Audio-visual deep learning methods for musical instrument classification and separation

Olga Slizovskaia

TESI DOCTORAL UPF / year 2020

THESIS SUPERVISORS

Dr. Emilia Gómez
Department of Information and Communication Technologies, Universitat Pompeu Fabra
Joint Research Centre, European Commission

Dr. Gloria Haro
Department of Information and Communication Technologies, Universitat Pompeu Fabra

Progress in unstoppable. Nothing will work anyway.
Ilya Segalovich (1964-2013)

Acknowledgements

My path to the PhD thesis was full of luck and coincidence. But more than that, it was full of happy encounters with truly inspiring people, whose influence can not be underestimated.

First of all, I want to express my gratitude to my supervisors, Emilia Gómez and Gloria Haro. They gave me a chance to start an unforgettable 4-years adventure, gave me their guidance, trust, and support. I am deeply thankful for their patience and understanding and I can't think of a better role model for myself. I would like to thank Xavier Serra and Coloma Ballester for giving this opportunity and making a friendly, open and inclusive environment at UPF. I was lucky to be a part of both research groups, the Music Technology Group and the Image Processing Group, and I have found it truly enriching.

This work was influenced by many people with whom I could collaborate with and to learn from, and whom I want to explicitly acknowledge: Marius Miron, Leo Kim, Juan Montesinos, Eduardo Fonseca, Daniel Michelsanti, Jordi Pons, Rong Gong, and Joan Serrà.

I would also like to thank my former and present colleagues for numerous conversations and ideas: Pritish Chandna, Helena Cuesta, Merlijn Blaauw, Juan Gómez, Andrés Pérez, Juanjo Bosch, Lorenzo Porcaro, Aggelos Gkiokas, Furkan Yesiler, Felipe Navarro, Jordi Bonada from the MIR Lab; Pablo Zinemanas, Dmitry Bogdanov, Vsevolod Eremenko, Albin Correya, Frederic Font, Georgi Dzhambazov, Minz Won, Sergio Oramas, Oriol Romaní, Zacharias Vamvakousis from the MTG and Venkatesh Kadandale, Lara Raad, Patricia Vitoria, Pablo Arias, and Adrià Arbués from the GPI.

My life would not be the same without friends who made these years in Barcelona fun and fulfilling: Maria Rauschenberger, Cecília Nunes, Vadim Fedorov, Tanya Radchenko, Federico Franzoni, Julia Olkhovskaia, Xavier Favory, Javier Vázquez – I am happy to have friends like you. I have plenty of warm memories with my former and present UPF colleagues: Amelia Jiménez-Sánchez, Zaira Pindado, Rasoul Nikbakht, Carlos Ruiz, Gabriela Ghimpeteanu, Raquel Gil, Arash Akbarinia and Sara Noureldin.

A special thanks go my former master students, form whom I learned so much: Siddharth Bhardwaj and Will Barleycorn, it was an inspirational experience.

A significant part of my thesis was conveyed under the María de Maeztu program, and I sincerely thank Aurelio Ruiz for all the efforts he put in coordinating the program. I also want to thank Lydia Garcia, Jana Safrankova, Ruth Temporal, Sonia Espi and Cristina Garrido for helping in all the administrative tasks.

During the months of confinement, I smoothly transitioned to Voctro Labs and found new and well-known teammates: Óscar Mayor, Jordi Janer, Álvaro Sarasúa, and Matan Gover. I am especially thankful to Jordi for giving me flexibility and understanding while adapting to the new work environment and finishing the thesis.

Looking backwards, I want to express my appreciation to the people who brought me to this point and without whom it would never happen. I am grateful to Jenia for teaching me brave and responsibility, and to Manuel for teaching me patience and planning. I thank Alexander and Alyona for sharing the excitement of this adventure and Zlata and Yaromir for the fondness that exceeds all limits. Above all, my deepest gratitude goes to my parent, Elena and Evgenii, for their unconditional love and support. For all that you gave me, for always challenging me, for always being there for me, I hope to be able to give something back.

Abstract

In music perception, the information we receive from a visual system and audio system is often complementary. Moreover, visual perception plays an important role in the overall experience of being exposed to a music performance. This fact brings attention to machine learning methods that could combine audio and visual information for automatic music analysis.

This thesis addresses two research problems: instrument classification and source separation in the context of music performance videos. A multimodal approach for each task is developed using deep learning techniques to train an encoded representation for each modality. For source separation, we also study two approaches conditioned on instrument labels and examine the influence that two extra sources of information have on separation performance compared with a conventional model. Another important aspect of this work is in the exploration of different fusion methods which allow for better multimodal integration of information sources from associated domains.

Resum

En la percepció visual, és habitual que rebem informacions complementàries des del nostres sistemes visual i auditiu. A més a més, la percepció visual té un paper molt important en la nostra experiència integral davant una interpretació musical. Aquesta relació entre àudio i visió ha fet créixer l'interàs en mètodes d'aprenentatge automàtic capaços de combinar ambdues modalitats per l'anàlisi musical automàtic.

Aquesta tesi se centra en dos problemes principals: la classificació d'instruments i la separació de fonts en el context dels vídeos musicals. Per a cadascú dels problemes, s'ha desenvolupat un mètode multimodal fent servir tècniques de Deep Learning. Això ens ha permès d'obtenir –gràcies a l'aprenentatge– una representació codificada per a cada modalitat. A més a més, en el cas del problema de separació de fonts, també proposem dos models condicionats a les etiquetes dels instruments, i examinem la influència que tenen dos fonts d'informació extra sobre el rendiment de la separació –tot comparant-les amb un model convencional–. Un altre aspecte d'aquest treball es basa en l'exploració de diferents models de fusió, els quals permeten una millor integració multimodal de fonts d'informació de dominis associats.

Resumen

En la percepción musical, normalmente recibimos por nuestro sistema visual y por nuestro sistema auditivo informaciones complementarias. Además, la percepción visual juega un papel importante en nuestra experiencia integral ante una interpretación musical. Esta relación entre audio y visión ha incrementado el interés en métodos de aprendizaje automático capaces de combinar ambas modalidades para el análisis musical automático.

Esta tesis se centra en dos problemas principales: la clasificación de instrumentos y la separación de fuentes en el contexto de videos musicales. Para cada uno de los problemas, se desarrolla un método multimodal utilizando técnicas de Deep Learning. Esto nos permite obtener –a través del aprendizaje– una representación codificada para cada modalidad. Además, para el problema de la separación de fuentes, también proponemos dos modelos condicionados a las etiquetas de los instrumentos, y examinamos la influencia que tienen dos fuentes de información extra en el rendimiento de la separación -comparándolas contra un modelo convencional-. Otro aspecto importante de este trabajo se basa en la exploración de diferentes modelos de fusión que permiten una mejor integración multimodal de fuentes de información de dominios asociados.

