Interpretable Convolutional Filters with SincNet

Mirco Ravanelli
Mila, Université de Montréal

Yoshua Bengio
Mila, Université de Montréal
CIFAR Fellow

Abstract

Deep learning is currently playing a crucial role toward higher levels of artificial intelligence. This paradigm allows neural networks to learn complex and abstract representations, that are progressively obtained by combining simpler ones. Nevertheless, the internal "black-box" representations automatically discovered by current neural architectures often suffer from a lack of interpretability, making of primary interest the study of explainable machine learning techniques.

This paper summarizes our recent efforts to develop a more interpretable neural model for directly processing speech from the raw waveform. In particular, we propose *SincNet*, a novel Convolutional Neural Network (CNN) that encourages the first layer to discover more meaningful filters by exploiting parametrized sinc functions. In contrast to standard CNNs, which learn all the elements of each filter, only low and high cutoff frequencies of band-pass filters are directly learned from data. This inductive bias offers a very compact way to derive a customized filter-bank front-end, that only depends on some parameters with a clear physical meaning. Our experiments, conducted on both speaker and speech recognition, show that the proposed architecture converges faster, performs better, and is more interpretable than standard CNNs.

1 Introduction

Deep learning has recently contributed to achieving unprecedented performance levels in numerous tasks, mainly thanks to the progressive maturation of supervised learning techniques [1]. The increased discrimination power of modern neural networks, however, is often obtained at the cost of a reduced interpretability of the model. Modern end-to-end systems, whose popularity is increasing in many fields such as speech recognition [2–4], often discover "black-box" internal representations that make sense for the machine but are arguably difficult to interpret by humans. The remarkable sensitivity of current neural networks toward adversarial examples [5], for instance, not only highlights how superficial the discovered representations could be but also raises crucial concerns about our capabilities to really interpret neural models. Such a lack of interpretability can be a major bottleneck for the development of future deep learning techniques. Having more meaningful insights on the logic behind network predictions and errors, in fact, can help us to better trust, understand, and diagnose our model, eventually guiding our efforts toward more robust deep learning. In recent years, a growing interest has been thus devoted to the development of interpretable machine learning [6, 7], as witnessed by the numerous works in the field, ranging from visualization [8, 9], diagnosis of DNNs [10], explanatory graphs [11], and explainable models [12], just to name a few.

Interpretability is a major concern for audio and speech applications as well [13]. CNNs and Recurrent Neural Networks (RNNs) are the most popular architectures nowadays used in speech and speaker recognition [2]. RNN can be employed to capture the temporal evolution of the speech signal [14–17], while CNNs, thanks to their weight sharing, local filters, and pooling networks are normally employed to extract robust and invariant representations [18]. Even though standard hand-crafted features such as FBANK and Mel-Frequency Cepstral Coefficients (MFCC) are still employed in many

state-of-the-art systems [19–21], directly feeding a CNN with spectrogram bins [22–24] or even with raw audio samples [25–37] is an approach of increasing popularity. The engineered features, in fact, are originally designed from perceptual evidence and there are no guarantees that such representations are optimal for all speech-related tasks. Standard features, for instance, smooth the speech spectrum, possibly hindering the extraction of crucial narrow-band speaker characteristics such as pitch and formants. Conversely, directly processing the raw waveform allows the network to learn low-level representations that are possibly more customized on each specific task.

The downside of raw speech processing lies in the possible lack of interpretability of the filter bank learned in the first convolutional layer. According to us, the latter layer is arguably the most critical part of current waveform-based CNNs. This layer deals with high-dimensional inputs and is also more affected by vanishing gradient problems, especially when employing very deep architectures. As will be discussed in this paper, the filters learned by CNNs often take noisy and incongruous multi-band shapes, especially when few training samples are available. These filters certainly make some sense for the neural network, but they do not appeal to human intuition, nor appear to lead to an efficient representation of the speech signal.

To help the CNNs discover more meaningful filters, this work proposes to add some constraints on their shape. Compared to standard CNNs, where the filter-bank characteristics depend on several parameters (each element of the filter vector is directly learned), SincNet convolves the waveform with a set of parametrized sinc functions that implement band-pass filters [38]. The low and high cutoff frequencies are the only parameters of the filter learned from data. This solution still offers considerable flexibility but forces the network to focus on high-level tunable parameters that have a clear physical meaning. Our experimental validation has considered both speaker and speech recognition tasks. Speaker recognition is carried out on TIMIT [39] and Librispeech [40] datasets under challenging but realistic conditions, characterized by minimal training data (i.e., 12-15 seconds for each speaker) and short test sentences (lasting from 2 to 6 seconds). With the purpose of validating SincNet in both clean and noisy conditions, speech recognition experiments are conducted on both the TIMIT and DIRHA dataset [41, 42]. Results show that the proposed SincNet converges faster, achieves better performance, and is more interpretable than a more standard CNN.

The remainder of the paper is organized as follows. The SincNet architecture is described in Sec. 2. Sec. 3 discusses the relation to prior work. The experimental activity on both speaker and speech recognition is outlined in Sec. 4. Finally, Sec. 5 discusses our conclusions.

