Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music

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

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

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

EUSIPCO2020
28th European Signal Processing Conference
January 18-22, 2021

List of Contents

1. Introduction
2. Augmentation Methods
3. Experimental Setup
4. Results & Contributions
5. Future Work

List of Contents

1 Introduction

2 Augmentation Methods

3 Experimental Setup

4 Results & Contributions

5 Future Work

Intuition

How do we recognize a specific instrument playing?

Intuition

How do we recognize a specific instrument playing?

- **Timbre**: Sound Quality that distinguishes 2 sounds of the same pitch, loudness and duration (associated with the identification of environmental sound sources)

Intuition

How do we recognize a specific instrument playing?

- **Timbre**: Sound Quality that distinguishes 2 sounds of the same pitch, loudness and duration (associated with the identification of environmental sound sources)

But is timbre identification enough?

- Instrument Family Recognition (ex. string, wind)

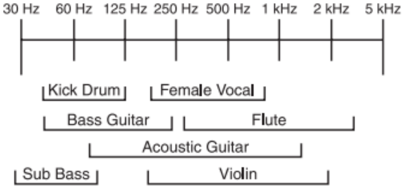
Intuition

How do we recognize a specific instrument playing?

- **Timbre**: Sound Quality that distinguishes 2 sounds of the same pitch, loudness and duration (associated with the identification of environmental sound sources)

But is timbre identification enough?

- Instrument Family Recognition (ex. string, wind)
- Pitch: Instruments play notes at different frequencies
- What if we combine these characteristics?



Instrument Classification

Instrument Classification

Subfield of **Music Information Retrieval**:

- Source Separation
- Music Generation
- Automatic Transcription
- Sentiment Analysis

Related Tasks: Audio Tagging, Categorization, Event Detection

Instrument Classification

Subfield of **Music Information Retrieval**:

- Source Separation
- Music Generation
- Automatic Transcription
- Sentiment Analysis

Related Tasks: Audio Tagging, Categorization, Event Detection

- **Monophonic**: Data typically include isolated notes or natural recordings of solo-playing instruments. Concentration given on handcrafted features ex. MFCCs
- **Polyphonic**: Various instruments co-play. Increased difficulty due to superposition of time-frequency features. Mainly deep modeling.

List of Contents

1 Introduction

2 Augmentation Methods

3 Experimental Setup

4 Results & Contributions

5 Future Work

Augmentation Principles - Mixing

Augmentation: Enhance data quality (and usually quantity)

Augmentation Principles - Mixing

Augmentation: Enhance data quality (and usually quantity)

Why mixing different tracks?

- Simple implementation, just overlaying data
- Substantially increased data: $\binom{n}{2}$ combinations for n instruments
- Good polyphonic music approximation
- Easily customizable mixtures

Augmentation Principles - Mixing

Augmentation: Enhance data quality (and usually quantity)

Why mixing different tracks?

- Simple implementation, just overlaying data
- Substantially increased data: $\binom{n}{2}$ combinations for n instruments
- Good polyphonic music approximation
- Easily customizable mixtures

We customize mixing in order to extract timbre-like features:

- Alignment based on Tempo
- Alignment based on Pitch
- Alignment based on Genre

Random & Genre-Aligned Mixing

- 11 instruments → 55 combinations (IRMAS dataset)
- Shuffle the audio segments and overlay one by one
- Number of segments equal to that of the smallest class

