Data Driven Filter Learning for Speaker Recognition using Raw waveforms

Sensing
Perception &
INtelligence (SPIN) Lab

Aditya Suryawanshi, Dr. Neeraj Sharma

a.suryawanshi@iitg.ac.in, neerajs@iitg.ac.in

Indian Institute of Technology Guwahati, India

Abstract

This work explores the use of a custom filter for convolutional layers of CNN to improve on the feature extraction of standard CNNs. The performance of the filter will be analysed using a speaker recognition CNN model with raw audio waveforms as input.

Problem Statements

- Despite advancements in speaker recognition using CNNs, traditional approaches relying on hand-crafted features like MFCCs limit the ability to fully exploit raw waveform data, obscuring critical speaker-specific characteristics. Conventional CNN filters trained on raw waveforms often become noisy and less interpretable, especially with limited data.
- There is a need for optimized filter designs, such as SincNet's [4] parametrized filters, which learn meaningful spectral features efficiently while improving model performance and interpretability.

Framework

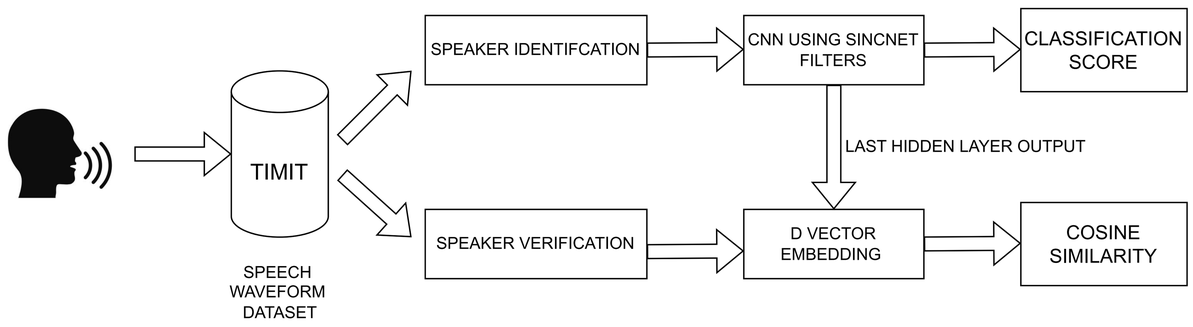


Figure 1: **Pipeline**: Utterences from various speakers are collected(TIMIT Dataset [2]).Two tasks, speaker identification and verification are performed using CNN.The classification score and cosine similarity are noted for identification and verification respectively.

Speaker Identification Architecture

- **Dataset Used**:The TIMIT [2] dataset has been used that features recordingsof 630 speakers of 8 dialects of American English each reading 10 phonetically-rich sentences.
- **Preprocessing**: Non-speech intervals at the beginning and end of each sentence were removed.To address text-independent speaker recognition, the calibration sentences of TIMIT (i.e.,the utterances with the same text for all speakers) have been removed.

