Evaluating the impact of multiscale temporal processing on sound-to-event recurrent neural networks

Michele Esposito (m.esposito@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Maastricht University
Maastricht, 6200MD The Netherlands

Bruno L. Giordano (bruno.giordano@univ-amu.fr)
Institut des Neurosciences de La Timone
UMR 7289, CNRS and Université Aix-Marseille, Marseille, France

Giancarlo Valente (g.valente@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Maastricht University
Maastricht, 6200MD The Netherlands

Elia Formisano (e.formisano@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Maastricht University
Maastricht, 6200MD The Netherlands

Abstract

This study investigates the impact of multiscale temporal processing on environmental sound recognition using deep neural networks (DNN). Inspired by the brain's capability to process auditory information across various time scales, we developed a multi-scale DNN architecture, integrating multiple parallel recurrent neural network (RNN) streams. Each stream processes the input spectrogram at a distinct temporal scale. The outputs of these streams are then combined and further processed to achieve sound categorization with a temporal resolution of 50 ms. This design aims to capture the diverse dynamics of natural sounds at large, ranging from transient, impulsive signals to repetitive and sustained sounds. We conducted a comparative analysis between the performance of this multiscale RNN network and networks trained on single-scale inputs. A comparison of our multiscale RNN network with single-scale networks reveals superior multiscale-RNN recognition of events. This performance advantage suggests that the combination of the unique information in multiple temporal scales achieves superior classification of natural sound events.

Keywords: Multiscale temporal processing; Natural sound recognition; Deep neural networks; Time-resolved event classification.

Introduction

A substantial body of behavioural, electrophysiological and neuroimaging research and biologically-inspired computational modelling has demonstrated the brain's ability to process different time scales of auditory information, ranging from milliseconds to seconds. The significance of multiscale temporal processing has been highlighted in diverse auditory tasks such as speech perception, music processing, and auditory scene analysis (Chi, Ru, & Shamma, 2005; Elhilali & Shamma, 2008; Santoro et al., 2014; Norman-Haignere et al., 2022).

Recent advancements in automated sound event detection (Mesaros, Heittola, Virtanen, & Plumbley, 2021) have highlighted the superior ability of deep learning models, including convolutional (CNN) (Hershey, Chaudhuri, Ellis, Gemmeke, et al., 2017; Esposito et al., 2023), recurrent (RNN), and convolutional-recurrent (CRNN) networks (Cakir, Parascandolo, Heittola, Huttunen, & Virtanen, 2017), at classifying complex, natural sounds and sound mixtures. Notably, however, the state of the art in natural sound DNN modelling lacks the incorporation of multiscale temporally-resolved mechanisms. Here, we describe our initial steps to fill this gap. We train a deep neural network (DNN) that enhances sound classification by exploiting the multiscale nature inherent in the temporal dynamics of natural sounds.

We considered a DNN architecture employing multiple RNN streams, each tailored to process inputs at different temporal scales. The outputs of these streams are then combined and further processed, leading to sound categorization. This design is intended to capture the distinct temporal dynamics of sound at various scales, from the quick chirp of a bird to the prolonged hum of urban noise. For our evaluation, we trained a *MultiScaleRNN* network with three streams, each receiving a different representation of the same sound computed at distinct time scales. We compared the performance of this network to that of networks trained on single-scale inputs. Importantly, the DNNs received the input spectrogram incrementally and generated an output with a temporal resolution of 50 ms. Our findings show that the multiscale approach exhibits superior ability to recognize events compared to the single-scale model. This performance advantage suggests that the multiscale network effectively combines the unique contributions of each temporal scale to classify sound events more accurately.

