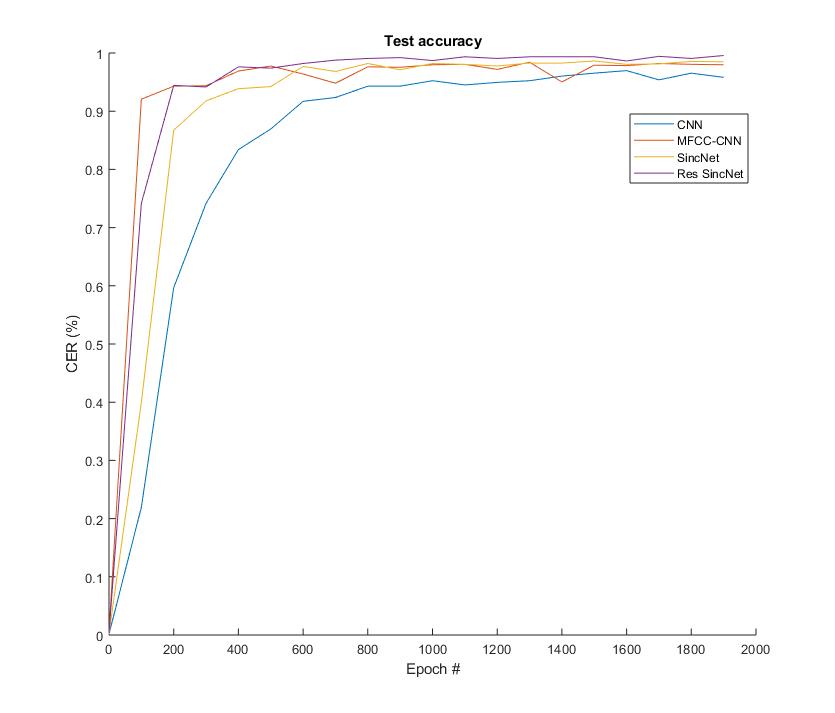


Figure 3: This graph compares the training proccess of each method.

##### **References**

1. J. A. C. Nunes, D. Macedo and C. Zanchettin, “Additive Margin SincNet for Speaker Recognition”, 2019.
2. D. sztaho, G. szaszak and A. Beke “Deep learning methods in speaker recognition: a review”.
3. M. Ravanelli, Y. Bengio, “Interpretable Convolutional Filters with SincNet”, 2019.
4. M. Ravanelli, Y. Bengio, “SPEECH AND SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET”, 2018.
5. M. Tripathi, D. Singh and S. Susan, “Speaker Recognition using SincNet and X-Vector Fusion”, 2020.
6. K. HE, X. Zhang, Z. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, 2015.
7. D. Synder, D. Gracia-Romero, G. Sell, D. povey and S. Khundanpur, "X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION".
8. A. Hossan, Sheeraz Memon and M. Gregory, "A Novel Approach for MFCC Feature Extraction".
9. Y. Jia, J. Yu and L. Wang, "Speaker recognition based on characteristic spectrograms and an improved self‑organizing feature map neural network".
10. N. Salwani Ibrahim, D. Athiar Ramli, "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction", 2018.
11. A. Ashar, M. Sahid Bhatii and U.Mushtaq, "Speaker Identification Using a Hybrid CNN-MFCC Approach".