Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms

Kleanthis Avramidis*, Agelos Kratimenos*,
Christos Garoufis, Athanasia Zlatintsi and Petros Maragos

National Technical University of Athens, School of ECE
Computer Vision, Speech Communication and Signal Processing Group

kle.avramidis@gmail.com; ageloskrat@yahoo.gr; cgaroufis@mail.ntua.gr; [nzlat, maragos]@cs.ntua.gr

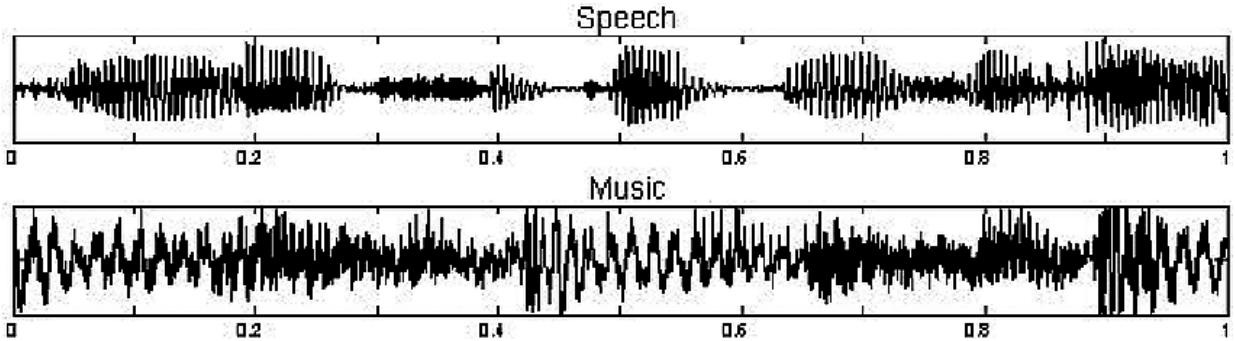
List of Contents

List of Contents

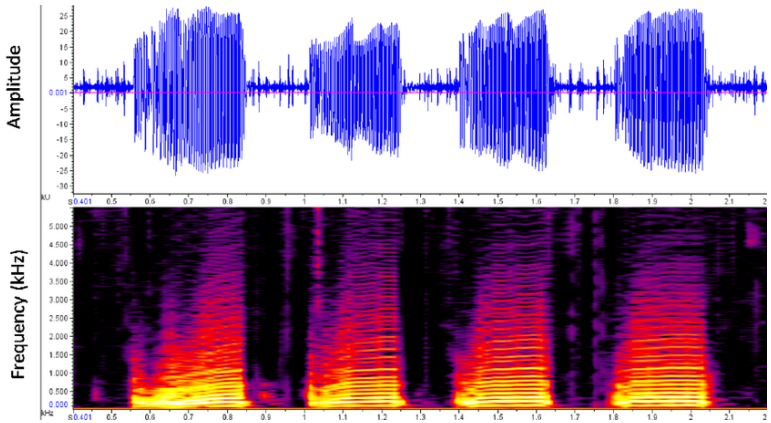
Waveforms & End-to-End Models

Waveform: Abstract representation of a sound wave

- Complex, non-intuitive structure
- Inherits noise from surroundings / equipment / sound event



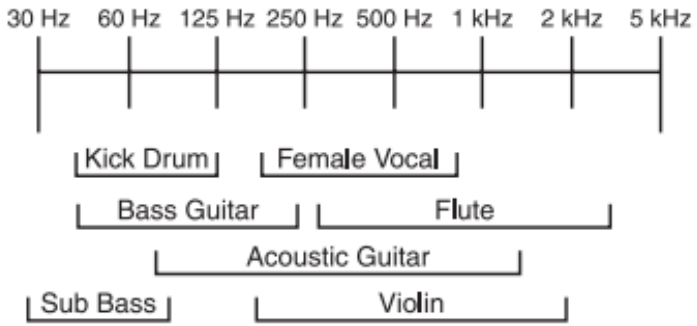
Instead: **Time-Frequency Representations** (i.e CQT, STFT)



But: Which should we use? What is their computational cost?

In Music Information Retrieval (MIR)

In MIR and Instrument Classification particularly, there is strong intuition into utilizing frequency-related representations, since notes and instruments are densely associated with specific frequency events.



Remark: Challenging and computationally expensive to design specialized feature representations for each different recognition task.
Proposal: Take advantage of Deep Learning methods to build efficient feature extractors from raw waveforms. Should handle:

- High input dimensionality and noisy structure
- Low-level temporal correlations and features
- Reduced computational cost without performance loss

List of Contents

Recurrent Networks (RNN)

Have been widely used in waveform and generally sequence modeling thanks to their ability to handle long-range temporal dependencies.