Methods

Model Architecture

We employed an RNN architecture consisting of three parallel streams of Gated Recurrent Units (GRUs). The streams received as input a spectrogram patch of 5, 20 and 40 timesteps respectively (Fig. 1). Each stream consisted of a series of three GRUs, with 1024, 512 and 256 units, respectively. The model comprises GRU layers with dropout (rate=0.2) for overfitting prevention, outputting data at each time-step. Outputs from the three streams concatenate into a single vector, processed by a feedforward (FF) network. The FF network consists of a 256-unit layer with ReLU activation, followed by a classification layer with 91 units using sigmoid activation. This setup enables frame-wise classification of sound sequences, with binary cross-entropy as the loss function.

Dataset

The dataset for evaluating the DNNs was derived from FSD50K (Fonseca, Favory, Pons, Font, & Serra, 2022), an open dataset featuring 51,197 audio clips, each human labelled across 200 categories. Based on data-quality consideration, we selected a subset of 90 categories, with over 19,689 sounds designated for the training set, 2,814 for the validation set and 7,055 for the evaluation set.

Pre-Processing

Each audio clip was resampled to 16 kHz mono and adjusted to a total duration of 6 seconds, with the initial second set to silence. Sounds shorter than 5 seconds were repeated to meet the required duration. Spectrograms were generated using the Short-Time Fourier Transform (STFT) with window sizes of 50 ms, 200 ms, and 400 ms, respectively for the three streams. These spectrograms were then converted to mel-spectrograms with 128 frequency bins (between 125 and 7500 Hz). The mel-spectrograms were segmented into 50 ms frames, with the first 20 frames labelled as “Silence” and the rest labelled with the sound-category label for the FSD50K clip.

Evaluation

To evaluate the performance of our models, we utilized the *F1-Score* macro metric. The F1 Score is calculated as the harmonic mean of precision and recall. The macro-average method computes the F1 Score independently for each class and then takes the average, thus treating all classes equally.

Results

Fig 1 provides an example of sound classification by our MultiScaleRNN. The panel below the waveform shows three spectrograms, each computed at a different scale, alongside their respective input patch spectrograms of varying lengths. The bottom panel shows the frame-by-frame scoring of the true label for the MultiScaleRNN and each of the single-scale networks. The MultiScaleRNN identifies the sound event more accurately, and outputs a declining true-label score as the barking sound ends, leading to the subsequent portion of silence. In contrast, the single-scale networks detect the event with a lower confidence below threshold and does not show the same strong decline in event recognition as the barking sound ends. These results demonstrate that the MultiScaleRNN provides a more accurate time-wise score than the other models.