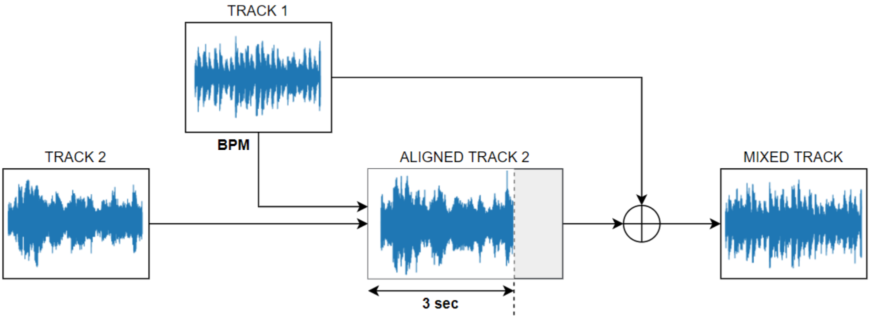
Random & Genre-Aligned Mixing

- 11 instruments → 55 combinations (IRMAS dataset)
- Shuffle the audio segments and overlay one by one
- Number of segments equal to that of the smallest class

- Utilized data already genre-labeled
- Alignment: only matching segments of the same genre
- **Genres available**: country-folk, classical, pop-rock, latin-soul

Tempo-aligned Mixing

- Instruments often hold the same beat when they co-play
- Tempo calculation (BPM) for each track at larger resolution
- The beat of the first track is the desired. The second track is stretched to match its BPM and then to 3 sec. After that, cut.



Pitch-aligned Mixing

- Timbre-like characteristics become distinguishable
- CREPE pitch prediction model [1] for pitch estimation
- Main frequencies f_1, f_2 of the two tracks are used to calculate the frequency shift needed for the second track:

$$s = 12 \log_2(f_1/f_2)$$

- Pitch is estimated at windows of 10 msec. In order to have realistic tracks we smooth the predictions and erase the short-time anomalies.
- We then apply pitch shift for each window of the second track. Resulting segments are being overlaid as before.

[1] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation", in Proc. ICASSP 2018, Calgary, AL, Canada, 2018.

List of Contents

1. Introduction
2. Augmentation Methods
3. Experimental Setup
4. Results & Contributions
5. Future Work

Dataset & Pre-Processing

The **IRMAS Dataset** [2]: 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

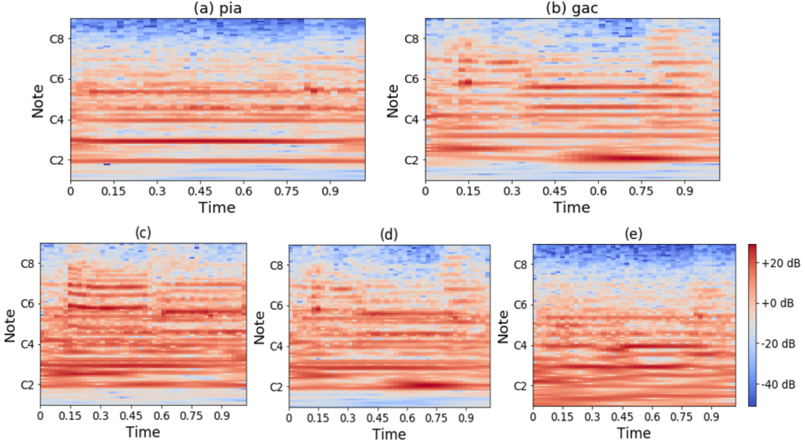
Each training track was:

- cut to 1-sec segments
- normalized and augmented
- turned into a spectrogram

[2] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals, in Proc. ISMIR, Porto, Portugal, 2012

Constant Q Transform

Feature representation of the audio segment, used at MIR tasks.
Each segment is transformed into a 96×87 time-frequency matrix.

