# TIMESCALENET : A MULTIRESOLUTION APPROACH FOR RAW AUDIO RECOGNITION

*Éric Bavu*[⋆] *Aro Ramamonjy*[⋆] *Hadrien Pujol*[⋆] *Alexandre Garcia*[⋆]

[⋆] Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC),
Conservatoire national des arts et métiers (Cnam), 292 rue Saint-Martin, 75003 Paris, France.

## ABSTRACT

In recent years, the use of Deep Learning techniques in audio signal processing has led the scientific community to develop machine learning strategies that allow to build efficient representations from raw waveforms for machine hearing tasks. In the present paper, we show the benefit of a multi-resolution approach : TimeScaleNet aims at learning an efficient representation of a sound, by learning time dependencies both at the sample level and at the frame level. At the sample level, TimeScaleNet's architecture introduces a new form of recurrent neural layer that acts as a learnable passband biquadratic digital IIR filterbank and self-adapts to the specific recognition task and dataset, with a large receptive field and very few learnable parameters. The obtained frame-level feature map is then processed using a residual network of depthwise separable atrous convolutions. This second scale of analysis allows to encode the time fluctuations at the frame timescale, in different learnt pooled frequency bands. In the present paper, TimeScaleNet is tested using the Speech Commands Dataset. We report a very high mean accuracy of $94.87 \pm 0.24\%$ (macro averaged F1-score : $94.9 \pm 0.24\%$) for this particular task.

***Index Terms*—** Machine hearing, Learnable Biquadratic filters, Deep Learning, Time domain modelling

## 1. INTRODUCTION

Machine hearing can involve the use of hand-crafted features, time-frequency representations, or raw audio. The use of the latter representation has emerged as an active area of research in the last few years : the unprocessed, time-domain audio signals contain all the information to be extracted for machine hearing tasks. In previously published studies [1–6], the models mostly used large filters, which can model passband filters approximating cochlear filter estimates. These studies, along with recent advances in machine learning for one-dimensional signals [7–9] have motivated the present work, which aims at showing the benefits of a multi-resolution approach for machine hearing. The proposed approach avoids using large convolutional kernels, by introducing a new form of recurrent neural cell, directly inspired by IIR digital signal processing.

## 2. METHODS

TimeScaleNet's architecture is detailed in 2.1. This architecture can be split in two subnets, acting at two different timescales. The architecture and the detailed implementation of these two subnets are explained in 2.2 and 2.3. The training procedure is also detailed in 2.4.

### 2.1. Global neural network architecture

As shown on Fig. 1, the first subnet of TimeScaleNet's architecture, BiquadNet (see 2.2), acts at the sample level, and aims at encoding the information for time scales in the range of [100 $\mu$s ; 20 ms], thus allowing to compute a time-frequency-like representation, that is fed to the next subnet of our architecture. The proposed "biquadratic" RNN filter can be thought as a set of infinite impulse-response (IIR) filters, expressed as a biquadratic filterbank [10] with learnable coefficients, that self-adapts to the audio dataset that has to be classified. This strategy allows a computationally-efficient IIR bandpass filtering using only two learnable parameters for arbitrarily long receptive fields, rather than 1-dimensional convolutional neural networks with wide kernels reported in previous studies [4, 6, 11].

The obtained time-frequency-like representation at the output of BiquadNet is then fed to the second subnet, referred in the following as "FrameNet" (see 2.3), because it acts at the frame level, in order to efficiently encode the time fluctations in the range of [20 ms ; 200 ms]. This second scale of analysis aims at extracting the relevant relationships between time fluctuations in different learnt pooled frequency channels, with a large receptive field. For this purpose, we propose the use of residual networks of one-dimensional depthwise separable atrous convolutions, which allow to operate on channel-wise frames in a computationally efficient way.

The output of FrameNet is then flattened, and fed to two full-connected layers with Selu activations, in order to compute a vector of dimension $N_{\text{classes}}$ representing the probability of belonging to the classes of the dataset.

### 2.2. BiquadNet : raw waveform processing

From a machine-learning point of view, the first layer of BiquadNet is a non-conventional recurrent neural network cell.