2   The SincNet Architecture

The first layer of a standard CNN performs a set of time-domain convolutions between the input waveform and some Finite Impulse Response (FIR) filters [43]. Each convolution is defined as follows[1]:

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l] \tag{1}$$

where $x[n]$ is a chunk of the speech signal, $h[n]$ is the filter of length L, and $y[n]$ is the filtered output. In standard CNNs, all the L elements (taps) of each filter are learned from data. Conversely, the proposed SincNet (depicted in Fig. 1) performs the convolution with a predefined function g that depends on few learnable parameters θ only, as highlighted in the following equation:

$$y[n] = x[n] * g[n, \theta] \tag{2}$$

A reasonable choice, inspired by standard filtering in digital signal processing, is to define g such that a filter-bank composed of rectangular bandpass filters is employed. In the frequency domain, the magnitude of a generic bandpass filter can be written as the difference between two low-pass filters:

$$G[f, f_1, f_2] = rect\Big(\frac{f}{2f_2}\Big) - rect\Big(\frac{f}{2f_1}\Big), \tag{3}$$

---

[1]Most deep learning toolkits actually compute *correlation* rather than *convolution*. The obtained flipped (mirrored) filters do not affect the results.
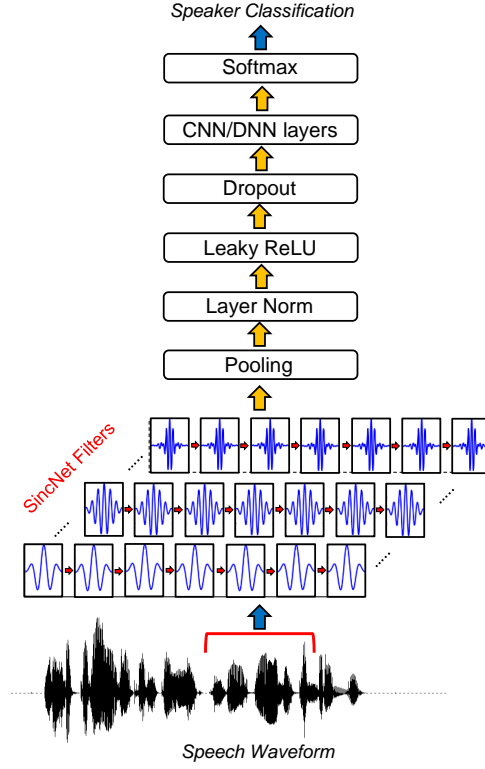
Figure 1: Architecture of SincNet.

where f_1 and f_2 are the learned low and high cutoff frequencies, and $rect(\cdot)$ is the rectangular function in the magnitude frequency domain[2]. After returning to the time domain (using the inverse Fourier transform [43]), the reference function g becomes:

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n), \tag{4}$$

where the sinc function is defined as $sinc(x) = sin(x)/x$.

The cut-off frequencies can be initialized randomly in the range $[0, f_s/2]$, where f_s represents the sampling frequency of the input signal. As an alternative, filters can be initialized with the cutoff frequencies of the mel-scale filter-bank, which has the advantage of directly allocating more filters in the lower part of the spectrum, where crucial speech information is located. To ensure $f_1 \geq 0$ and $f_2 \geq f_1$, the previous equation is actually fed by the following parameters:

$$f_1^{abs} = |f_1| \tag{5}$$

$$f_2^{abs} = f_1 + |f_2 - f_1| \tag{6}$$

Note that no bounds have been imposed to force f_2 to be smaller than the Nyquist frequency, since we observed that this constraint is naturally fulfilled during training. Moreover, the gain of each filter is not learned at this level. This parameter is managed by the subsequent layers, which can easily attribute more or less importance to each filter output.

An ideal bandpass filter (i.e., a filter where the passband is perfectly flat and the attenuation in the stopband 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 passband and limited attenuation in the stopband. A popular solution to mitigate this issue is windowing [43]. Windowing

---

[2]The phase of the $rect(\cdot)$ function is considered to be linear.
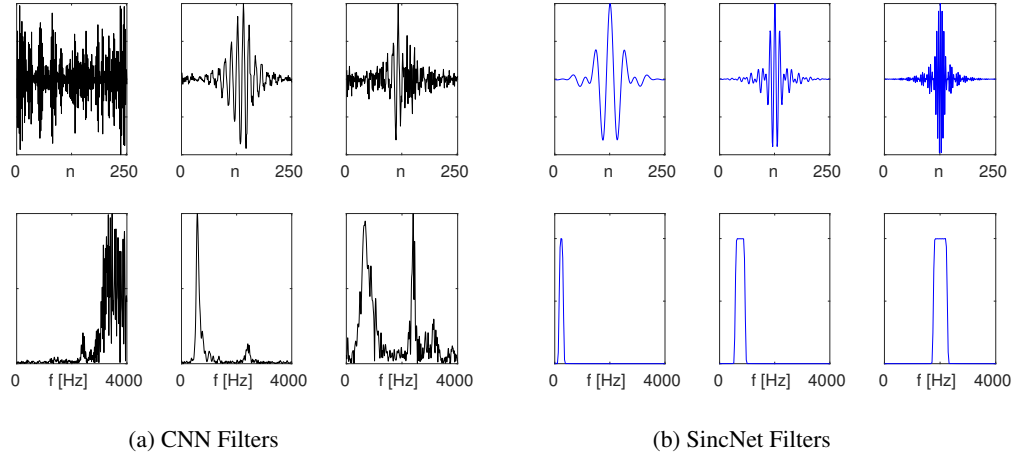
(a) CNN Filters   (b) SincNet Filters

Figure 2: Examples of filters learned by a standard CNN and by the proposed SincNet (using the Librispeech corpus on a speaker-id task). The first row reports the filters in the time domain, while the second one shows their magnitude frequency response.