Contents

List of Figures

List of Tables

List of Abbreviations

ASR Automatic Speech Recognition. 32

AV Audio-Visual. 3, 5, 32

AVO Audio-Visual Object. 33

BCE Binary Cross-Entropy. 94, 95

CCA Canonical Correlation Analysis. 34

CNN Convolutional Neural Network. 11, 23, 39, 47, 58, 60, 61, 70

CQT Constant-Q Transform. 42, 51

DBN Dynamic Bayesian Network. 33

DTW Dynamic Time Warping. 41, 42

EPS Energy at Predicted Silence. 98

GMM Gaussian Mixture Model. 33, 34, 39, 48

HMM Hidden Markov Model. 33, 42

HPSS Harmonic-Percussive Source Separation. 72, 74

IBM Ideal Binary Mask. 101

LSTM Long Short-Term Memory. 49

MIR Music Information Retrieval. 1, 3, 5, 32, 57, 71

NMF Non-negative Matrix Factorization. 39, 86

PES Predicted Energy at Silence. 98, 101

SCSS Single Channel Source Separation. 10, 49, 81

SD Scale-Dependent. 98

SDR Signal-to-Distortion Ratio. 49, 87, 90, 98, 101, 102, 107, 109

SGD Stochastic Gradient Descent. 62

SI Scale-Independent. 49, 98, 101, 102, 107

SIFT Scale Invariant Feature Transform. 11, 22, 73

SIR Signal-to-Inference Ratio. 87, 90, 98, 107

SIR Signal-to-Artifacts Ratio. 87, 90, 98, 107

STFT Short-Time Fourier Transform. 10, 16, 58, 64, 94

Chapter 1

Introduction

1.1. Motivation

This research is mostly driven by the fact that neither machine perception nor human perception is perfect. Although there has been a long and successful history of the first attempting to mimic the second, many problems remain open, especially for complex phenomena that involve several perception channels.

Without a doubt, music is one such phenomenon. Several studies emphasize the multimodal nature of music perception, which can involve listening, learning, dancing, or performing [Leman, 2017, Thompson et al., 2005]. Curiously, music has *always* been a multimodal art. It first shifted to an audio-only mode with a technological advance in the late nineteenth century, but later returned to a somewhat altered audio-visual mode with a technological advance in the twentieth century. Despite this, most research in computational music analysis and processing focuses on the auditory component. Until recently, the research community disregarded its multimodal complexity.

Our goal as music information researchers is to untangle that complexity. Thus, one of the first calls for multimodal strategies for Music Information Retrieval (MIR) dates at least as far back as 2011 [Liem et al., 2011, Essid and Richard, 2012]. It touches upon the sophisticated relations be-

tween different music modalities, from creation to performance. It is important to consider every available source of information, especially if they are complementary and each modality brings additional meaning, or if the obtained meaning of each modality is subjective.

However it is not only music that is complex. The world itself is complex, and everything that humanity has done so far to understand it falls well under the `divide-and-conquer` paradigm. For digital multimedia, we normally take the `divide` part for granted, having audio, music sheets, MIDI, metadata, visual and context music components. Therefore, we can assume that this part of the puzzle has nearly been solved. At the same time, the `conquer` part has advanced significantly and is reaching its limits for classical analysis tasks with state-of-the-art machine learning and signal processing methods. However, algorithmically speaking, `divide-and-conquer` has another vital step: `merge`. As will be shown in the thesis, when it comes to analysing complex multimodal phenomena, `merge` is where our efforts should be.

Thanks to evolution, humans are very good at merging different sources of information, doing so constantly and mostly unconsciously. Algorithms are not so advanced at the moment and there are multiple reasons why. First, although various integration techniques have been discussed for decades already [Naphade et al., 2002, Essid and Richard, 2012], there is still no agreement and no single and straightforward method of multimodal fusion. Second, in some cases the information obtained from one modality may contradict information from another modality. This is especially the case for digital multimedia when some data in one of the modalities may be missing or incomplete. Some examples are a corrupted video stream with preserved audio stream or just a video focusing on a lead vocalist in a band. Finally, there are still gaps in understanding multimodal human perception, which sometimes produces curious mismatches. This is illustrated by the McGurk effect [McGurk and MacDonald, 1976], which consists of speech misunderstanding that results from a mismatch of visual and audio stimuli. It is also shown by ventriloquism, which involves a misattribution of speech source location due to misleading visual stimuli. At a more general level, audio stimuli can provoke visual illusions as reported for

light flashes and sound beeps in [Shams et al., 2002].

Regarding musical performance, we should emphasize the special importance of the visual modality in perception of this type of multimedia. Thus, a meta-analysis by [Platz and Kopiez, 2012] reveals that musical performances are rated higher in terms of overall impression, likability, expressiveness, and overall quality if the visual component is present. Moreover, a study of [Griffiths and Reay, 2018] shows the effect of congruence between audio and visual information in evaluations of musical performances (recordings with professional video and amateur audio are rated higher than recordings made of amateur video and professional audio in terms of musicality, technical proficiency, and overall performance quality). Once again, this highlights the importance of audio-visual analysis of musical performances.

Research in multi-modal, especially audio-visual (AV), MIR is not only interesting from the theoretical perspective but also from the practical point of view. The following three data analysis tasks can be considered under the umbrella of AV MIR:

- **Audio-visual classification and detection:** both are widely used for indexing and organizing video content in big data repositories on the internet to facilitate information retrieval. For example, there is evidence that a music genre can be recognized just by looking at the album cover [Schindler, 2019].

- **Audio-visual source separation:** classical blind source separation may benefit from including additional sources of information such as music sheets or corresponding video stream as they provide extra distinguishing clues, which is particularly important when separating similar sources [Parekh et al., 2017, Zhao et al., 2019]. Moreover, parallel and synchronized data streams within a dataset make possible the use of unsupervised learning techniques and mitigate the need for annotated data.

- **Audio-visual music transcription:** in some cases, such as multi-source recordings and for musical instruments whose sound is de-

fined by hand position, it is virtually impossible to install a separate microphone for each player and achieve accurate transcription with audio analysis; however, music transcription can be accomplished using image processing techniques [Zinemanas et al., 2017, Koepke et al., 2020].

