# CGCNN: COMPLEX GABOR CONVOLUTIONAL NEURAL NETWORK ON RAW SPEECH

*Paul-Gauthier Noé*[1], *Titouan Parcollet*[1,2], *Mohamed Morchid*[1]

[1]LIA, Université d'Avignon, France
[2]University of Oxford, UK

## ABSTRACT

Convolutional Neural Networks (CNN) have been used in Automatic Speech Recognition (ASR) to learn representations directly from the raw signal instead of hand-crafted acoustic features, providing a richer and lossless input signal. Recent researches propose to inject prior acoustic knowledge to the first convolutional layer by integrating the shape of the impulse responses in order to increase both the interpretability of the learnt acoustic model, and its performances. We propose to combine the complex Gabor filter with complex-valued deep neural networks to replace usual CNN weights kernels, to fully take advantage of its optimal time-frequency resolution and of the complex domain. The conducted experiments on the TIMIT phoneme recognition task shows that the proposed approach reaches top-of-the-line performances while remaining interpretable.

***Index Terms*—** SincNet, complex neural networks, Gabor filters, speech recognition.

## 1. INTRODUCTION

The task of Automatic Speech Recognition (ASR) is far from being solved and represents an active research field [1, 2, 3, 4]. More precisely ASR systems are either hybrid DNN-HMM, with multiple sub-blocks trained separately [5, 6], or End-to-End (E2E), with various Neural Networks (NN) learnt accordingly to a joint training procedure [7, 8, 9]. Despite numerous architectures and training investigations, the input feature representation remains mostly unchanged with traditional handcrafted acoustic features such as Mel-filter-banks.

Nonetheless, another field of speech recognition recently obtained promising performances while operating at the raw waveform level [10, 11, 12, 13]. In these works, Convolutional Neural Networks (CNN) have been used directly on the raw speech signal to rapidly consume the large number of data points (*e.g.* $16,000$ per second for an audio sampled at 16kHz). Unfortunately, and as demonstrated in [14, 15] traditional CNN kernel filters are not efficient at learning common acoustic features due to the lack of constraint on neural parameters. To alleviate the latter issue, the authors proposed a novel convolutional layer named SincNet incorporating prior speech processing knowledge in an efficient and interpretable operation. More precisely, SincNet filters are initialized and constrained following specific acoustic filters to produce easily interpretable filtering of the input waveform [14]. As a matter of fact, processing the raw waveform theoretically has numerous advantages including a higher richness of the input signal alongside with a lossless natural representation of the features. Recently, other filters have been proposed for SincNet to further increase its performances [16]. In the latter, Gaussian filters have been suggested without motivations on their major properties in terms of time and frequency localization [17]. Furthermore, the authors only proposed to consider the real-part to feed the standard SincNet architecture. We propose to extend this work to the complete complex Gabor filters to take advantage of both the complex-valued representation and a better time-frequency resolution due to the Gaussian shape. Other recent works have used Gabor filters instead of the triangular ones to operate over FBANK acoustic features [18]. In this case, cutoff frequencies of the filters are fixed on the Mel-scale and require fine tuning. Gabor filters have also been used to initialize end-to-end filterbanks learning [19]. In the latter works, the filter frequency positions are fixed. We propose to learn all these parameters jointly to the rest of the neural network model to allow the Gabor filters to perfectly match the considered task.

Nonetheless, complex Gabor filters produce a complex-valued filtered signal that must be processed with a dedicated complex-valued model to further fully exploit this input representation. Fortunately, Complex-Valued Convolutional Neural Networks (CVCNN) have been recently introduced with applications to image and speech processing [20, 21]. Thus, we propose to enhance the original real-valued SincNet with the proposed complex Gabor filters, and to process the filtered signal with a CVCNN respectful of the complex algebra. Contributions of the paper are summarized as:

1. The optimal time and frequency localization compromise of Gabor filters is presented (Section 2).

2. SincNet [14] is extended to the complex-valued space with a complex Gabor filter and CVCNNs. The model is also released for reproducibility [1] (Section 3).