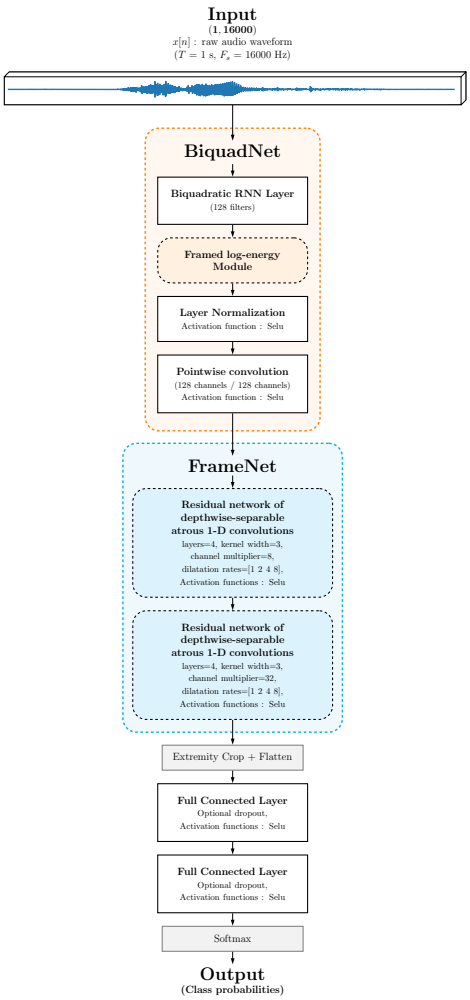


**Fig. 1**. Schematical representation of TimeScaleNet

From a digital signal processing however, this RNN cell is directly derived from the widely used infinite impulse response (IIR) biquadratic filters, which are very efficient types of filters to implement, because they require less computation and memory than FIR filters in order to perform similar filtering operations. It is also useful to note that passband biquadratic filters (also referred as two-poles two-zeros in the litterature) have been demonstrated to be good numerical models of auditory filterbanks [12,13]. In the present work, we implemented a bidirectional biquadratic RNN cell, which allows to achieve forward-backward filtering [14], in order to perform a perfect zero-phase filtering in the time domain. In order to prevent a potential numerical instability we only use second-order IIR filters. This is the main reason why most digital signal processors implement stacks of biquadratic IIR filters. This kind of topology can be easily transposed to machine learning, where deep neural network topologies often use stacking of similar layers. Using the $Z$-transform, biquadratic filters exhibit two zeros and two poles, and can be represented as in Figure 2 :

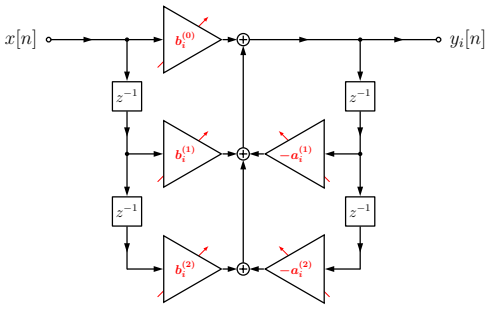


**Fig. 2**. Flow graph of the learnable biquadratic IIR filters

Since we aim at obtaining a "time-frequency"-like representation at the output of BiquadNet, we restrict the possible values of the coefficients of the learnable IIR filterbank to correspond to passband versions of a biquadratic IIR filter. This allows to simplify the stability properties of the learnt filters and ensures that the learnt biquadratic filters are stable, since $a^{(1)}$ and $a^{(2)}$ are constrained to stay inside the "stability triangle" during the whole learning process. Each biquadratic bandpass filter of the learnable filterbank can be fully determined using only two parameters, $K^{(i)} = \tan\left(\pi f_c^{(i)}/f_s\right)$ and $Q^{(i)}$, where $f_s$ is the sample frequency, $f_c^{(i)}$ is the central frequency of the i$^{\text{th}}$ bandpass filter, and $Q^{(i)}$ is the quality factor of the i$^{\text{th}}$ bandpass filter . The five coefficients represented in Figure 2 can be expressed directly using these two meaningful variables $K^{(i)}$ and $Q^{(i)}$, that are chosen to be the learnable variables in TimeScaleNet. The corresponding custom RNN cell has been implemented using high order operations of the Tensorflow open source software library that allow to recursively scan functions over arbitrarily long sequences and to unfold dynamically the computational graph at runtime. This implementation is compatible with a back-propagation-through-time process, in order to compute the derivative chain rule and to update the neural network parameters at each iterations of the machine learning process. The expression of the custom biquadratic bidirectional RNN is fully differentiable, which allows to be compatible with the proposed machine learning approach for audio recognition, while being directly linked to standard digital audio signal processing approaches.