From the practical perspective, there are several application domains which can benefit from audio-visual analysis:

- In the field of music education, AV MIR analysis can both increase involvement of students in the educational process and provide better feedback on assignments.
- For video post-processing, apart from classical remixing and synchronization tasks, the scene rearrangement problem could be addressed so that an editor could automatically substitute an object in a scene together with its audio component, or change its location and corresponding sound direction.
- For augmented reality, a more immersive experience can be achieved using joint models to automatically synthesize an audio stream from visual events and vice versa.

To conclude, it is worth mentioning that the scope of audio-visual analysis is much broader than just musical performance analysis. Even though this research focuses on the music information retrieval domain, specifically, on classification and source separation in instrumental musical performances, we believe the methods presented in this thesis can impact other domains as well, such as audio-visual speech and multimedia processing.

1.2. Goals and research questions

This research focuses on providing better understanding of underlying relationships between audio and visual modalities in musical performance

videos. Our goals are to identify areas of research interest for AV-MIR and to improve classic audio-only methods by making use of visual and contextual information in the settings of audio-visual musical instrumental performances widely available in a form of video recordings on the internet. The research aims to answer the following questions:

- **RQ1** Which MIR tasks can benefit from audio-visual analysis? What is a potential improvement that we can obtain?
- **RQ2** Can we extract meaningful and effective audio-visual features useful for MIR?
- **RQ3** How can audio-visual strategies and features help us to better understand underlying multimodal relationships? How much impact does visual information have on musical performance analysis?
- **RQ4** What are optimal strategies for multimodal fusion? How sensitive are different machine learning methods to incorrect data fusion techniques and the missing modalities problem?

The principal objective of this study is to ***design a machine learning method for musical performance video analysis which enables use of audio-visual information available on the internet***. In order to achieve this and answer the above-mentioned research questions, the following two tasks has been selected:

- **T1** Audio-visual musical instrument classification.
- **T2** Audio-visual source separation in musical performance videos.

The selected tasks represent two different viewpoints and problems in a broader scope of AV-MIR as described in Section 1.1. This selection allows the work to be done under the general paradigm of analysis, transformation and synthesis. Here, the classification task represents analysis and pre-transformation stages and the source separation task states for synthesis.

The further decomposition of the generic tasks into specific objectives looks as follows:

- Perform a state-of-the-art review of available techniques in audio, video, and audio-visual processing. Define a methodology for experimental research.
- Identify available data collections and requirements. Carry out data acquisition and analysis.
- Define a technical approach for audio-visual instrument classification. Propose a tool for post-hoc interpretability analysis of classification decisions.
- Define a technical approach for audio-visual music source separation and propose a tool for multimodal source separation.
- Prepare demos and convey outreach activities in order to spread the outcomes of the research.

1.3. Challenges

Challenges in AV-MIR fall into two principal categories: 1) related to the data itself, and 2) related to its processing.

Data-related challenges

Shortage of dedicated datasets. Multimodal analysis is a relatively new research area, therefore we must first look at datasets in each of the subfields we want to study. In the field of music and audio analysis, the majority of datasets represent the audio-only domain, and a similar observation can be made for image analysis. Although there are several video datasets that include both components, they are made with a general tagging problem in mind and contain a very small fraction of music data. On the other hand, the large-scale datasets are often gathered automatically and lack careful curation. We could also find audio-visual speech analysis datasets, but there is no dedicated, large, and well-annotated multimodal dataset for audio-visual analysis of musical performances.

Low quality of available data. Most of the available datasets have been gathered from YouTube and other video streaming services [Aytar et al., 2016, Abu-El-Haija et al., 2016, Jiang et al., 2018, Zhao et al., 2018] which means that videos are often of low quality. Common problems also include but are not limited to:

- presence of video artifacts due to video post-processing;
- weak and incomplete annotations;
- disagreement between modalities such that the annotated information is present only in one of the modalities but not in the other;
- asynchrony between audio and video such that a sounding object can appear only during a few seconds of a recording.

Large diversity in data. Within available datasets, we can observe a number of variations in visual and audio characteristics within the same category of musical instruments (or audio-visual objects) due to different recording conditions, such as illumination, viewpoints, or properties of the recording devices [Parekh, 2019]. Moreover, similar instruments within a single family tend to have very few differences.

Processing-related challenges

Dimensionality mismatch problem. While music data is a highly multimodal concept, each of its components is multidimensional, with a high variance in the number of dimensions. In order to perform joint audio-visual analysis, we have to provide a method which could handle the dimensionality mismatch, meaning a data preprocessing algorithm that allows both audio and visual data to be treated at a comparable scale.

Streaming data processing. That music evolves over time, is one of its remarkable characteristics. This puts an additional requirement on making sure that information from all modalities is aggregated with correct timing. Besides this, differing sample rates for audio and video processing should be taken into account.

Data aggregation problem. Once converted to a common representation, the multimodal data has to be merged and jointly processed. Although several studies have been conducted on the optimal fusion strategies for multimodal data [Ngiam et al., 2011, Essid and Richard, 2012, Srivastava and Salakhutdinov, 2012, Sohn et al., 2014, Zhang et al., 2016], there is still no agreement among them and it is currently an active research area.

1.4. Outline of the thesis

In Chapter 2 we provide the necessary background information on the techniques used in this thesis. In Chapter 3 we set the stage of audio-visual music information research and place our work within the area by discussing a historical evolution of audio, visual and audio-visual methods, specifically focusing on techniques that have been developed for musical instrument classification and source separation.

This is followed by Chapter 4, where we detail our audio-visual musical instrument classification method based on a deep learning framework, providing a complete method description, experimental results and a case-study on the method's explainability.

In Chapter 5 we continue development on audio-visual coupling strategies within the deep learning framework and investigate context-conditioned and visual-conditioned source separation. We take advantage of the availability of visual and contextual information, which allows improvement of source separation quality in complex auditory scenarios.

In Chapter 6 we reflect on the achieved results, provide a summary of out contributions, and discuss future research perspectives.

Chapter 2

Background

2.1. Introduction

This work focuses on two major MIR problems and tackles multiple aspects of data processing and data integration in multimodal machine learning. In this chapter we give an overview of background technologies. This includes formal definitions of classification and source separation tasks, basic principles of convolutional neural networks and an introduction to encoder-decoder framework. We then describe common data representations for audio and video data, along with the idea of representation learning and how it has been strengthened by modern deep learning techniques. Lastly, we provide a summary of multimodal data fusion methods.