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:

$$g_w[n, f_1, f_2] = g[n, f_1, f_2] \cdot w[n].\tag{7}$$

This paper uses the popular Hamming window [44], defined as follows:

$$w[n] = 0.54 - 0.46 \cdot cos\Big(\frac{2\pi n}{L}\Big).\tag{8}$$

The Hamming window is particularly suitable to achieve high frequency selectivity [44]. However, results not reported here reveal no significant performance difference when adopting other functions, such as Hann, Blackman, and Kaiser windows.

All operations involved in SincNet are fully differentiable and the cutoff frequencies of the filters can be jointly optimized with other CNN parameters using Stochastic Gradient Descent (SGD) or other gradient-based optimization routines. As shown in Fig. 1, a standard CNN pipeline (pooling, normalization, activations, dropout) can be employed after the first sinc-based convolution. Multiple standard convolutional, fully-connected or recurrent layers [15–17, 45] can then be stacked together to finally perform a classification with a softmax classifier.

Fig. 2 shows some examples of filters learned by a standard CNN and by the proposed SincNet for a speaker identification task trained on Librispeech (the frequency response is plotted between 0 and 4 kHz). As observed in the figures, the standard CNN does not always learn filters with a well-defined frequency response. In some cases, the frequency response looks noisy (see the first CNN filter), while in others assuming multi-band shapes (see the third CNN filter). SincNet, instead, is specifically designed to implement rectangular bandpass filters, leading to more a meaningful filter-bank.

2.1 Model properties

The proposed SincNet has some remarkable properties:

- **Fast Convergence:** SincNet forces the network to focus only on the filter parameters with a major impact on performance. The proposed approach actually implements a natural inductive bias, utilizing knowledge about the filter shape (similar to feature extraction methods generally deployed on this task) while retaining flexibility to adapt to data. This prior knowledge makes learning the filter characteristics much easier, helping SincNet to converge significantly faster to a better solution. Fig. 3 shows the learning curves of SincNet and CNN obtained in a speaker-id task. These results are achieved on the TIMIT dataset and highlight a faster decrease of the Frame Error Rate ($FER\%$) when SincNet is used.
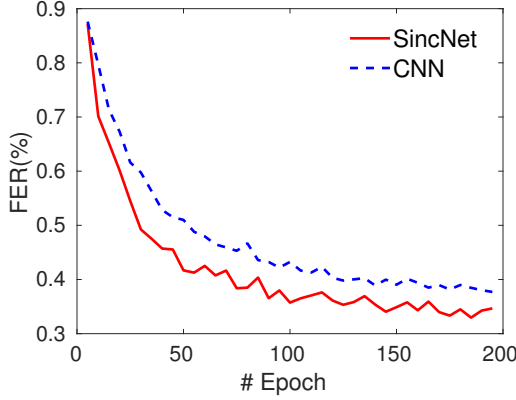
Figure 3: Frame Error Rate (%) obtained on speaker-id with the TIMIT corpus (using held-out data).
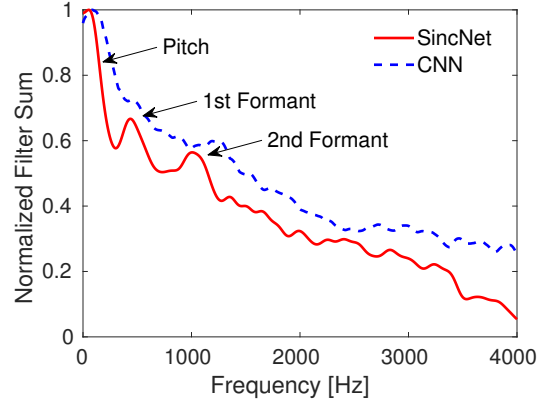

Figure 4: Cumulative frequency response of SincNet and CNN filters on speaker-id.

Moreover, SincNet converges to better performance leading to a FER of 33.0% against a FER of 37.7% achieved with the CNN baseline.

- **Few Parameters:** SincNet drastically reduces the number of parameters in the first convolutional layer. For instance, if we consider a layer composed of F filters of length L, a standard CNN employs $F \cdot L$ parameters, against the $2F$ considered by SincNet. If $F = 80$ and $L = 100$, we employ 8k parameters for the CNN and only 160 for SincNet. Moreover, if we double the filter length L, a standard CNN doubles its parameter count (e.g., we go from 8k to 16k), while SincNet has an unchanged parameter count (only two parameters are employed for each filter, regardless its length L). This offers the possibility to derive very selective filters with many taps, without actually adding parameters to the optimization problem. Moreover, the compactness of the SincNet architecture makes it suitable in the few sample regime.