[1]https://github.com/NOEPG/pytorch-kaldi

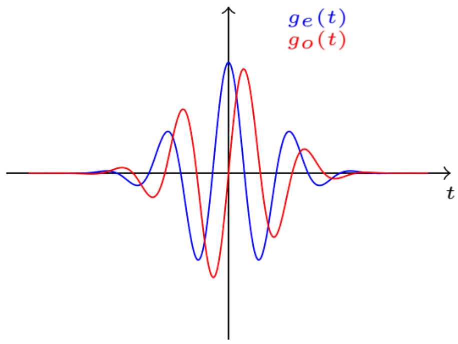


**Fig. 1**: Illustration of the real ($g_e$) and imaginary ($g_o$) parts of the complex Gabor filter impulse response.

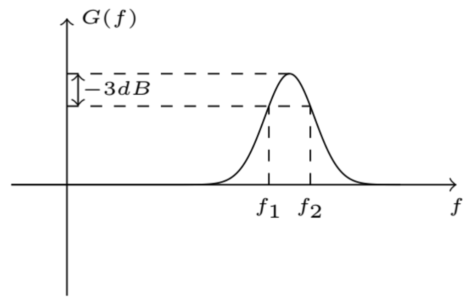


**Fig. 2**: Illustration of the frequency response of the complex Gabor filter with $f_1$ and $f_2$ respectively the low and high cut-off frequencies.

3. Evaluate and analyse the proposed complex gabor-based SincNet on a phoneme recognition task with the TIMIT dataset (Section 4).

The conducted experiments show that our approach reaches promising performances with a more natural and efficient filtered representation of the raw waveform, while retaining the interpretability property of the original SincNet.

## 2. UNCERTAINTY PRINCIPLE AND GABOR FILTERING

This section motivates the use of Gabor filters to replace the standard sinc approach of SincNet [14].

First, it is important to notice that time-frequency resolution is limited by the uncertainty principle. More precisely, the product of the time and frequency spread of a filter is superior to a constant. Given any function $y \in L^2(\mathbb{R})$ and its Fourier transform $\hat{y}$, their time $e_y$ and frequency $e_{\hat{y}}$ localizations are respectively:

$$\begin{aligned} e_y &= \frac{1}{||y||^2}\int_{-\infty}^{+\infty} t|y(t)|^2 dt. \\ e_{\hat{y}} &= \frac{1}{||\hat{y}||^2}\int_{-\infty}^{+\infty} f|\hat{y}(f)|^2 df. \end{aligned} \tag{1}$$

Then, their time and frequency spreads are defined by their variance $v_y$ and $v_{\hat{y}}$:

$$\begin{aligned} v_y &= \frac{1}{||y||^2}\int_{-\infty}^{+\infty} (t-e_y)^2|y(t)|^2 dt. \\ v_{\hat{y}} &= \frac{1}{||\hat{y}||^2}\int_{-\infty}^{+\infty} (f-e_{\hat{y}})^2|\hat{y}(f)|^2 df. \end{aligned} \tag{2}$$

The uncertainty principle states that $v_y v_{\hat{y}} \geq \frac{1}{16\pi^2}$ [22]. Thus a function can not have an infinite narrow localization both in time and frequency domain. However, this inequality becomes an equality for Gaussian shape functions such as Gabor filter [22] and thus provides the best time-frequency resolution compromise.

In particular, the complex impulse response of the Gabor filter is defined as follows:

$$\begin{aligned} g(t) &= w_\sigma(t)e^{i2\pi f_0 t}, \\ &= g_e(t) + i g_o(t), \\ w_\sigma(t) &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{t^2}{2\sigma^2}}, \end{aligned} \tag{3}$$

with $\sigma$ the standard deviation of the temporal Gaussian window $w_\sigma(t)$, $f_0$ the center frequency of the filter and $g_e(t)$ and $g_o(t)$ the real and imaginary part of the impulse response respectively (Figure 1).