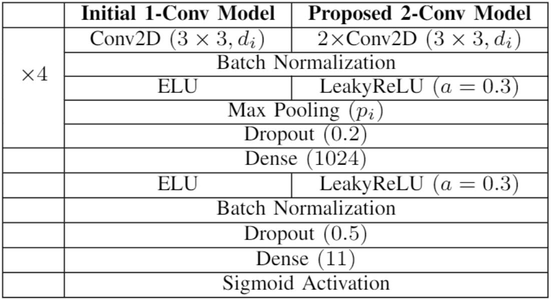


CQT spectrograms of 2 segments derived from piano and guitar.
(a) Piano, (b) Acoustic Guitar, (c) Random Mix, (d) Tempo Mix, (e) Pitch Mix

Architectures

	Initial 1-Conv Model	**Proposed 2-Conv Model**
×4	Conv2D $(3 \times 3, d_i)$	2×Conv2D $(3 \times 3, d_i)$
	Batch Normalization	
	ELU	LeakyReLU $(a = 0.3)$
	Max Pooling (p_i)	
	Dropout (0.2)	
	Dense (1024)	
	ELU	LeakyReLU $(a = 0.3)$
	Batch Normalization	
	Dropout (0.5)	
	Dense (11)	
	Sigmoid Activation	

Architectures



<table>
<tr><th></th><th>Initial 1-Conv Model</th><th>Proposed 2-Conv Model</th></tr>
<tr><td rowspan="5">×4</td><td>Conv2D $(3 \times 3, d_i)$</td><td>2×Conv2D $(3 \times 3, d_i)$</td></tr>
<tr><td colspan="2">Batch Normalization</td></tr>
<tr><td>ELU</td><td>LeakyReLU $(a = 0.3)$</td></tr>
<tr><td colspan="2">Max Pooling (p_i)</td></tr>
<tr><td colspan="2">Dropout (0.2)</td></tr>
<tr><td></td><td colspan="2">Dense (1024)</td></tr>
<tr><td></td><td>ELU</td><td>LeakyReLU $(a = 0.3)$</td></tr>
<tr><td></td><td colspan="2">Batch Normalization</td></tr>
<tr><td></td><td colspan="2">Dropout (0.5)</td></tr>
<tr><td></td><td colspan="2">Dense (11)</td></tr>
<tr><td></td><td colspan="2">Sigmoid Activation</td></tr>
</table>

Baseline proposed by Gururani [3]:

- Single 2D convolutional layer
- ELU activation function

Our proposed network model:

- Double 2D convolutional layers
- Leaky ReLU activation function

[3] S. Gururani, C. Summers, and A. Lerch, “Instrument Activity Detection in Polyphonic Music Using Deep Neural Networks”, in Proc. ISMIR 2018, Paris, France, 2018.

Experimental Protocol: Training Parameters

The set of spectrograms is partitioned into 5 equal-size subsets, We perform multiple training sessions, using each subset in rotation as validation data and the remaining 4 for fitting the classifier.

- every sample is utilized at both training and validation
- more representative estimations and statistical inference

Configuration:

- Binary Cross-Entropy Loss (Multi-label Task)
- Adam Optimizer (10^{-4} learning rate)
- Learning Rate Reduction & Early Stopping

Experimental Protocol: Evaluation

Utilized evaluation metrics:

- **Label Ranking Average Prediction (LRAP)**: Suitable for multi-label tasks, ranking intuition, threshold independent
- **Area Under the ROC Curve (AUC)**: The probability that a classifier will rank a randomly chosen positive instance higher than a negative instance
- $\boldsymbol{F_1}$ **Score**: Comparable evaluation, class imbalance

Experimental Protocol: Evaluation

Utilized evaluation metrics:

- **Label Ranking Average Prediction (LRAP)**: Suitable for multi-label tasks, ranking intuition, threshold independent
- **Area Under the ROC Curve (AUC)**: The probability that a classifier will rank a randomly chosen positive instance higher than a negative instance
- 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

1. Introduction
2. Augmentation Methods
3. Experimental Setup
4. Results & Contributions
5. Future Work

Architecture Comparison

- Significantly improved LRAP score

Mixing Dataset	Model	LRAP	AUC
Monophonic	Gururani et al. [3]	0.738	0.839
	Proposed	**0.767**	**0.843**
Random	Gururani et al. [3]	0.750	0.853
	Proposed	**0.776**	**0.861**