2.2. Deep learning methods

2.2.1. Basic learning setup

Classification

For the classification setup, we have a set of objects $X \in \mathbb{R}^N$ and a set of class labels $C \in \{0, 1\}^K$. We are interested in predicting a final subset of labels $y = (y_1, y_2, ..., y_K) \in C$ for a sample $x = (x_1, x_2, ..., x_N) \in X$

which is an N-dimensional representation of an object x. Therefore, we are looking for a function such as

$$\hat{y} = f_\theta(x),$$

where θ are the function parameters and $\hat{y}$ is a probability vector for the class estimates. A common mapping function f could be a complex non-linear neural network, where the final probability scores can be obtained via softmax function

$$\hat{y_j}(z) = \frac{e^{z_j}}{\sum_{i=1}^{L} e^{z_i}},$$

which takes a vector of arbitrary real-valued scores $z = (z_1, ...z_L) \in \mathbb{R}^L$ and transforms it to a vector of values between zero and one that sum to one.

A common choice to learn the parameters θ is through a backpropagation algorithm [Kelley, 1960, Rumelhart et al., 1986] where one of the common optimization techniques [Le et al., 2011] can be used to minimize a *loss function*. For multi-labels classification, the loss function can be defined as categorical cross-entropy between ground truth and estimated probability distributions of class labels:

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)).$$

Source Separation

Single channel source separation (SCSS) consists of estimating the individual sources $x_i : \mathbb{R} \to \mathbb{R}$, given a mono mixture time-domain signal $y : \mathbb{R} \to \mathbb{R}$ of N sources:

$$y(t) = \sum_{i=1}^{N} x_i(t). \tag{2.1}$$

Although we can predict the signals directly, a general approach for solving SCSS involves the estimation of N masks for Short-Time Fourier transform (STFT) values of the mixture, as defined in Eq. 2.4. In this

case, we consider a time-frequency representation of the mixture $\boldsymbol{Y}$ and the sources $\boldsymbol{X_i}$, with the goal of the source separation method being to learn a real-valued (or complex-valued) mask M_i for each source i .

In this work we consider only two types of *real-valued* masks, namely *ideal ratio or soft* masks M_i^{ir}:

$$M_i^{ir}(\tau, \omega) = \frac{|\boldsymbol{X_i}(\tau, \omega)|}{|\boldsymbol{Y}(\tau, \omega)|}, \tag{2.2}$$

and *ideal binary* masks M_i^{ib}:

$$M_i^{ib}(\tau, \omega) = \begin{cases} 1, & \text{if } \frac{|\boldsymbol{X_i}(\tau,\omega)|}{|\boldsymbol{Y}(\tau,\omega)|-|\boldsymbol{X_i}(\tau,\omega)|} \geq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

where $|\boldsymbol{X_i}(\tau, \omega)|$ and $|\boldsymbol{Y}(\tau, \omega)|$ indicate the magnitude of the STFT values, of $\boldsymbol{X_i}$ and $\boldsymbol{Y}$ respectively, at frequency ω and time frame τ.

We obtain the STFT magnitude values of separated sources by multiplying the STFT magnitude of the mixture by the estimated masks $\hat{M}_i$, i.e. $|\hat{\boldsymbol{X}}_{\boldsymbol{i}}| = \hat{M}_i|\boldsymbol{Y}|$. Then, the waveforms of the source signals are recovered by applying the inverse STFT transformation on the predicted magnitude $|\hat{\boldsymbol{X}}_{\boldsymbol{i}}|$ and using the phase of the mixture $\boldsymbol{Y}$.

2.2.2. Convolutional architectures

Convolutions are linear filters that have been known for more than 250 years [Domínguez, 2015] and have been used extensively in many fields of science and engineering, including digital signal processing algorithms. One notable application of convolutions in the image processing field is smoothing, which underlies such algorithms as SIFT keypoint detection or edge detection methods [Lowe, 1999]. In audio processing, convolutions are widely used to model reverberation effects and frequency filtering [Reilly and McGrath, 1995].

In convolutional neural networks (CNNs) we typically use *discrete*

finite convolutions:

$$(x * w)[n] = \sum_{i=-K}^{K} x[i]\, w[n-i],$$

where $x : \mathbb{Z} \to \mathbb{R}$ is the discrete input signal and $w : \mathbb{Z} \to \mathbb{R}$ is a convolutional kernel of size K. In practice, most CNN implementations use a cross-correlation operation instead of a convolutional operation [Mishra, 2019] which is defined as:

$$(x * w)[n] = \sum_{i=-K}^{K} x[i]\, w[n+i].$$

Usually, a typical convolutional layer is represented by not just the convolutional operation, but rather an affine transformation:

$$f(x) = Wx + b,$$

where x is an input vector, W is a learned convolutional kernel and b is a learned bias vector. Conventional CNNs take input data as a multidimensional vector and apply a cascade of several affine transformations at each layer, typically alternated with non-linearities and pooling operations between layers. In modern CNN architectures, the bias term is often omitted due to integration of additional regularization techniques such as Batch Normalization [Ioffe and Szegedy, 2015].

Some important characteristics of CNNs are translation invariance and local connectivity which have always been desirable features in image recognition algorithms.

2.2.3. Autoencoders and Encoder-Decoder framework

An autoencoder is a type of neural network that is used for efficient, unsupervised data representation learning by reconstructing its input. A basic architecture consists of an encoder, which transforms input data

into a latent code, and a decoder, which subsequently does a backward transformation from the latent code to the original data point.

Typically, autoencoders are trained through backpropagation to minimize the reconstruction error

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2.$$

In order to discourage memorization and overfitting, an additional regularization term on the network activations is often added to the loss. One common regularization term could be an L1-regularization over the neuron activation values:

$$\sum_i \left| a_i^{(h)} \right|,$$

where $a_i^{(h)}$ is an activation value of a neuron i at the level h. More advanced regularization terms include the Kullback-Leibler divergence (KL-divergence) which enforces that distributions of different neuron activations are different from one another. We can also constrain the derivatives of the activations to be small in order to ensure the latent codes for similar inputs are close.

Autoencoders have been mostly used for compression, denoising and as a pre-training step for other tasks. Based on the same idea of having an encoder for data-to-latent-code transformation and a decoder for latent-code-to-data transformation, the encoder-decoder framework was used for many different tasks where desired output data did not correspond to the input.

In the audio processing domain, encoder-decoder framework was widely used for denoising and source separation problems. It has also been applied in image processing to solve the segmentation problem, in natural language processing for translation, as well as for visual question-answering.