The Gabor filter has the following Gaussian frequency response:

$$G(f) = e^{-2\pi^2\sigma^2(f-f_0)^2}. \tag{4}$$

For the sake of interpretability, it is feasible to express $\sigma$ and $f_0$ in terms of the cutoff frequencies:

$$\sigma = \frac{A}{\pi(f_2 - f_1)}, \; f_0 = \frac{f_1 + f_2}{2}, \tag{5}$$

with $f_1$ and $f_2$ the low and high $(-3dB)$ cutoff frequencies respectively, and $A = \sqrt{\frac{3ln(10)}{10}}$ a constant. The Gabor frequency response is depicted in Figure.2

## 3. CONVOLUTIONAL NEURAL NETWORKS ON THE RAW WAVEFORM

This section details the proposed complex-valued model to directly operate of the raw speech waveform. First, SincNet is presented (Section 3.1). Then, we introduce the complex Gabor filtering alongside with the complex-valued architecture (Section 3.2).

### 3.1. SincNet

Traditional parametric CNNs operate over the raw waveform by performing multiple time-domain convolutions between the input signal and a certain finite impulse response. Therefore, SincNet [14] proposes to use a sinc function to obtain the impulse response $h(t)$ at time $t$ of a CNN as:

$$h(t) = 2f_2 sinc(2\pi f_2 t) - 2f_1 sinc(2\pi f_1 t), \tag{6}$$

with sinc($x$)=sin($x$)$/x$, $f_1$ and $f_2$ the two cutoff frequencies and $f_2 \geq f_1$. Such filter has a rectangular frequency response described as:

$$H(f) = rect_{2f_2}(f) - rect_{2f_1}(f), \tag{7}$$

with $rect_w(.)$ the rectangular function of width $w$ centered in $0$. Finally, cutoff frequencies are the two only trainable parameters reducing drastically the number of neural parameters compared to CNNs. Then, multiple CNN layers are stacked to reduce the signal dimension, before being fed into a classifier.

### 3.2. Complex Gabor Filtering with CVCNN

We propose to replace the sinc based filter of Eq. 6 with a complex-valued Gabor one, for a better time-frequency resolution. Therefore, let $g(t)$ be the complex Gabor filter impulse response with $g_e(t)$ and $g_o(t)$ the corresponding real and imaginary parts. The convolution of an input signal $x(t)$ with $g(t)$ at time $t$ is written as:

$$x(t) * g(t) = x(t) * g_e(t) + i(x(t) * g_o(t)). \tag{8}$$

The obtained filtered signal is of the form $z = a + \mathbf{i}b$ lying on the complex plane, and is an approximation of analytic signal (*i.e.* without negative frequencies) [23]. We propose to fully take this complex representation into consideration by further processing it with complex-valued neural networks layers only.

Therefore, the extracted features are then fed into multiple CVCNNs followed by complex layer-normalization to reduce the signal dimension. Finally, the reduced complex features are fed into a complex fully-connected classifier (CVDNN). In the same manner as SincNet [2], the two cutoff frequencies of the filters are the only parameters learnt in the first layer, thus reducing the number of neural parameters.

## 4. EXPERIMENTS

The proposed approach is evaluated on the TIMIT phoneme recognition task [24]. The latter dataset is composed of a standard $462$-speaker training dataset, a 50-speakers development dataset and a core test dataset of $192$ sentences. During the experiments, the dialect sentences of the training dataset are removed.

### 4.1. Models Achitecture

Our model starts with four CVCNNs with respectively $128$, $60$, $60$ and $60$ complex filters of kernel size $129$, $5$, $5$, $3$ with Complex Gabor filters used at the first layer only. Then, a Complex-Valued Multilayer Perceptrons (CMLP) composed of $5$ layers with $1024$ complex hidden unit is added. The ReLU activation function [25] is used accross all the layers. Complex layer-normalization is done on each layer of the CVCNN, while complex batch-normalization is applied on the CMLP layers. Maxpooling is applied on the convolutional part to further reduce the dimension of the signal. Finally the $softmax$ function is applied to obtain the posterior probability over the HMM states. Monophone regularization is also used to smooth the training [26].