As shown on Figure 1, the 128 time-domain outputs of the biquadratic RNN layer are then fed to a deterministic module that allows to compute a framed log-energy, in order to obtain a time-frequency-like representation, with overlapping windows of 23.2 ms and a stride of 5.8 ms. This non-learnable module consists in the computation of a stabilized logarithmic compression of a sliding mean quadratic value over successive overlapping timeframes. This module is followed by layer normalization [15], which allows to compute layer-wise

statistics and to normalize the Selu [16] nonlinear activations across all summed inputs within the layer, instead of within the batch. The features are then pooled across the whole frequency channels using pointwise convolutions, in order to obtain a representation that has the ability to encode well phonemes such as vowels formants and consonants, by aggregating relevant learnt frequency channels together.

### 2.3. FrameNet : large-scale time relationship learning

FrameNet acts at the time frame level, in order to efficiently encode the relevant relationships between time fluctuations in different pooled frequency channels, with a large time receptive field, thanks to one-dimensional atrous convolutions [7–9]. The stacked residual dilated convolutions therefore allow the network to operate on multiple time scales in the range of $[20\text{ms}; 200\text{ms}]$ without impacting too much the computational efficiency (Figure 3).

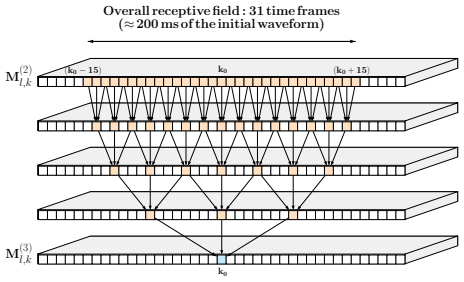


**Fig. 3**. Schematics of one of the two stacks of depthwise separable atrous layers used in FrameNet, from data point of view. Only the depthwise convolution is shown here, with arrows showing the frame indexes involved in dilated convolutions for the computation of the output at frame index $k_0$.

The convolutions are performed independently over every pooled channel (depthwise separable convolutions). This approach has been motivated by preliminary analysis of the energy fluctuations in different frequency channels using classical spectrogram representations. These computed depthwise convolutions are then projected onto a new channel space for each layer using a pointwise convolution, and residual connections are used between each layer, in order to offer increased representation power, by circumventing some of the learning difficulties introduced by deep layers.

### 2.4. Training procedure

In our experiments, TimeScaleNet is trained with one-hot encoded labels, therefore allowing to compute the cross-entropy loss between estimated labels and ground truth labels. The learning and backpropagation of errors through the neural network is optimized using the Adaptive Moment Estimation (Adam) [17] algorithm. The models have been implemented and tested using the Tensorflow open source software library, and computations were carried out on four Nvidia GTX 1080Ti GPU cards, using mini-batches of 70 raw waveforms for each training steps. All the weights involved in layers followed with Selu activations were initialized using the He initialization [18]. $K^{(i)}$ and $Q^{(i)}$ were initialized using clipped random values matching the range of the equivalent rectangular bandwidth cochlear model introduced by Patterson [19], for central frequencies spanning from $40$ Hz to $f_s/2.1$.