One particularly interesting architecture under the encoder-decoder framework is called U-Net [Ronneberger et al., 2015]. It was designed to solve a medical image segmentation problem. The principal difference with respect to standard encoder-decoder architectures is the use of skip

connections between parallel levels of the encoder-decoder pyramid (see Figure 2.1).

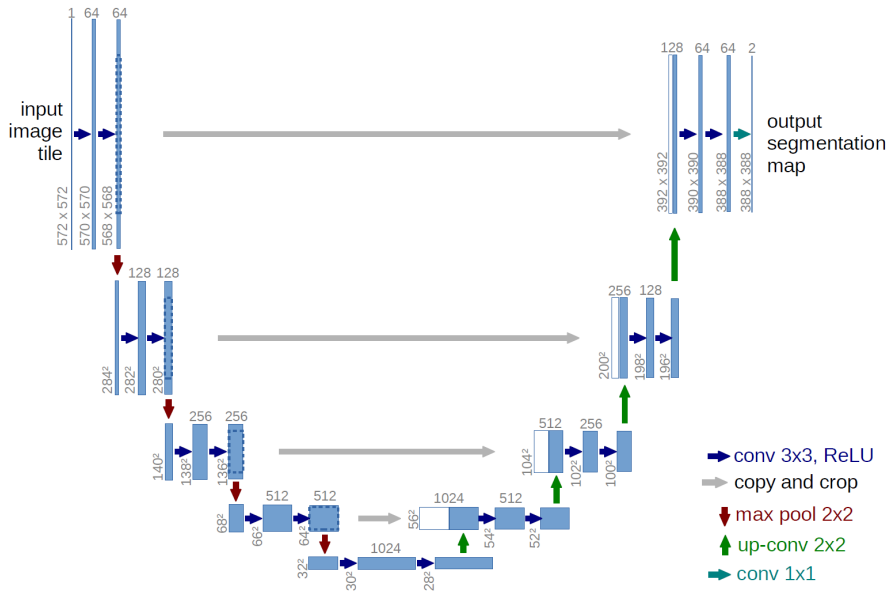


Figure 2.1: U-Net architecture with skip connections. Illustration from [Ronneberger et al., 2015]

The skip connections were proposed earlier for convolutional networks [He et al., 2016a] and have been shown to be efficient for avoiding the problem of vanishing gradient and, consequently, for helping the information to propagate better between different layers, thus preserving global and local feature structures.

2.3. Data representation

In machine learning, if we want our algorithm to learn and generalize well, we must ensure that the data we use for training is informative, not redundant, and easy to interpret. This task is addressed at multiple stages

in a machine learning pipeline, from constructing a dataset to designing an ML-algorithm. At one of the stages, we have to make an important choice about how to represent our data. The underlying data representation is a cornerstone of any machine learning solution. In fact, the art of feature engineering has a share in every successful machine learning technique, and each field has its own specialities when it comes to feature extraction.

In this section, we provide an overview of classical and modern data representations for audio and video data. In the context of this study, the terms *data representation* and *feature* will be used interchangeably unless explicitly mentioned that the term *feature* refers to a non-learnable representation constructed by a deterministic algorithm.

2.3.1. Audio representation

In digital signal processing, a waveform is the principal form to represent audio data. However, this is a very low-level time-domain representation that, until recently, was rarely used directly for analysis in machine learning algorithms. A routine for audio analysis has always included a data transformation pipeline and a construction of a set of interpretable and easy-to-analyze features. An example of hand-crafted features and audio representations is shown in Figure 2.2.

For decades, audio researchers mainly focused on constructing analytical methods for task-specific data representations in audio, speech and music domains. A classical pipeline usually includes two steps:

- a transformation of time-domain data into a time-frequency (TF) representation;
- feature extraction in time and TF domains.

With the advent of modern deep learning techniques, the *learned representations* have become popular. The idea consists of including a dedicated subnetwork that aims to do an optimal feature learning from the underground representation, be it a waveform or a TF representation. Therefore, the feature extraction part is always learnable and can take advantage of, for example, supervised or unsupervised pretraining.

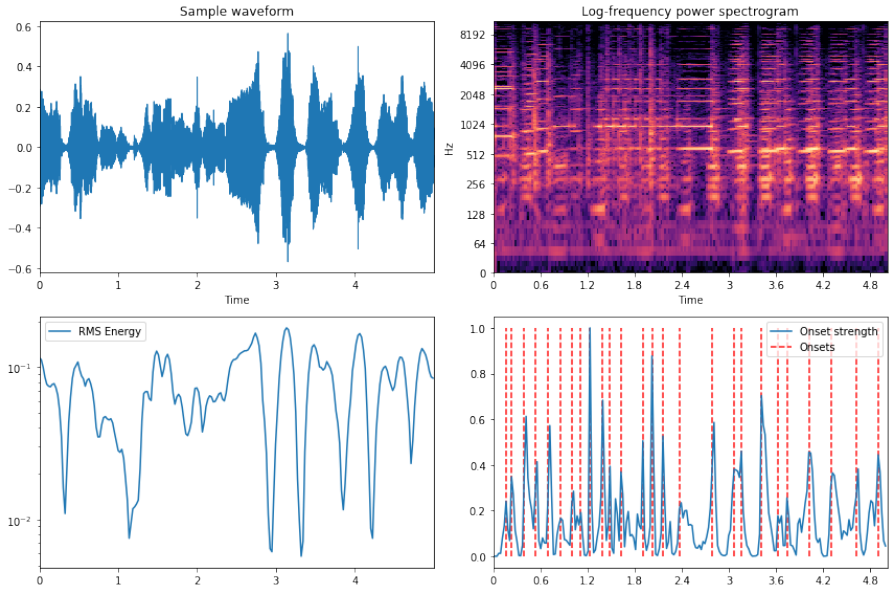


Figure 2.2: Audio representations and hand-crafted features: (a, top-left) a sample waveform, (b, top-right) its log-scaled time-frequency representation, (c, bottom-left) its root mean square (RMS) energy level, (d, bottom-right) aligned onsets.

Time-Frequency transforms

In audio signal processing, several techniques that transform an original time-domain signal into a time-frequency domain are widely used, including Short Time Fourier Transform (STFT), Wavelet Transform, Constant-Q Transform (CQT), and Wigner Distribution Function (WDF). Among them, the STFT stands out as an efficient technique for time-frequency analysis of music signals.