We propose to compare this approach with a real-valued equivalent baseline. Thus, the input layer has the same number of filters. Then, the dimensions of the other layers are increased to obtain a comparable number of neural parameters $20M$.

Both models are fed with $200$ms speech signal chunks with an overlap of $10$ms. They are trained using standard stochastic gradient descent with a batch size of $128$ for $20$ epochs. The learning rate is annealed with respect to a certain threshold on the validation loss evolution, alleviating the risk of overfitting and ensuring an optimal convergence for both models. A dropout probability of $0.15$ is also set for each layer except the classification ones.

Models and experiments are run within the Pytorch-Kaldi toolkit [2].

### 4.2. Results and Discussions

Models are evaluated by averaging the Phone Error Rate (PER) observed on the validation and test datasets for $5$ runs with the TIMIT phoneme recognition task [24]. Results are shown in Table 1.

First, it is worth underlying that the obtained results are comparable to state-of-the-art performances on the TIMIT task from the raw waveform. Then, while both real and complex-valued Gabor models obtain similar averaged performances, a best PER of $16.7\%$ is obtained with the complex alternative compared to $16.9\%$ for the real-valued one. In fact, both real and complex filters have the same frequency response in the positive domain. But using complex quadratic filters that produce analytic signal for which the complex Gabor filtered signal is an approximation could help for instantaneous frequency estimation [23] and preserves the phase information that can be useful for other tasks such as speaker recognition.

Then, Figure 3 illustrates the learnt cutoff frequencies. As expected, distributions are similar along a straight line for both models. In fact, both models roughly learn the same couples of cutoff frequencies due to an equivalent frequency response shape.

**Table 1**: Results obtained with different ASR systems on the TIMIT phoneme recognition tasks. Valid. denotes the validation dataset, and CTC the Connectionist Temporal Classification training scheme. Results are expressed in Phoneme Error Rate (*i.e.* lower is better).

| Model | Valid.% | Avg. Test% | Best Test% |
|---|---|---|---|
| Gabor-CNN-CTC [18] | - | 18.8 | 18.5 |
| SincNet [2] | - | 17.2 | - |
| GaborReal | 15.2 | 17.2 | 16.9 |
| GaborComplex | 15.2 | 17.1 | 16.7 |

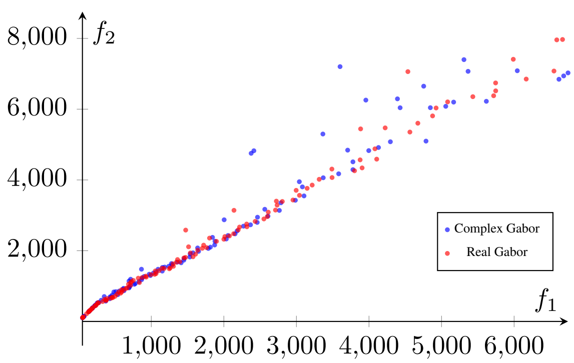


**Fig. 3**: Illustration of the cutoff frequencies learnt with the Real and the Complex Gabor CNN.

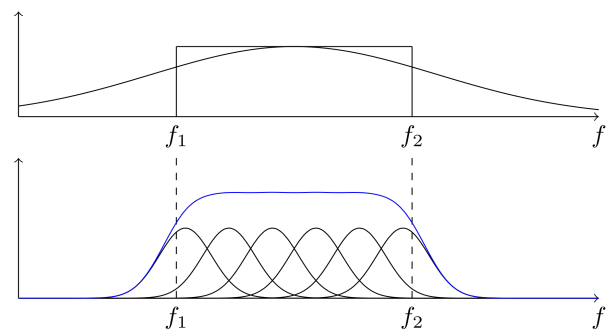


**Fig. 4**: Illustration large band filters phenomenon. The top row shows how Gaussian filters can be flattened when the bandwidth is too large. The bottom row illustrates how smaller bandwidth Gaussian filters can be combined to obtain a better large band filter.