- **Computational Efficiency**: The proposed function g is symmetric. This means we can perform convolution in a very efficient way by only considering one side of the filter and inheriting the results for the other half. This saves 50% of the first-layer computation over a standard CNN.

- **Interpretability**: The SincNet feature maps obtained in the first convolutional layer are definitely more interpretable and human-readable than other approaches. The filter bank, in fact, only depends on parameters with a clear physical meaning. Fig. 4, for instance, shows the cumulative frequency response of the filters learned by SincNet and CNN on a speaker-id task. The cumulative frequency response is obtained by summing up all the discovered filters and is useful to highlight which frequency bands are covered by the learned filters.

  Interestingly, there are three main peaks which clearly stand out from the SincNet plot (see the red line in the figure). The first one corresponds to the pitch region (the average pitch is 133 Hz for a male and 234 for a female). The second peak (approximately located at 500 Hz) mainly captures first formants, whose average value over the various English vowels is indeed 500 Hz. Finally, the third peak (ranging from 900 to 1400 Hz) captures some important second formants, such as the second formant of the vowel $/a/$, which is located on average at 1100 Hz. This filter-bank configuration indicates that SincNet has successfully adapted its characteristics to address speaker identification. Conversely, the standard CNN does not exhibit such a meaningful pattern: the CNN filters tend to correctly focus on the lower part of the spectrum, but peaks tuned on first and second formants do not clearly appear. As one can observe from Fig. 4, the CNN curve stands above the SincNet one. SincNet, in fact, learns filters that are, on average, more selective than CNN ones, possibly better capturing narrow-band speaker clues.

  Fig. 5 shows the cumulative frequency response of a CNN and SincNet obtained on a noisy speech recognition task. In this experiment, we have artificially corrupted TIMIT with a significant quantity of noise in the band between 2.0 and 2.5 kHz (see the spectrogram) and
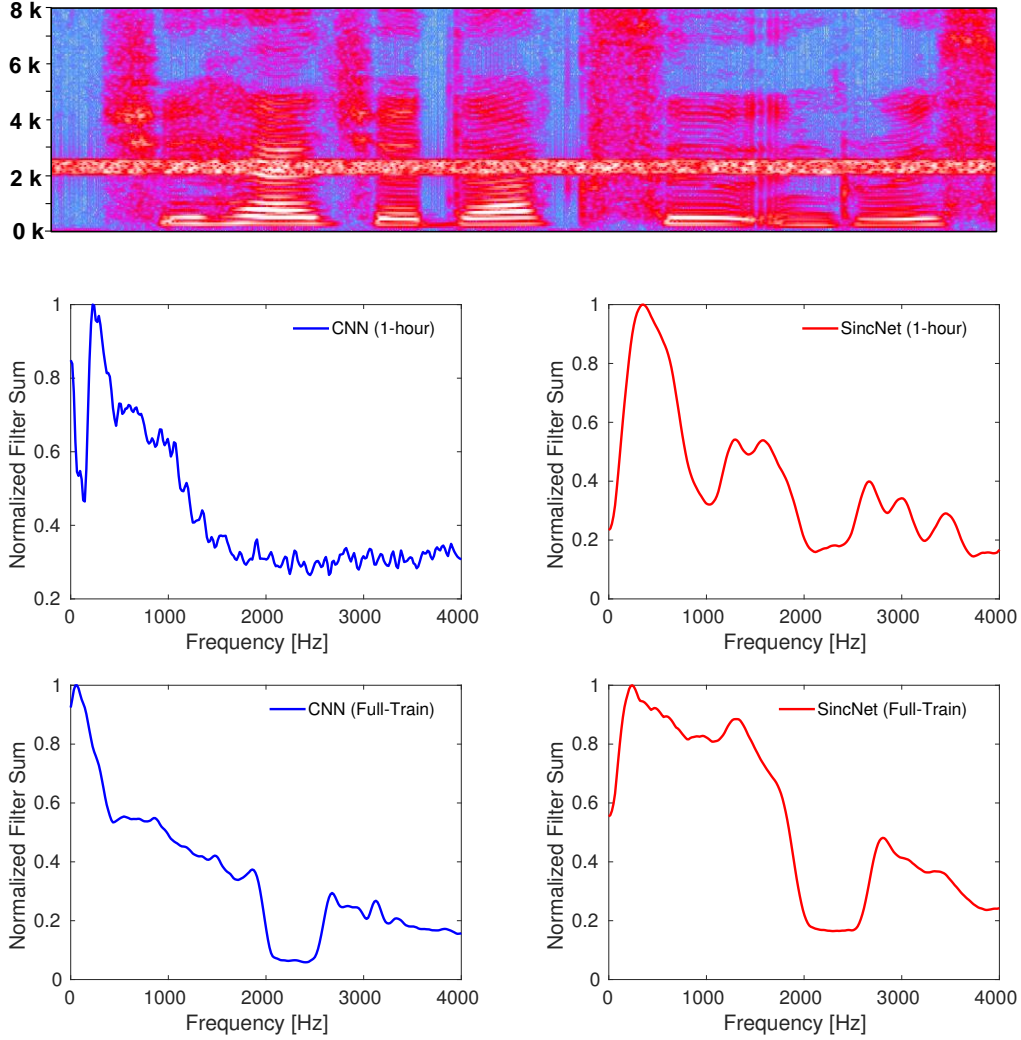
5