Short Time Fourier Transform (STFT) is a particular type of Fourier transform that allows one to obtain a frequency distribution of the original time-domain signal $x(t)$. In the discrete time case, the transform is defined

as

$$\mathbf{STFT}\{x[n]\}(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}, \tag{2.4}$$

where $x[n]$ is the discrete time signal, $w[n]$ is the window function, m is the new time index, and ω is the angular frequency. The magnitude and phase components of the STFT can be computed and, often in audio analysis, only the spectrogram representation is used which is the squared magnitude of the STFT:

$$\text{spectrogram}\{x(t)\}(\tau,\omega) = |STFT(x(t))(\tau,\omega)|^2.$$

The STFT is usually calculated with some overlap between segments in order to reduce the number of artifacts at the window's boundary. Various window functions can be employed.

Feature extraction

As feature extraction techniques go far beyond the scope of this thesis, we refer the reader to an extensive review of different audio features [Alías et al., 2016]. In this section, we outline only a few features used for the analysis in Chapter 4, namely `loudness`, `onset rate`, and `harmonic-percussive source separation (HPSS)`.

Loudness is the subjective perception of sound pressure [Wikipedia contributors, 2020]. The sound pressure level is a logarithmic measure of the effective pressure of a sound relative to a reference value:

$$L_p = 20\log_{10}\left(\frac{p}{p_0}\right) \text{ dB},$$

where p stands for root mean square (RMS) of the sound signal

$$p = \sqrt{\frac{1}{n}\sum_{i=0}^{n}|x(i)|^2},$$

and p_0 stands for the reference sound pressure.

The empirical relationship between the sound pressure level and human subjective perception is represented by Steven's power law where the power coefficient a for loudness is 0.67, therefore the loudness L is defined as:

$$L = (L_p)^{0.67}.$$

The particular implementation of loudness that we use, utilize the energy of the signal

$$E_s(x) = \sum_{i=0}^{n} |x(i)|^2$$

instead of the sound pressure level [Bogdanov et al., 2013]. The RMS itself can also be used for measuring loudness.

Onset rate is the number of onsets per second. It is based on the onset detection algorithm available in the Essentia library [Bogdanov et al., 2013]. The onsets are computed with two methods:

- The High Frequency Content detection function over the discrete STFT spectrum of a signal $\boldsymbol{X}(t, f)$:

$$\mathrm{HFC(t)} = \sum_{i=0}^{N-1} i|\boldsymbol{X}(t, i)|,$$

  where t is time and f is frequency. The HFC is used to characterize the amount of high frequency content in the signal.

- The Complex-Domain spectral difference function [Bello et al., 2004] takes into account changes in magnitude and phase.

The onsets obtained with both algorithms when postprocessed and the onset rate is computed.

Harmonic-percussive source separation (HPSS) is the algorithm that decomposes the signal spectrogram into harmonic and percussive components using a set of precomputed filters. The HPSS assigns energy of each time-frequency bin according to whether a harmonic or percussive filter responds higher at this TF bin [McFee et al., 2015].

Learned representation

In a deep learning pipeline, we have the option of designing a desired feature representation by learning it from raw data. It gives the advantage of the feature representation being more flexible and optimized with respect to the task that we want to solve. The early review of representation learning techniques [Bengio et al., 2013] highlights its importance to such fields as speech recognition and signal processing, object recognition, natural language processing, multi-task and transfer learning. Nowadays, the learned representation is a practical standard for the aforementioned fields and problems, with more and more researches opting for end-to-end approaches without any pretraining. The two principal ways to obtain a learned representation include *supervised and unsupervised training strategies*.

The supervised training pipeline is often a good choice, as it is easy to construct and generally provides good results. The representation is obtained indirectly by optimizing the loss function of the task of choice. The key assumption underlying this method is that the model learns some relevant characteristics of the data that are transferable to other tasks. One of the advantages of this method is that we do not have to be concerned about the model structure, or to put extra constraints on the hidden representation (e.g. independence of the variables of the learned representation), or to use a special network architecture to enforce the sparsity or convexity of the hidden space.

In the supervised case we can consider two techniques, whenever we train a feature extractor for the target task from scratch or fine-tune a feature extractor trained on an adjacent problem.

The first technique, which is purely supervised learned representation, is predominant in cases when the data is abundant and easy to collect, and often leads to better performance on the reference task. However, it also has a few drawbacks such as:

- the latent representation may have memorized irrelevant characteristics or noise from the data,
- the latent representation may not be optimal in case we want to use

it for another task without fine-tuning; because once a model has found the solution and relevant features for solving the donor task, it will stop, discarding other important underlying data properties that might be useful for the recipient task.

The second technique is used to overcome the data scarcity problem. Adapting a slightly different representation leads to the use of fine-tuning and transfer learning methods which help to adjust the pretrained feature extractor from the reference problem to the target problem. It is important that either the reference problem is as close as possible to the target problem, or the reference problem domain is broader and provides enough diversity in the latent representation.

The unsupervised training pipeline for obtaining a good data representation consists of training a model optimized to solve an artificially constructed problem which reveals the hidden structure of the data. Some examples include data reconstruction using the Encoder-Decoder framework [Fang et al., 2018], autoregressive models trained to predict the next sample in a sequence [van den Oord et al., 2016], reconstructing the order of randomly permuted segments, and so on [Bengio et al., 2013].

There are several aspects that must be taken into account while proposing a model trained to obtain an unsupervised learned representation. First, the feature extractor has to be able to disentangle different aspects of the data. Second, it should not throw away any hidden concepts as we do not know whether they can be important for the target tasks. Those aspects are desirable but not easy to accomplish. For example, naive unsupervised pretrainings can be obtained with the Encoder-Decoder framework, which minimizes the distance between the original and reconstructed objects. However, in this case we have no control whatsoever on the hidden space, just the compression property.

To improve upon this idea, several tactics can be discussed and employed. To give an example, a good representation should be able to capture different valuable characteristics of the data at different scales. In music, we can learn a representation where one feature of it may be associated with onsets (and it can be further used for onset detection) but another one could be linked with tempo, which is a higher level property. At the

same time, another dimension of the same representation could be relevant to the fundamental frequency of given audio (and later be transferred to a melody estimation model), or to its compressed timbre representation for instrument identification. As such, different facets of the data could be learned at different frequency and temporal scales. Both supervised and unsupervised schemes can take advantage of this idea: by adapting the model architecture to comply with that multi-scale constraint in the first case, or by integrating a multi-scale loss function in the second case.

In this thesis, we mostly use the learned representations trained in a pure supervised way with subsequent fine-tuning, and we leave unsupervised pretraining strategies as an important task to explore in the future.