Bidirectional GRU:

- Lower computational cost compared to LSTM
- Comparable performance to LSTMs for audio sequences
- Considers both past and future features for dependencies

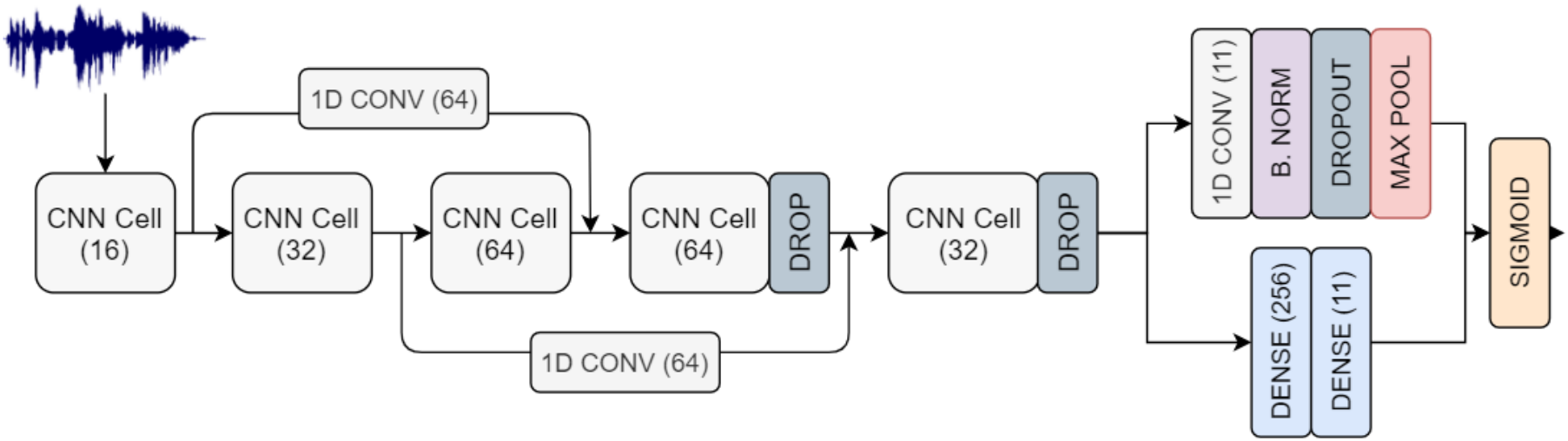
We experiment on the number of layers and utilized GRU units:

Number of Layers	**Number of Units**
1	128 or 256
2	128, 64
Dropout (0.5)	
Output Dense	

Convolutional Networks (CNN)

- Traditionally operate on images or time-frequency features.
- Already exhibited results in audio waveform processing [1].

Network based on [2] with alterations:



- **DCNN**: 2 dense layers to predict - *many trainable parameters*
- **FCN**: Dense layers *!* unit-kernel convolutions and filter pooling
- **RFCN**: embed skip connections to the previous model

[1] W.Dai et al, in Proc. ICASSP 2017 [2] A.Kratimenos et al, in Proc. EUSIPCO 2020

Combined Networks

- CNNs concentrate on temporally **local correlations** in waveforms, while RNNs are useful in modeling **longer-term** temporal structure.
- We expect that by efficiently combining these networks we will combine **di erent kinds** of discriminative features.
- We attach the best-performing RNN model of our experiments to the RFCN model in various positions.
- **Connection**: The embedded model takes the output of the corresponding CNN cell and its output is reduced to classes through convolution and Global Average Pooling. The final representation is the **average** of the 2 modules' outputs.
- Empirically search the optimal way of integrating the recurrent model information into a robust classifier.

List of Contents

Dataset & Pre-Processing

The **IRMAS Dataset** [3]: 11 instruments/classes

[cello, clarinet, flute, acoustic/electric guitar, organ, piano, saxophone, trumpet, violin, voice]

- **Training Set**: A set of 3-sec monophonic audio chunks (music tracks with a predominant instrument) for each class
- **Testing Set**: A set of multilabeled polyphonic tracks

Each training track was:

- cut to 1-sec segments
- downsampled and downmixed
- normalised by RMS energy

[3] J.J.Bosch et al, in Proc. ISMIR 2012.