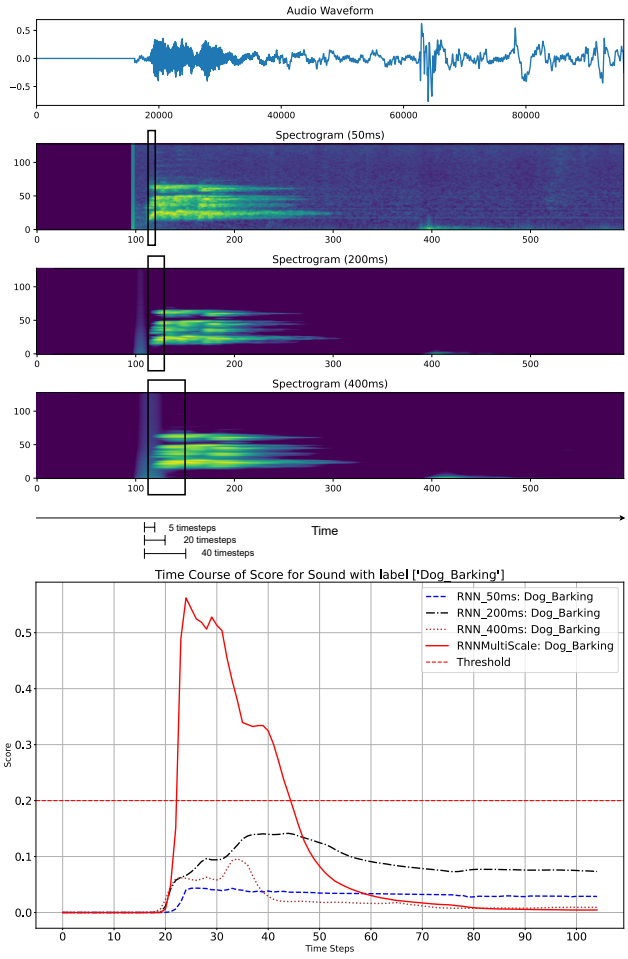

Figure 1: **Time-Wise Score Prediction across frames** Audio waveform of a dog barking twice, followed by a period containing silence and low-amplitude background noise (top panel). Spectrograms of the barking sound at three scales (middle panels). Predictions scores, with MultiScaleRNN outperforming other models at recognizing this sound event (lower panel).

Fig. 2 shows the frame-by-frame F1-Macro score, computed across all test sounds, and excluding the first 20 silence frames. The MultiScaleRNN consistently outperforms single-scale networks.

Conclusions

Our results suggest that integrating information across from three distinct acoustic scales improves sound classification in

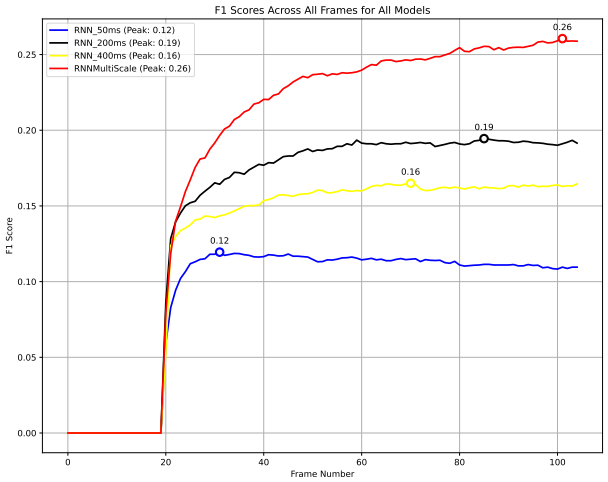

Figure 2: **F1-Score Trends Across Frames.** F1-Macro scores, excluding the first 20 silence frames, with MultiScaleRNN showing consistently higher performance levels compared to single-scale models.

RNN models. The proposed architecture generates frame-by-frame predictions and can be used in model-based analysis of time-resolved brain measurements, such as (intracranial)EEG or MEG. Ongoing work includes improvements in up-scaling the network to larger datasets, the inclusion of time-resolved convolutional layers for optimizing spectro-temporal feature extraction (Cakir et al., 2017) and the replacement of categorical labels with continuous semantic representations (Esposito et al., 2023).

Acknowledgments

This work was supported by the Dutch Research Council (NWO 406.20.GO.030 to EF), the French National Research Agency (ANR-21-CE37-0027-01 to BLG; ANR-16-CONV-0002 – ILCB; ANR11-LABX-0036 – BLRI), Data Science Research Infrastructure (DSRI; Maastricht University).

References

Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017, June). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(6), 1291–1303. Retrieved from http://dx.doi.org/10.1109/TASLP.2017.2690575 doi: 10.1109/taslp.2017.2690575

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, *124*(6), 3751–3771.

Esposito, M., Valente, G., Plasencia-Calaña, Y., Dumontier, M., Giordano, B. L., & Formisano, E. (2023). Semantically-informed deep neural networks for sound recognition. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2022). *Fsd50k: An open dataset of human-labeled sound events.*

Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., et al. (2017). CNN architectures for large-scale audio classification. In *Proc. icassp 2017* (pp. 131–135).

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021, September). Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, *38*(5), 67–83. Retrieved from http://dx.doi.org/10.1109/MSP.2021.3090678 doi: 10.1109/msp.2021.3090678

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., ... Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*, 455–469.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol*, *10*, e1003412.