2.3.2. Video representation

A video recording is, in essence, a sequence of static RGB images (frames) taken in a chronological order with an average rate of 25-30 frames per second (FPS). The approaches that use visual information can be broadly classified into two categories: (1) those consider a video as a set of frames, (2) those make use of temporal information already in the initial layers. In the fist case, an algorithm analyzes spatial content of each frame independently and the extracted information is aggregated later. In the second case, a method is explicitly designed to utilize temporal evolution of the visual content in a sequence.

Considering our tasks of interest, namely instrument classification and source separation, we provide related approaches for video data representation. We make connections to the following relevant computer vision areas: object recognition, human activity recognition and motion description from RGB data.

Spatial data representation

There is a direct analogy for the instrument classification task in MIR that is the object recognition problem in computer vision. It has been typically addressed in the image domain and several solutions have been

developed. In very general terms, a classical pipeline consisted of extracting appearance characteristics from an image, converting them to a feature vector, and constructing a statistical model capable of capturing important patterns in the feature vectors and matching a new vector to a predefined set of object categories. In particular, sparse *local descriptors* (or *interest points*), such as SIFT [Lowe, 1999] defined the success of object recognition methods in the early 2000s.

We outline the computation skeleton of SIFT as follows: (1) first, the keypoints candidates (scale-space extrema) of an image are detected by convolving the original image with Gaussian filters at different scales, (2) then many low-contrast extrema are filtered out, (3) at the next step each keypoint is assigned to one or more orientations based on local image gradients, and (4) finally, the keypoint descriptor is computed from the orientation histograms of the keypoint neighborhood.

Most local descriptors widely adopted in computer vision are invariant to scale and rotation, and stable under illumination changes making this group of methods robust. These properties gained them certain popularity, especially for the object recognition and pattern matching tasks. However, SIFT descriptors are easily affected by changes of point of view and are not stable to background changes. Thus, with the great success of convolutional neural networks in object recognition, the underlying learned features have become the new standard for image data representation.

Usually, researchers use one of the popular computer vision architectures [Voulodimos et al., 2018] pretrained on ImageNet [Deng et al., 2009] and either fine-tune the learned representation or employ it directly.

Spatio-temporal data representation

Unlike the spatial representation for images, in video space we can capture not only appearance characteristics of objects, but also describe interactions between them. The primary type of interaction that we have in our data is a person playing a musical instrument. On one hand, as we only track a single activity, it bounds us again to the task of object recognition.

On the other hand, as we have more than a single visual sample per segment, the use of spatio-temporal representation can improve the robustness of instrument detection. For the source separation problem, even though the activity type remains the same, the complexity of interaction and the amount of detail that we have to track and synchronize in order to facilitate the task require rigorous attention to both appearance and motion analysis.

In the area of human activity recognition, several kinds of spatio-temporal data representation have been employed, such as sparse spatio-temporal interest points that generalize 2D descriptors to 3D volumes, motion trajectories, motion energy and motion history images. Several extensive reviews of hand-crafted and deep learned spatio-temporal features for human action recognition can be found in the literature [Aggarwal and Ryoo, 2011, Zhang et al., 2019].

Spatio-temporal features that represent actions can be automatically learned from video data and, as in the case of images, be more robust to camera movements, occlusions, and complex scenes. Following [Aggarwal and Ryoo, 2011], we adapt the classification taxonomy based on the way the video sequence is treated. We classify video representations into three types: sequential, space-time and hybrid approaches. The sequential approaches treat each frame independently and describe the activity by analyzing changes in a sequence of individual frame representations. The space-time approaches treat an input as a joint 3D volume and operate on it directly, extracting motion or spatio-temporal features. The hybrid approaches first extract spatial or spatio-temporal features from individual frames or 3D volumes, and then analyze them jointly.

The examples of sequential approach are Single Frame and Late Fusion CNN architectures from [Karpathy et al., 2014], Convolutional and Late Pooling architectures from [Ng et al., 2015], and a number of hybrid CNN-LSTM approaches where a per-frame spatial representation is obtained with a CNN, and the analysis is performed with an LSTM network [Hori et al., 2017, Wu et al., 2015].

The representative work in the space-time approach include 3D CNN architectures (C3D, [Tran et al., 2015]), Slow Fusion CNN from

[Karpathy et al., 2014], and Pseudo-3D ResNet [Qiu et al., 2017].

Most two-stream networks use a single RGB image and an Optical Flow image ([Simonyan and Zisserman, 2014, Feichtenhofer et al., 2016]) for each stream and therefore represent the hybrid approach. Other examples include an approach combining C3D features and RNN analysis [Montes et al., 2016], 2D-CNN, Optical Flow and LSTM analysis [Ng et al., 2015, Li et al., 2018b].

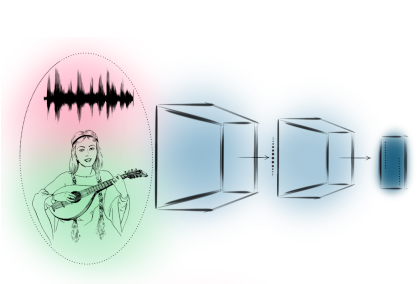
2.4. Multimodal data fusion

In this section we outline several important concepts of multimodal information fusion which we use over the thesis. Throughout the years of development of multimodal algorithms, several principal paths for joining multimodal information have been developed. Depending on how much the data from individual modalities are processed before being mixed together, several fusion strategies can be distinguished (see Figure 2.3):

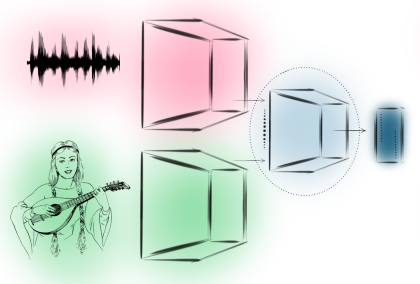
- *early fusion*, when raw or minimally processed data are combined;
- *hybrid (a.k.a. slow, joint) fusion*, when mid-level embeddings or features are combined;
- *late fusion*, when decision-level embeddings are combined.

Each fusion strategy has its areas of applicability. Early fusion is most suitable for data aggregation from the same domain or in case of similar structure when two modalities have direct relations between values (e.g. in two time series for audio loudness and light intensity sampled with the same frame rate). Late fusion is appropriate in most cases when the data is heterogeneous and the final decision has to be made based on high-level characteristics extracted from each modality.