### 2.5. Dataset and evaluation metrics

In the present paper, TimeScaleNet is evaluated for raw audio recognition, using the publicly available Google speech commands dataset v2 [20]. This dataset has recently served a competition hosted by Kaggle, which consisted in recognizing the ten words “Yes”, “No”, “Up”, “Down”, “Left”, “Right”, “On”,“Off”, “Stop”, and “Go” along with the “silence” class (*i.e.* no word spoken) and “unknown” class, which is randomly sampled from the remaining 25 keywords from the dataset. The dataset is split into training, validation and test sets in the ratio of 80:10:10 while making sure that the audio clips from the same person stays in the same set, using the exact procedure detailed by the maintainer of the dataset in [20]. In order to analyze precisely the performances of the proposed TimeScaleNet for the task of supervised multi-class classification, several evaluation metrics will be used in the following : The class recall $R_i$, the class precision $P_i$, and their macro-averaged versions $\text{R}_M$ and $\text{P}_M$. The macro-averaged $F_1$ score is also derived, in order to evaluate the relations between data’s positive labels and those given by the classifier, which allow full understanding of the overall classification task achieved by the neural network.

## 3. RESULTS AND DISCUSSION

For the Speech Commands Dataset, the learning process has been performed using TimeScaleNet during 45 epochs, without dropout regularization. These 45 epochs correspond to a total of 1200 hours of audio waveforms processed by the proposed model. Table 1 shows the obtained evaluation metrics on the Speech Commands dataset.

The evaluation metrics shown on Table 1 show that for speech recognition, TimeScaleNet appears to classify the 12 classes with a very high accuracy ($94.87\%$ for the evaluation set, $94.78\%$ for the testing set, after 45 epochs of learning), with a very good homogeneity for all the classes.

For reference, we also evaluated the performances of TimeScaleNet on the Speech Commands dataset with a frozen BiquadNet, using a deterministic (non-learnable) biquadratic filterbank matching the Patterson’s cochlear model with Glasberg and Moore parameters, which achieved $92.4\%$ accuracy over the testing set. A similar experiment has also been per-

**Table 1**. Evaluation metrics obtained after convergence (45 epochs of learning), for the Speech Commands Dataset [20].

| Data | Cardinality | Accuracy | Precision$_M$ | Recall$_M$ | $F_{1,M}$ |
|---|---|---|---|---|---|
| Speech Evaluation Set | 4916 | $\mathbf{94.87 \pm 0.24}\%$ | $\mathbf{94.91 \pm 0.22}\%$ | $\mathbf{94.88 \pm 0.26}\%$ | $\mathbf{94.9 \pm 0.24}\%$ |
| Speech Testing Set | 5157 | $\mathbf{94.78 \pm 0.26}\%$ | $\mathbf{94.87 \pm 0.25}\%$ | $\mathbf{94.87 \pm 0.25}\%$ | $\mathbf{94.87 \pm 0.25}\%$ |

formed using a log-mel-spectrogram as an input to FrameNet, which achieved $89.7\%$ accuracy over the testing set. These two preliminary experiments mainly motivated the development of the BiquadNet part of TimeScaleNet, because this time domain approach allows to achieve a significant performance boost over handcrafted time-frequency features representations. The $94.78\%$ accuracy achieved on the testing set using the proposed TimeScaleNet matches some of the highest values found in [21]. To the best authors knowledge, the only published models that significantly outperforms TimeScaleNet on this particular dataset are DS-CNN in [21] and $res15$ [22], which exhibits the best results to date with a mean accuracy of 95.8 $\%$.