Figure 5: Cumulative frequency responses obtained on a speech recognition task trained with a noisy version of TIMIT. As shown in the spectrogram, noise has been artificially added into the band 2.0-2.5 kHz. Both the CNN and SincNet learn to avoid the noisy band, but SincNet learns it much faster, after processing only one hour of speech.

we have analyzed how fast the two architectures learn to avoid such a useless band. The second row of sub-figures compares the CNN and the SincNet at a very early training stage (i.e., after having processed only one hour of speech in the first epoch), while the last row shows the cumulative frequency responses after completing the training. From the figures emerges that both CNN and SincNet have correctly learned to avoid the corrupted band at end of training, as highlighted by the holes between 2.0 and 2.5 kHz in the cumulative frequency responses. SincNet, however, learns to avoid such a noisy band much earlier. In the second row of sub-figures, in fact, SincNet shows a visible valley in the cumulative spectrum even after processing only one hour of speech, while CNN has only learned to give more importance to the lower part of the spectrum.

3   Related Work

Several works have recently explored the use of low-level speech representations to process audio and speech with CNNs. Most prior attempts exploit magnitude spectrogram features [22–24, 46–48]. Although spectrograms retain more information than standard hand-crafted features, their design still

requires careful tuning of some crucial hyper-parameters, such as the duration, overlap, and typology of the frame window, as well as the number of frequency bins. For this reason, a more recent trend is to directly learn from raw waveforms, thus completely avoiding any feature extraction step. This approach has shown promise in speech [25–29], including emotion tasks [30], speaker recognition [35], spoofing detection [34], and speech synthesis [31, 32].

Similar to SincNet, some previous works have proposed to add constraints on the CNN filters, for instance forcing them to work on specific bands [46, 47]. Differently from the proposed approach, the latter works operate on spectrogram features and still learn all the L elements of the CNN filters. An idea related to the proposed method has been recently explored in [48], where a set of parameterized Gaussian filters are employed. This approach operates on the spectrogram domain, while SincNet directly considers the raw waveform in the time domain.

Some valuable works have recently proposed theoretical and experimental frameworks to analyze CNNs [49, 50]. In particular, [51, 35, 52] feed a standard CNN with raw audio samples and analyze the filters learned in the first layer on both speech recognition and speaker identification tasks. The authors highlight some interesting properties emerged from analyzing the cumulative frequency response and propose a spectral dictionary interpretation of the learned filters. Similarly to our findings, the latter works noticed that the filters tend to focus more on the lower part of the spectrum and they can sometimes highlight some peaks that likely corresponds to the fundamental frequency. In this work, we argue that all of these interesting properties can be observed more clearly and at an earlier training stage with SincNet.

This paper extends our previous studies on the SincNet [38]. To the best of our knowledge, this paper is the first that shows the effectiveness of the proposed SincNet in a speech recognition application. Moreover, this work not only considers standard close-talking speech recognition, but it also extends the validation of SincNet to distant-talking speech recognition [53–55].

4   Results

The proposed SincNet has been evaluated on both speech and speaker recognition using different corpora. This work considers a challenging but realistic speaker recognition scenario: for all the adopted corpora, we only employed 12-15 seconds of training material for each speaker, and we tested the system performance on short sentences lasting from 2 to 6 seconds. In the spirit of reproducible research, we release the code of SincNet for speaker identification[3] and speech recognition[4] (under the PyTorch-Kaldi project [56]). More details on the adopted datasets as well as on the SincNet and baseline setups can found in the **appendix**.

4.1   Speaker Recognition

Table 1 reports the Classification Error Rates (CER%) achieved on a speaker-id task. The table shows that SincNet outperforms other systems on both TIMIT (462 speakers) and Librispeech (2484 speakers) datasets. The gap with a standard CNN fed by raw waveform is larger on TIMIT, confirming the effectiveness of SincNet when few training data are available. Although this gap is reduced when LibriSpeech is used, we still observe a 4% relative improvement that is also obtained with faster convergence (1200 vs 1800 epochs). Standard FBANKs provide results comparable to SincNet only on TIMIT, but are significantly worse than our architecture when using Librispech. With few training data, the network cannot discover filters that are much better than that of FBANKs, but with more data a customized filter-bank is learned and exploited to improve the performance.

Table 2 extends our validation to speaker verification, reporting the Equal Error Rate (EER%) achieved with Librispeech. All DNN models show promising performance, leading to an EER lower than 1% in all cases. The table also highlights that SincNet outperforms the other models, showing a relative performance improvement of about 11% over the standard CNN model. Note that the speaker verification system is derived from the speaker-id neural network using the *d-vector* technique. The *d-vector* [19, 24] is extracted from the last hidden layer of the speaker-id network. A speaker-dependent d-vector is computed and stored for each enrollment speaker by performing an L2 normalization and averaging all the d-vectors of the different speech chunks. The cosine distance

---

[3] at `https://github.com/mravanelli/SincNet/`.
[4] at `https://github.com/mravanelli/pytorch-kaldi/`.