Training Protocol & Evaluation

- 5-fold Cross-Validation
- Binary Cross-Entropy Loss (Multi-label Task)
- Adam Optimizer (10^{-3} learning rate)
- Learning Rate Reduction & Early Stopping

Training Protocol & Evaluation

- 5-fold Cross-Validation
- Binary Cross-Entropy Loss (Multi-label Task)
- Adam Optimizer (10^{-3} learning rate)
- Learning Rate Reduction & Early Stopping

Utilized evaluation metrics:

- **Label Ranking Average Prediction (LRAP)**: Suitable for multi-label tasks, ranking intuition, threshold independent
- F_1 **Score**: Comparable evaluation, class imbalance

IRMAS Testing Set: Tracks ranging from 5-20 sec. We average the per-sec predictions to obtain a single prediction for each track. Labeled instruments are active throughout the track.

List of Contents

Architecture Comparison

A simple recurrent network cannot su ciently decode the information included in a waveform

BiGRU	F1-micro %		F1-macro %		LRAP %		#Params
1 (128)	43.76	1.95	37.37	1.90	57.26	3.28	103.4K
1 (256)	43.51	2.46	39.19	2.23	58.47	2.73	403.4K
2	49.28	2.45	43.18	3.11	67.07	1.81	225.6K

Architecture Comparison

A simple recurrent network cannot su ciently decode the information included in a waveform

BiGRU	F1-micro %		F1-macro %		LRAP %		#Params
1 (128)	43.76	1.95	37.37	1.90	57.26	3.28	103.4K
1 (256)	43.51	2.46	39.19	2.23	58.47	2.73	403.4K
2	49.28	2.45	43.18	3.11	67.07	1.81	225.6K

1D CNNs are capable of extracting the most discriminative features from raw waveforms, almost as well as 2D models on spectrograms

FCN: in the absence of a dense layer, the network generalizes bette upon the information from spatial processing + less parameters

Models	F1-micro %		F1-macro %		LRAP %		#Params
DCNN	55.32	0.55	48.30	0.31	73.48	0.38	1.14M
FCN	58.45	0.36	49.96	0.29	75.13	0.32	81.8K
RFCN	58.55	0.22	50.22	0.35	75.14	0.23	85K

Architecture Comparison - Combination

Simply averaging the RNN and CNN model outputs lowers accuracy
! inadequate standalone performance of the BiGRU

We thus inserted the BiGRU in various locations in the RFCN mode

Models	F1-micro %		F1-macro %		LRAP %		#Params
$CRNN_2$	59.80	0.66	53.20	0.52	74.16	0.66	1.03M
$CRNN_3$	60.77	0.26	54.31	0.35	74.74	0.39	1.07M
$CRNN_4$	60.07	0.67	53.73	0.59	74.11	0.50	1.08M
$CRNN_5$	59.21	0.56	52.18	0.46	74.32	0.65	1.03M

Table: The subscript denotes the CNN layer in which the RNN was connected

No observed improvement in performance for the LRAP metric, steady increase however for F1 scores

The combined models consist of signi cantly more parameters

Literature Comparison

Models	F1-micro	F1-macro	LRAP	#Params
Bosch et al. [3]	0.503	0.432	{	{
Pons et al. [5]	0.589	0.516	{	{
Han et al. [4]	0.602	0.503	{	{
Kratimenos et al. [2]	0.616	0.506	0.767	24.3M
Reduced [2]	0.520	0.458	0.689	1.20M
Proposed	0.608	0.543	0.747	1.07M

Table: Comparison of our work with previous studies on the IRMAS Dataset

F1 micro surpasses most studies on the task, while we observe dominant performance at the more competitive F1 macro score.

Results obtained with a signi cantly reduced number of trainable parameters, low training - testing time and minimal pre-processing.

[3] J.J. Bosch et al, in Proc. ISMIR 2012. [4] Y.Han et al, in IEEE/ACM Trans. Audio, Speech and Language Processing, 2017.

[5] J.Pons et al, in Proc. EUSIPCO 2017. [2] A. Kratimenos et al, in Proc. EUSIPCO 2020.

Instrument-wise Analysis

We use the per-class F_1 score for this experiment

We examine how each instrument can be discriminative in either waveform or time-frequency representation.

Brass instruments (ex. clarinet, flute, saxophone) → Waveforms

Predominant and leading instruments (ex. guitars, piano, voice) → Constant Q Transform Spectrograms

List of Contents

Contributions & Future Work

Experiments with various architectures that are favourable towards waveform modeling, like Fully Convolutional and Residual Nets and information fusion.

A residual FCN-BiGRU model (1M parameters) outperforms the state-of-the-art with CQT spectrograms (24M parameters)

Brass instruments are being identi ed easier through waveforms, while leading instruments bene t more from time-frequency features.

Future work : alternate methods to exploit RNNs / enhance performance of predominant instruments / ways to deal with inherent noise

Thank you