### 3.1. Analysis of the learnt representation from raw waveforms using BiquadNet

In this subsection, we analyse the variables learnt by BiquadNet, in order to give further insight on the learning process involved. The architecture of BiquadNet has been specifically developed to automatically build a 2D map that can be interpreted as an energy-like representation in 128 pooled frequency channels. Figure 4 shows the computed magnitude response at the output of BiquadNet, sorted by ascending order of frequency at which the maximum occurs for each filters. BiquadNet learns to build a selective filterbank which pools several frequency bands together, in order to pass them to FrameNet, which then encodes the time fluctuations in those pooled frequency bands at the frame level. Some of the channels shown on Figure 4 exhibit frequency patterns that could be linked to vowels or nasals, whereas the last channels exhibit a frequency patterns that could serve the purpose of encoding fricatives or plosives only, with wideband, high frequency content. It is also interesting to note that the pooled frequency channels representation build by BiquadNet for speech recognition further increases the density of activations for frequencies between 200 Hz and 1000 Hz, and may explain why TimeScaleNet allows a better accuracy than with a frozen version of BiquadNet with the Patterson's cochlear model using the parameters of Glasberg and Moore, shown as a dotted line on Figure 4. Interestingly, for the 100 first channels, which may mainly encode vowels and nasals, the learnt channels follow a very similar evolution than the Mel scale of 128 filters between 150 Hz and 5000 Hz. This is a really interesting property, since there was initially no intent to use the mel scale in the present study. However, for the highest channel numbers depicted on Figure 4, where the frequencies at which the maximum magnitude occurs at a larger frequency than 2500 Hz, the learnt filterbanks switches back to a Glasberg model, and clusters high frequencies together, which could help in recognizing consonants.

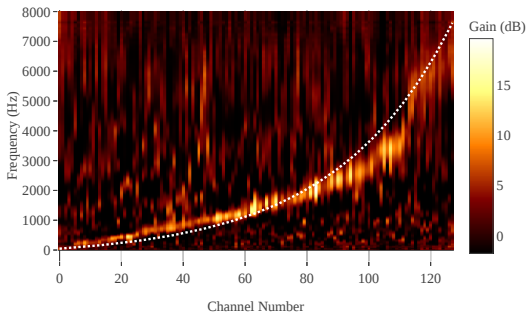


**Fig. 4**. Magnitude response of the equivalent filterbank at the output of BiquadNet, after convergence for the Speech Commands Dataset [20].

## 4. CONCLUSION

In this paper, we presented a machine learning approach of multiresolution modelling of unprocessed, time domain audio waveforms. The proposed deep neural network (TimeScaleNet) aims at merging digital signal processing techniques with new machine learning techniques. We show that this whole process allows to achieve speech recognition with a very high accuracy, which matches the performances of the best models to date on the Speech Commands dataset. By analyzing the learnt parameters in BiquadNet for this particular task and by deriving the equivalent filterbank magnitudes from the frozen model after convergence, we give further interpretability of the proposed machine hearing process. We also show that on this particular task, the proposed neural network builds a representation that both encodes the frequency content between 200 Hz and 3000 Hz with a pattern matching the mel-scale, and encodes higher frequency content with a pattern matching the Patterson's model. The approach also allows to pool frequency bands together, which can efficiently encode nasals, vowels, fricatives, and plosives for speech recognition. These results allow to interpret the machine learning task in light of cognitive models of audition, while standing on both machine learning and digital signal processing solid basis.

## 5. REFERENCES


[1] Sander Dieleman and Benjamin Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.

[2] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 421–425.

[3] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[4] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam, "Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, pp. 150, 2018.

[6] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrener, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[8] Lukasz Kaiser, Aidan N. Gomez, and Francois Chollet, "Depthwise separable convolutions for neural machine translation," in *International Conference on Learning Representations*, 2018.

[9] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[10] Lawrence R Rabiner and Bernard Gold, *Theory and application of digital signal processing*, Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975.

[11] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4624–4628.

[12] Malcolm Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[13] Richard F Lyon, "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3893–3904, 2011.

[14] Julius O. Smith, *Introduction to Digital Filters with Audio Applications*, W3K Publishing, 2007.

[15] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[16] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 971–980.

[17] Diederik P Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[19] Roy D Patterson, John Holdsworth, and Michael Allerhand, "Auditory models as preprocessors for speech recognition," in *The Auditory Processing of Speech: from Auditory Periphery to Words*, pp. 67–89. Mouton de Gruyler, Berlin, 1992.

[20] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[21] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra, "Hello edge: Keyword spotting on microcontrollers," *arXiv preprint arXiv:1711.07128*, 2017.

[22] Raphael Tang and Jimmy Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.