	TIMIT	LibriSpeech
DNN-MFCC	0.99	2.02
CNN-FBANK	0.86	1.55
CNN-Raw	1.65	1.00
SincNet	**0.85**	**0.96**

Table 1: Classification Error Rate (CER%) of speaker identification systems trained on TIMIT (462 spks) and Librispeech (2484 spks) datasets. SincNet outperforms the competing alternatives.

	EER(%)
DNN-MFCC	0.88
CNN-FBANK	0.60
CNN-Raw	0.58
SINCNET	**0.51**

Table 2: Speaker Verification Equal Error Rate (EER%) on Librispeech datasets over different systems. SincNet outperforms the competing alternatives.

between enrolment and test d-vectors is then calculated, and a threshold is then applied on it to reject or accept the speaker. Ten utterances from impostors were randomly selected for each sentence coming from a genuine speaker. To assess our approach on a standard open-set speaker verification task, all the enrolment and test utterances were taken from a speaker pool different from that used for training the speaker-id DNN.

For the sake of completeness, experiments have also been conducted with standard i-vectors. Although a detailed comparison with this technology is out of the scope of this paper, it is worth noting that our best i-vector system achieves an EER=1.1%, rather far from what is achieved with DNN systems. It is well-known in the literature that i-vectors provide competitive performance when more training material is used for each speaker and when longer test sentences are employed [57–59]. Under the challenging conditions faced in this work, neural networks achieve better generalization.

4.2 Speech Recognition

Tab. 3 reports the speech recognition performance obtained by CNN and SincNet using the TIMIT and the DIRHA dataset [41]. To ensure a more accurate comparison between the architectures, five experiments varying the initialization seeds were conducted for each model and corpus. Table 3 thus reports the average speech recognition performance. Standard deviations, not reported here, range between 0.15 and 0.2 for all the experiments.

	TIMIT	DIRHA
CNN-FBANK	18.3	40.1
CNN-Raw waveform	18.3	40.5
SincNet-Raw waveform	**18.0**	**38.2**

Table 3: Speech recognition performance obtained on the TIMIT and DIRHA datasets.

For all the datasets, SincNet outperforms CNNs trained on both standard FBANK and raw waveforms. The latter result confirms the effectiveness of SincNet not only in close-talking scenarios but also in challenging noisy conditions characterized by the presence of both noise and reverberation. As emerged in Sec.2, SincNet is able to effectively tune its filter-bank front-end to better address the characteristics of the noise.

5 Conclusions and Future Work

This paper proposed SincNet, a neural architecture for directly processing waveform audio. Our model, inspired by the way filtering is conducted in digital signal processing, imposes constraints on the filter shapes through efficient parameterization. SincNet has been extensively evaluated on challenging speaker and speech recognition tasks, consistently showing some performance benefits.

Beyond performance improvements, SincNet also significantly improves convergence speed over a standard CNN, is more computationally efficient due to the exploitation of filter symmetry, and it is more interpretable than standard black-box models. Analysis of the SincNet filters, in fact, revealed

that the learned filter-bank is tuned to the specific task addressed by the neural network. In future work, we would like to evaluate SincNet on other popular speaker recognition tasks, such as VoxCeleb. Inspired by the promising results obtained in this paper, in the future we will explore the use of SincNet for supervised and unsupervised speaker/environmental adaptation. Moreover, although this study targeted speaker and speech recognition only, we believe that the proposed approach defines a general paradigm to process time-series and can be applied in numerous other fields.

Acknowledgement

References

[1] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[2] D. Yu and L. Deng. *Automatic Speech Recognition - A Deep Learning Approach*. Springer, 2015.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *Proc. of ICASSP*, pages 4945–4949, 2016.

[4] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proc. of ICML*, pages 1764–1772, 2014.

[5] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proc. of ICLR*, 2015.

[6] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2018.

[7] S. Chakraborty et al. Interpretability of deep learning models: A survey of results. In *Proc. of SmartWorld*, 2017.

[8] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. of ECCV*, 2014.

[9] Q.-S. Zhang and S.-C. Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, Jan 2018.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proc. of ACM SIGKDD*, pages 1135–1144, 2016.

[11] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu. Interpreting CNN Knowledge via an Explanatory Graph. In *Proc. of AAAI*, 2018.

[12] S. Sabour, N. Frosst, and G. E Hinton. Dynamic routing between capsules. In *Proc. of NIPS*, pages 3856–3866. 2017.

[13] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek. Interpreting and explaining deep neural networks for classification of audio signals. *CoRR*, abs/1807.03418, 2018.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[15] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proc. of NIPS*, 2014.

[16] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. Improving speech recognition by revising gated recurrent units. In *Proc. of Interspeech*, 2017.

[17] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102, April 2018.

[18] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, London, UK, UK, 1999. Springer-Verlag.

[19] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *Proc. of ICASSP*, pages 4052–4056, 2014.

[20] F. Richardson, D. A. Reynolds, and N. Dehak. A unified deep neural network for speaker and language recognition. In *Proc. of Interspeech*, pages 1146–1150, 2015.

[21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Proc. of Interspeech*, pages 999–1003, 2017.

[22] C. Zhang, K. Koishida, and J. Hansen. Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(9):1633–1644, 2018.

[23] G. Bhattacharya, J. Alam, and P. Kenny. Deep speaker embeddings for short-duration speaker verification. In *Proc. of Interspeech*, pages 1517–1521, 2017.