- Surpassing published efforts

Models	F_1micro	F_1macro
Bosch et al. [4]	0.503	0.432
Han et al. [5]	0.602	0.503
Pons et al. [6]	0.589	**0.516**
Proposed - Baseline	**0.616**	0.506

Key features:

- Consecutive Convolutional & Batch Normalization Layers
- Minor differences in audio pre-processing

Augmentation Impact

- Random Mixing Augmentation slightly improves the baseline
- BUT: Does this improvement reflect the mixtures or just the increased number of data? → Additional samples through pitch shifting

Augmentation Impact

- Random Mixing Augmentation slightly improves the baseline
- BUT: Does this improvement reflect the mixtures or just the increased number of data? → Additional samples through pitch shifting

- **Genre Alignment:** Most closely related to random.
  Only minor differences.

Metrics\Mix Methods	Monophonic	Random	Genre
LRAP	0.767 ± 0.008	**0.776** ± 0.006	0.774 ± 0.005
Mean AUC	0.843 ± 0.002	0.861 ± 0.001	**0.862** ± 0.003
F_1 micro	0.616 ± 0.009	**0.624** ± 0.006	0.617 ± 0.004
F_1 macro	0.506 ± 0.006	**0.528** ± 0.006	0.519 ± 0.004

Augmentation Impact

- Random Mixing Augmentation slightly improves the baseline
- BUT: Does this improvement reflect the mixtures or just the increased number of data? → Additional samples through pitch shifting

- **Tempo Alignment:** Cancelling out tempo variations helps instrument differences to be better understood.

Metrics\Mix Methods	Monophonic	Random	Tempo
LRAP	0.767 ± 0.008	0.776 ± 0.006	**0.790** ± 0.003
Mean AUC	0.843 ± 0.002	0.861 ± 0.001	**0.865** ± 0.002
F_1 micro	0.616 ± 0.009	0.624 ± 0.006	**0.628** ± 0.003
F_1 macro	0.506 ± 0.006	0.528 ± 0.006	**0.531** ± 0.003

Augmentation Impact

- Random Mixing Augmentation slightly improves the baseline
- BUT: Does this improvement reflect the mixtures or just the increased number of data? → Additional samples through pitch shifting

- **Pitch Alignment:** Intuitive improvement. Instrument characteristics are decoded **at best**, when instruments perform at the same pitch.

Metrics\Mix Methods	Monophonic	Random	Pitch
LRAP	0.767 ± 0.008	0.776 ± 0.006	**0.795** ± 0.002
Mean AUC	0.843 ± 0.002	**0.861** ± 0.001	0.856 ± 0.001
F_1 micro	0.616 ± 0.009	0.624 ± 0.006	**0.635** ± 0.002
F_1 macro	0.506 ± 0.006	0.528 ± 0.006	**0.532** ± 0.003

Ensemble Models

Back to intuition: We expect that some instruments are sensitive to certain of the examined characteristics (ex. bass - violin)

- We thus expect that aggregated predictions will combine most of the discriminative features of 11 different instruments.

Ensemble Models

Back to intuition: We expect that some instruments are sensitive to certain of the examined characteristics (ex. bass - violin)

- We thus expect that aggregated predictions will combine most of the discriminative features of 11 different instruments.

We average the predictions made by each of the examined strategies and then performed the testing. The best performing ensemble model consists of aggregated predictions of all models excluding the randomly mixtures.

- Ensemble improves LRAP by 4% compared to the monophonic
- Combining genre, pitch and tempo features assists recognition the most, while pitch-sync seems to be the key trait.