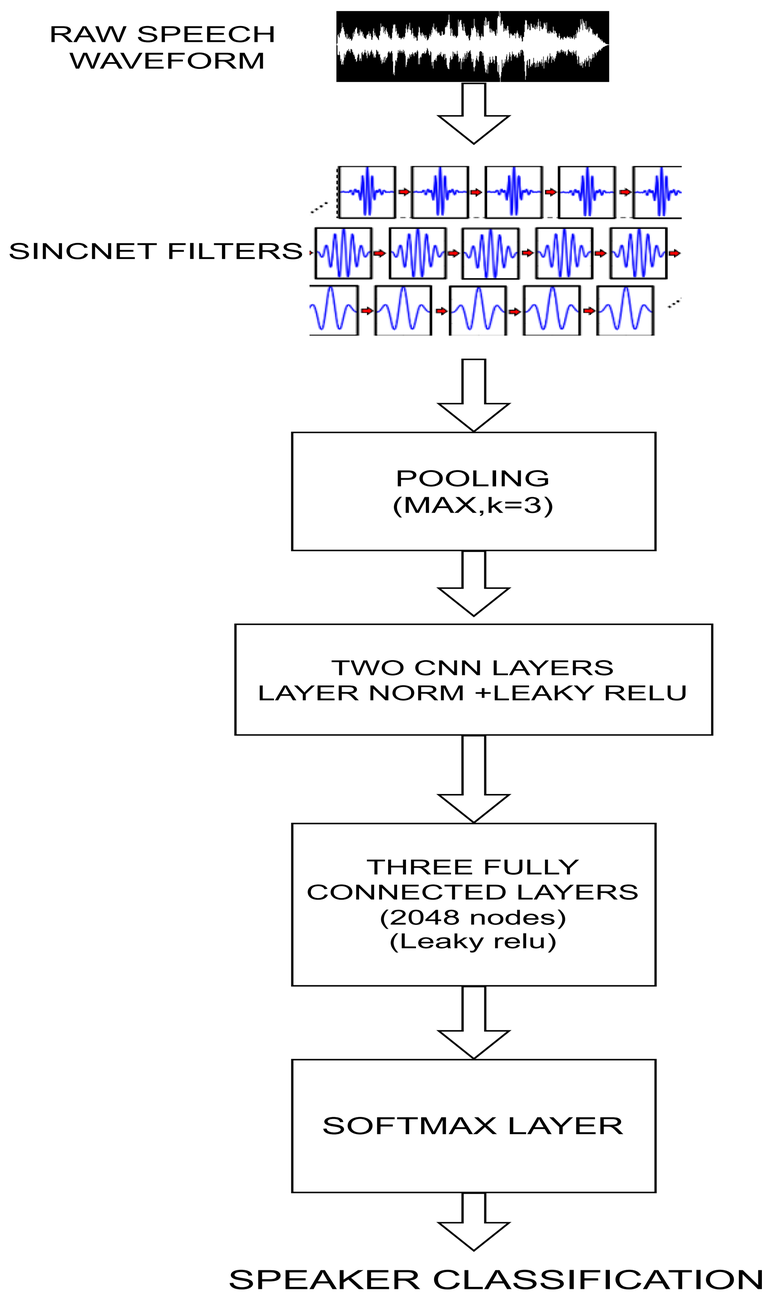


Figure 2: **SINCNET Architecture**:The SincNet architecture uses sinc functions for convolution in the first layer of CNN this creates a band pass filter only allowing important features to pass to the following layers ensuring better performance than a standard CNN architecture.

Speaker Verification Architecture

- **Dataset Used**:The Librispeach [3] dataset has been used that features recordings of 2484 speakers. The Librispeech sentences with internal silences lasting more than 125 ms were split into multiple chunks
- **Preprocessing**: The training and test material have been randomly selected to exploit 12-15 seconds of training material for each speaker and test sentences lasting 2-6 seconds.The waveform of each speech sentence was split into chunks of 200 ms (with 10 ms overlap), which were fed into the SincNet architecture.

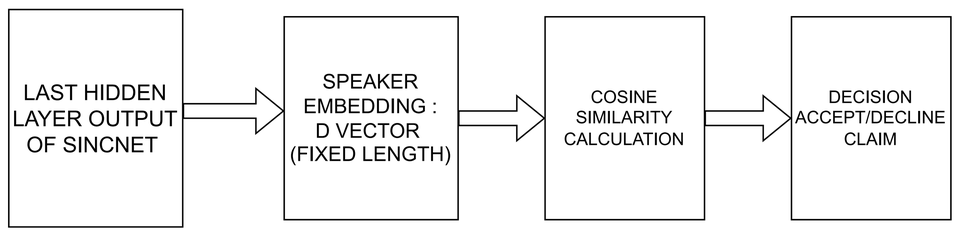
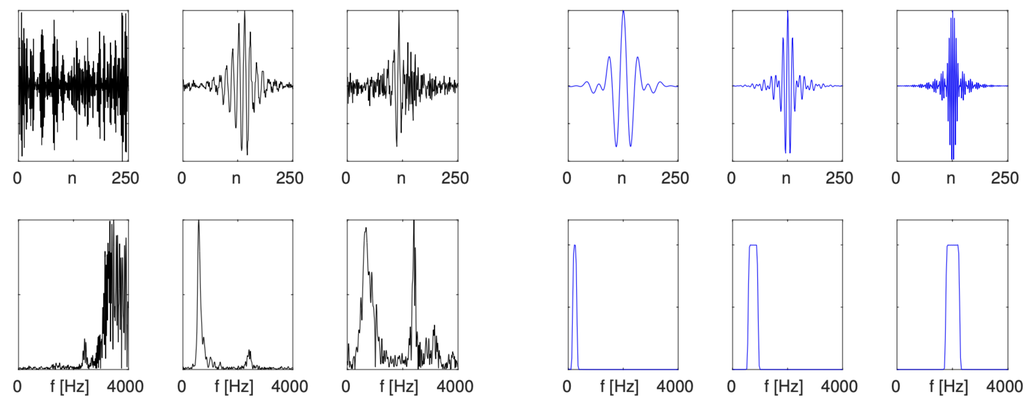


Figure 3: **Speaker Verification architecture:** We take the last hidden layer and generate a fixed length d-vector to uniquely represent the speaker. This vector is than compared with the d-vector of the claimed speaker and a threshold is applied to determine if the test sample is indeed from the claimed speaker or not.

Results and Discussion

- Compared to traditional methods, the SincNet-based CNN model gives better results for short speech raw inputs .
- Due to its design, which applies constraints to the learned filters (focusing on relevant frequency bands), SincNet tends to converge faster during training compared to traditional CNN models. This is particularly beneficial when working with limited data or computational resources.
- .As shown in experiments speaker identification on TIMIT dataset, SincNet outperforms standard CNNs, achieving lower classification error rates (1.24% compared to 1.72% for CNNs) .
- Speaker verification on Librispeach dataset, SincNet outperforms standard CNNs and i-vector[1] baseline, achieving lower verification Equal Error Rate (0.51% compared to 0.58% for CNNs) .
- Visualising the filters of CNN and SincNet and comparing we can note that the standard CNN does not always learn filters with a well-defined frequency response, in some cases the frequency response looks noisy, while in others assuming multiband shapes.



(a) CNN Filters (b) SincNet Filters

Figure 4: **Filters Learned:** The first row reports the filters in the time domain, while the second one shows their magnitude frequency response.

Future Work

- Apply the same method to define custom filters in 2D CNNs.
- Perform speaker recognition by converting the TIMIT and librispeach dataset to their spectrogram representation and compare it with the current results.
- Create different methods to choose custom filters for 2D CNNs to improve on feature extraction.

References

[1] N. Dehak et al. "Context- dependent pre-trained deep neural networks for large vocabulary speech recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* (2012).

[2] et al. Garofolo John S. " TIMIT Acoustic-Phonetic Continuous Speech Corpus". In: *Linguistic Data Consortium* (1993).

[3] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *IEEE International Conference on Acoustics* (2015).

[4] Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with SincNet". In: *Arvix* (2019). DOI: 10.1109/slt.2018.8639585.