[24] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Proc. of Interspech*, 2017.

[25] D. Palaz, M. Magimai-Doss, and R. Collobert. Analysis of CNN-based speech recognition system using raw speech as input. In *Proc. of Interspeech*, 2015.

[26] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals. Learning the speech front-end with raw waveform CLDNNs. In *Proc. of Interspeech*, 2015.

[27] Y. Hoshen, R. Weiss, and K. W. Wilson. Speech acoustic modeling from raw multichannel waveforms. In *Proc. of ICASSP*, 2015.

[28] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior. Speaker localization and microphone spacing invariant acoustic modeling from raw multichannel waveforms. In *Proc. of ASRU*, 2015.

[29] Z. Tüske, P. Golik, R. Schlüter, and H. Ney. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Proc. of Interspeech*, 2014.

[30] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. of ICASSP*, pages 5200–5204, 2016.

[31] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *Arxiv*, 2016.

[32] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *CoRR*, abs/1612.07837, 2016.

[33] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur. Acoustic modelling from the signal domain using CNNs. In *Proc. of Interspeech*, 2016.

[34] H. Dinkel, N. Chen, Y. Qian, and K. Yu. End-to-end spoofing detection with raw waveform CLDNNS. *Proc. of ICASSP*, 2017.

[35] H. Muckenhirn, M. Magimai-Doss, and S. Marcel. Towards directly modeling raw speech signal for speaker verification using CNNs. In *Proc. of ICASSP*, 2018.

[36] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, , and H.-J. Yu. A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. In *Proc. of ICASSP*, 2018.

[37] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu. Avoiding Speaker Overfitting in End-to-End DNNs using Raw Waveform for Text-Independent Speaker Verification. In *Proc. of Interspeech*, 2018.

[38] M. Ravanelli and Y. Bengio. Speaker Recognition from raw waveform with SincNet. In *Proc. of SLT*, 2018.

[39] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM, 1993.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proc. of ICASSP*, pages 5206–5210, 2015.

[41] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo. The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments. In *Proc. of ASRU 2015*, pages 275–282.

[42] M. Ravanelli, P. Svaizer, and M. Omologo. Realistic multi-microphone data simulation for distant speech recognition. In *Proc. of Interspeech*, 2016.

[43] L. R. Rabiner and R. W. Schafer. *Theory and Applications of Digital Speech Processing*. Prentice Hall, NJ, 2011.

[44] S. K. Mitra. *Digital Signal Processing*. McGraw-Hill, 2005.

[45] M. Ravanelli, D. Serdyuk, and Y. Bengio. Twin regularization for online speech recognition. In *Proc. of Interspeech*, 2018.

[46] T. N. Sainath, B. Kingsbury, A. R. Mohamed, and B. Ramabhadran. Learning filter banks within a deep neural network framework. In *Proc. of ASRU*, pages 297–302, 2013.

[47] H. Yu, Z. H. Tan, Y. Zhang, Z. Ma, and J. Guo. DNN Filter Bank Cepstral Coefficients for Spoofing Detection. *IEEE Access*, 5:4779–4787, 2017.

[48] H. Seki, K. Yamamoto, and S. Nakagawa. A deep neural network integrated with filterbank learning for speech recognition. In *Proc. of ICASSP*, pages 5480–5484, 2017.

[49] V. Papyan, Y. Romano, and M. Elad. Convolutional neural networks analyzed via convolutional sparse coding. *Journal of Machine Learning Research*, 18:83:1–83:52, 2017.

[50] S. Mallat. Understanding deep convolutional networks. *CoRR*, abs/1601.04920, 2016.

[51] D. Palaz and R. Magimai-Doss, M.and Collobert. End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition. 2016.

[52] H. Muckenhirn, M. Magimai-Doss, and S. Marcel. On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs. In *Proc. of Interspeech*, 2018.

[53] M. Ravanelli. *Deep learning for Distant Speech Recognition*. PhD Thesis, Unitn, 2017.

[54] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. A network of deep neural networks for distant speech recognition. In *Proc. of ICASSP*, pages 4880–4884, 2017.

[55] M. Ravanelli and M. Omologo. Contaminated speech training methods for robust DNN-HMM distant speech recognition. In *Proc. of Interspeech 2015*, pages 756–760.

[56] M. Ravanelli, T. Parcollet, and Y. Bengio. The PyTorch-Kaldi Speech Recognition Toolkit. In *arXiv:1811.07453*, 2018.

[57] A. K. Sarkar, D Matrouf, P.M. Bousquet, and J.F. Bonastre. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *Proc. of Interspeech*, pages 2662–2665, 2012.

[58] R. Travadi, M. Van Segbroeck, and S. Narayanan. Modified-prior i-Vector Estimation for Language Identification of Short Duration Utterances. In *Proc. of Interspeech*, pages 3037–3041, 2014.

[59] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason. i-vector based speaker recognition on short utterances. In *Proc. of Interspeech*, pages 2341–2344, 2011.

[60] M. Matassoni, R. Astudillo, A. Katsamanis, and M. Ravanelli. The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In *Proc. of Interspeech 2014*, pages 1616–1617.