Metrics\Mix Methods	Monophonic	Genre	Tempo	Pitch	Combined
LRAP	0.767 ± 0.008	0.774 ± 0.005	0.790 ± 0.003	0.795 ± 0.002	**0.805** ± 0.002
Mean AUC	0.843 ± 0.002	0.862 ± 0.003	0.865 ± 0.002	0.856 ± 0.001	**0.874** ± 0.001
F_1 micro	0.616 ± 0.009	0.617 ± 0.004	0.628 ± 0.003	0.635 ± 0.002	**0.647** ± 0.003
F_1 macro	0.506 ± 0.006	0.519 ± 0.004	0.531 ± 0.003	0.532 ± 0.003	**0.546** ± 0.004

Literature Comparison

Table: Comparison on IRMAS Dataset Performance

Models	F_1micro	F_1macro
Bosch et al. [4]	0.503	0.432
Han et al. [5]	0.602	0.503
Pons et al. [6]	0.589	0.516
Proposed - Baseline	0.616	0.506
Proposed - Combined	**0.647**	**0.546**

[4] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals", in Proc. ISMIR, Porto, Portugal, 2012.

[5] Y. Han, J. Kim, and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music", in IEEE/ACM Trans. Audio, Speech, and Language Processing, 25(1):208– 221, 2017.

[6] J. Pons, O. Slizovskaia, R. Gong, E. G omez, and X. Serra, "Timbr eAnalysis of Music Audio Signals with Convolutional Neural Networks", in Proc. EUSIPCO 2017, Kos, Greece, 2017.

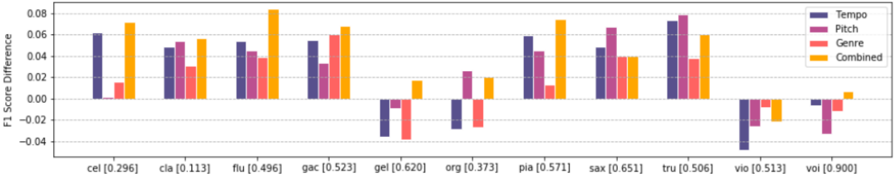
Per-Instrument Analysis

- We use the per-class F_1 score for this experiment
- We examine how each instrument responds to each method

Per-Instrument Analysis

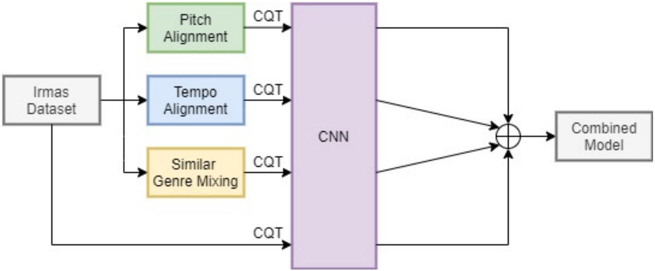
- We use the per-class F_1 score for this experiment
- We examine how each instrument responds to each method

- Augmentation mainly improves recognition in instruments that are of supporting role or tend to play in unison (ex. woodwind)
- Predominant and leading instruments (ex. electric guitar, violin) favor models trained on monophonic data



Contributions

- Address a polyphonic dataset through initial monophonic data
- Efficiency of augmentation based on mixing audios
- Timbre-like characteristics isolation
- Efficiency of ensemble classifiers in a multi-label task
- Profiling of the per-instrument sensitivities



List of Contents

1. Introduction
2. Augmentation Methods
3. Experimental Setup
4. Results & Contributions
5. Future Work

Current Research & Directions

Current Research & Directions

- How can we get rid of the need for time-frequency representations?
- End-to-end models without pre-processing
- Fully convolutional and recurrent models for feature extraction

K. Avramidis, A. Kratimenos, C. Garoufis, A. Zlatintsi, and P. Maragos, "Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms," submitted to ICASSP 2021.

Current Research & Directions

- How can we get rid of the need for time-frequency representations?
- End-to-end models without pre-processing
- Fully convolutional and recurrent models for feature extraction

K. Avramidis, A. Kratimenos, C. Garoufis, A. Zlatintsi, and P. Maragos, "Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms," submitted to ICASSP 2021.

Other proposed future directions:

- Alignment according to musical elements (ex. key)
- Multi-instrument mixtures / instrument clustering

Thank you

STAY SAFE!