Another drawback implied by Gaussian form functions could be the learning of large bandwidths. More precisely, the Gaussian shape is not ideal for large bandwidth filtering as depicted in figure 4. Indeed the frequency response tends to be flat in comparison to a rectangular filter. It is feasible to alleviate this issue by enabling the model to combine smaller band filters (Figure 4). However, and as expected, filters with larger bands are more distributed in the higher frequencies.

## 5. CONCLUSION

**Summary.** This paper introduces and releases a complex-valued and optimal time-frequency resolution alternative to the SincNet architecture. It is based on a complex Gabor filter learning process for automatic speech recognition. The conducted experiments show that the proposed approach is able to produce results comparable with state-of-the-art systems while operating on the raw waveform.

**Future Work.** The base of the Gabor filters in non-orthogonal leading to redundancy in the filtered signals. It is therefore crucial to investigate other filters to produce an orthogonal base for better performances [22]. Furthermore, the interest of the phase still has to be established and this model must be evaluated on tasks relying on the phase and local information preserved by complex Gabor filters.

## 6. ACKNOWLEDGEMENTS


This work was supported by the JST-ANR VoicePersonae Project and by the Engineering and Physical Sciences Research Council (EPSRC) under Grant: MOA (EP/S001530/).


## 7. REFERENCES


[1] E. Battenberg C. Case J. Casper B. Catanzaro J. Chen M. Chrzanowski A. Coates G. Diamos E. Elsen J. Engel L. Fan C. Fougner T. Han A. Hannun B. Jun P. LeGresley L. Lin S. Narang A. Ng S. Ozair R. Prenger J. Raiman S. Satheesh D. Seetapun S. Sengupta Y. Wang Z. Wang C. Wang B. Xiao D. Yogatama J. Zhan Z. Zhu D. Amodei, R. Anubhai, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595v1*, 2015.

[2] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6465–6469.

[3] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner,

N. Chen, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[4] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O Kuchaiev, J. Cohen, H. Nguyen, and R.T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[6] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Improving speech recognition by revising gated recurrent units," *Proc. Interspeech 2017*, pp. 1308–1312, 2017.

[7] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[8] Y Zhang, M Pezeshki, P. Brakel, S. Zhang, C. Laurent Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[9] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[10] D. Palaz, R. Collobert, and M.M Doss, "End-to-end phoneme sequence recognition using convolutional neural networks," *arXiv preprint arXiv:1312.2137*, 2013.

[11] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *Fifteenth annual conference of the international speech communication association*, 2014, pp. 890–894.

[12] Y. Hoshen, R. Weiss, and K. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4624–4628.

[13] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," in *Interspeech 2018*, 2018, pp. 781–785.

[14] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. of Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[15] M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1812.05920*, 2018.

[16] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable cnns with parametric modulated kernel-based filters," in *Proc. of Interspeech*. ISCA, 2019, pp. 3480–3484.

[17] D. Gabor, "Theory of communication," in *Journal of the Institute of Electrical Engineers*, 1946, vol. 93, pp. 429–457.

[18] S. Robertson, G. Penn, and Y. Wang, "Improving speech recognition with drop-in replacements for f-bank features," in *Proc. of SLSP 2019*, 2019, pp. 210–222.

[19] I. Kokkinos T. Schatz G. Synnaeve E. Dupoux N. Zeghidour, N. Usunier, "Learning filterbanks from raw speech for phoneme recognition," in *Proc. of ICASSP*. IEEE, 2018, pp. 5509–5513.

[20] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J.F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J Pal, "Deep complex networks," in *Proc. of ICLR 2018*, 2018.

[21] J. Huh A. Kim J.W. Ha K. Lee H.S. Choi, J.H. Kim, "Phase-aware speech enhancement with deep complex u-net," in *Proc. of ICLR 2019*, 2019.

[22] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Elsevier Science, 2008.

[23] G. Sommer, *Geometric Computing with Clifford Algebras*, Springer, Berlin, Heidelberg, 2001.

[24] J. S. Garofolo, L. F. Lamel, and W. M. Fisher, *Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1*, NASA STI/Recon technical report, vol. 93, 1993.

[25] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[26] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *Proc. of ICCASP*. IEEE, 2015, pp. 4290–4294.