[61] E. Zwyssig, M. Ravanelli, P. Svaizer, and M. Omologo. A multi-channel corpus for distant-speech interaction in presence of known interferences. In *Proc. of ICASSP 2015*, pages 4480–4484.

[62] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos. The DIRHA simulated corpus. In *Proc. of LREC 2014*, pages 2629–2634.

[63] Douglas P. and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*, Proc. of HLT, pages 357–362, 1992.

[64] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo. Impulse response estimation for robust speech recognition in a reverberant environment. In *Proc. of EUSIPCO 2012*.

[65] M. Ravanelli and M. Omologo. On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *Proc. of Interspeech 2014*, pages 1028–1032.

[66] J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[67] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*, pages 448–456, 2015.

[68] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. Batch-normalized joint training for dnn-based distant speech recognition. In *Proc. of SLT*, 2016.

[69] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. of ICML*, 2013.

[70] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of AISTATS*, pages 249–256, 2010.

[71] D. Povey et al. The Kaldi Speech Recognition Toolkit. In *Proc. of ASRU*, 2011.

[72] A. Larcher, K. A. Lee, and S. Meignier. An extensible speaker identification sidekit in python. In *Proc. of ICASSP*, pages 5095–5099, 2016.

Appendix

Corpora

To provide experimental evidence on datasets characterized by different numbers of speakers, this paper considers the TIMIT (462 spks, *train* chunk) [39] and Librispeech (2484 spks) [40] corpora. For speaker verification experiments, non-speech intervals at the beginning and end of each sentence were removed. Moreover, the Librispeech sentences with internal silences lasting more than 125 ms were split into multiple chunks. 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. For the latter dataset, five sentences for each speaker were used for training, while the remaining three were used for test. For the Librispeech corpus, 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. To evaluate the performance in a challenging distant-talking scenario, speech recognition experiments have also considered the DIRHA dataset [41]. This corpus, similarly to the other DIRHA corpora [60, 61], has been developed in the context of the DIRHA project [62] and is based on WSJ sentences [63] recorded in a domestic environment. Training is based on contaminating WSJ-5k utterances with realistic impulse responses [64, 65], while the test phase test phase consists of 409 WSJ sentences recorded by native speakers in a domestic environment (the average SNR is 10 dB).

SincNet Setup

The waveform of each speech sentence was split into chunks of 200 ms (with 10 ms overlap), which were fed into the SincNet architecture. The first layer performs sinc-based convolutions as described in Sec. 2, using 80 filters of length $L = 251$ samples. The architecture then employs two standard convolutional layers, both using 60 filters of length 5. Layer normalization [66] was used for both the input samples and for all convolutional layers (including the SincNet input layer). Next, three fully-connected layers composed of 2048 neurons and normalized with batch normalization [67, 68] were applied. All hidden layers use leaky-ReLU [69] non-linearities. The parameters of the sinc-layer were initialized using mel-scale cutoff frequencies, while the rest of the network was initialized with the well-known "Glorot" initialization scheme [70]. Frame-level speaker and phoneme classifications were obtained by applying a softmax classifier, providing a set of posterior probabilities over the targets. For speaker-id, a sentence-level classification was simply derived by averaging the frame predictions and voting for the speaker which maximizes the average posterior. Training used the RMSprop optimizer, with a learning rate $lr = 0.001$, $\alpha = 0.95$, $\epsilon = 10^-7$, and minibatches of size 128. All the hyper-parameters of the architecture were tuned on TIMIT, then inherited for Librispeech as well. The speaker verification system was derived from the speaker-id neural network using the *d-vector* approach [19, 24], which relies on the output of the last hidden layer and computes the cosine distance between test and the claimed speaker d-vectors. Ten utterances from impostors were randomly selected for each sentence coming from a genuine speaker. Note that to assess our approach on a standard open-set speaker-id task, all the impostors were taken from a speaker pool different from that used for training the speaker-id DNN.

Baseline Setups

We compared SincNet with several alternative systems. First, we considered a standard CNN fed by the raw waveform. This network is based on the same architecture as SincNet, but replacing the sinc-based convolution with a standard one.

A comparison with popular hand-crafted features was also performed. To this end, we computed 39 MFCCs (13 static+Δ+$\Delta\Delta$) and 40 FBANKs using the Kaldi toolkit [71]. These features, computed every 25 ms with 10 ms overlap, were gathered to form a context window of approximately 200 ms (i.e., a context similar to that of the considered waveform-based neural network). A CNN was used for FBANK features, while a Multi-Layer Perceptron (MLP) was used for MFCCs. Note that CNNs exploit local correlation across features and cannot be effectively used with uncorrelated MFCC features. Layer normalization was used for the FBANK network, while batch normalization was employed for the MFCC one. The hyper-parameters of these networks were also tuned using the aforementioned approach.

For speaker verification experiments, we also considered an i-vector baseline. The i-vector system was implemented with the SIDEKIT toolkit [72]. The GMM-UBM model, the Total Variability (TV) matrix, and the Probabilistic Linear Discriminant Analysis (PLDA) were trained on the Librispeech data (avoiding test and enrollment sentences). GMM-UBM was composed of 2048 Gaussians, and the rank of the TV and PLDA eigenvoice matrix was 400. The enrollment and test phase is conducted on Librispeech using the same set of speech segments used for DNN experiments.