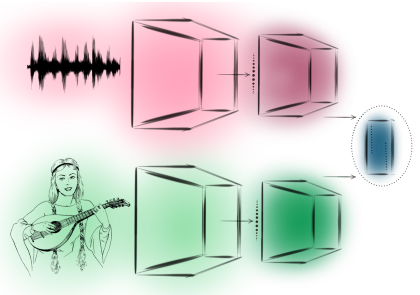
Hybrid fusion strategies, especially with fine-tuned embeddings, seem to be more flexible. It could be a silver bullet choice for many tasks where more than two modalities have to be fused, e.g. audio, visual-spatial, visual-temporal, and context information. It allows one to integrate



(a) Early fusion



(b) Hybrid fusion



(c) Late fusion

Figure 2.3: Common strategies for multimodal data fusion. The early fusion in (a) combines information from two modalities at a raw data representation or at an early stage of joint training. The hybrid fusion in (b) preprocesses the data streams individually and passes the features to a decision learning block. The late fusion in (c) has two separate network to process each modality independently and combine the learned decisions or high-level representations. Darker colors indicate deeper stages of learning and higher levels of representation. The dotted circles mark the joining phase for each strategy.

different representations at different stages of the training process and thus provides greater flexibility.

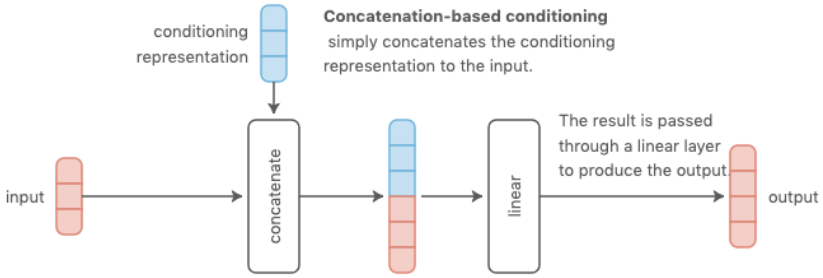
2.4.1. Conditioning techniques

Multimodal data fusion techniques can also be classified by the way of combining the representation vectors. Three methods that we use in this work include concatenation, multiplication, and FiLM [Perez et al., 2018] conditioning (see Figure 2.4 and Figure 2.5). We further explain these concepts under the *conditioning framework* and with an assumption that the data from two modalities are being integrated. Following [Dumoulin et al., 2018], we use the terms *content representation* and *conditioning representation* to indicate the primary data stream and an auxiliary data stream, respectively. Notwithstanding, the described techniques can be applied for more than two modalities and does not necessarily favor one representation over another.

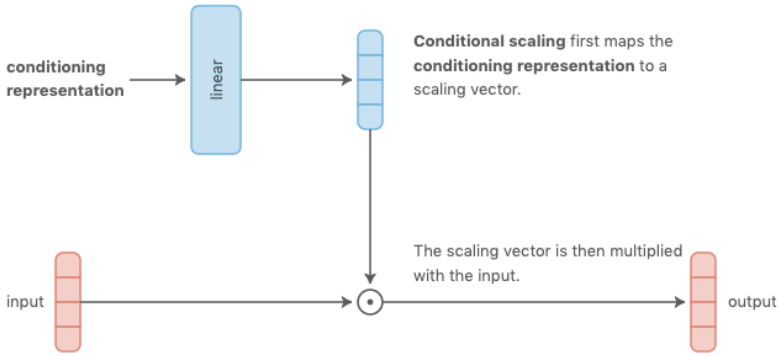
Concatenation

Concatenation-based conditioning (see Figure 2.4(a)) is a method that simply concatenates vectors representing different modalities. It is a parameter-efficient solution for the data fusion as the number of parameters in the layer, following the concatenation step, increases linearly with respect to the input. It is especially efficient in case of a relatively small conditioning vector and a significantly bigger content representation.

However, as noted in [Dumoulin et al., 2018], this method implies some domain knowledge on the explicit position where "the model needs to use the conditioning information." It is fair to use the concatenation-based conditioning, while holding this assumption, for the late fusion technique. Yet, it may not be optimal for early fusion. One possible solution is to concatenate the vector representations at each layer of the models, which brings us to a kind of hybrid fusion. However, in this case we would lose the efficiency of the method as we would need to concatenate a comparable-size representations of the modalities which would eventually lead to the overhead in the number of parameters and the training cost.



(a) Concatenation-based conditioning.



(b) Multiplication-based conditioning, or conditional scaling.

Figure 2.4: Concatenation-based and multiplication-based conditioning. Illustration from [Dumoulin et al., 2018].

Multiplication

Multiplication-based conditioning (see Figure 2.4(b)) is another method to obtain a joint multimodal representation. In order to mitigate the common problem of dimensionality mismatch, the conditioning vector can be passed through a linear layer. The obtained coefficients are then used for an element-wise multiplication with the content representation.

FiLM conditioning

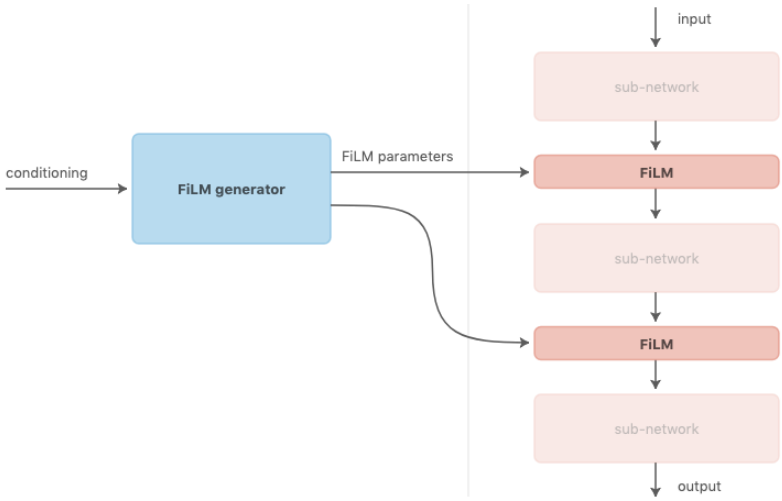
Another possibility to modulate activations of the content network by a conditioning vector extracted from another modality is known as **F**eature-w**i**se **L**inear **M**odulation (**FiLM**) [Perez et al., 2018]. The conceptual idea of FiLM conditioning is simple: it takes a set of learned features and scales and shifts them accordingly to a context vector. Scaling and shifting parameters (γ, β) are learned based on an input context vector $\mathbf{c}$ by an arbitrary function f which is called FiLM-generator:

$$(\gamma, \beta) = f(\mathbf{c}). \tag{2.5}$$

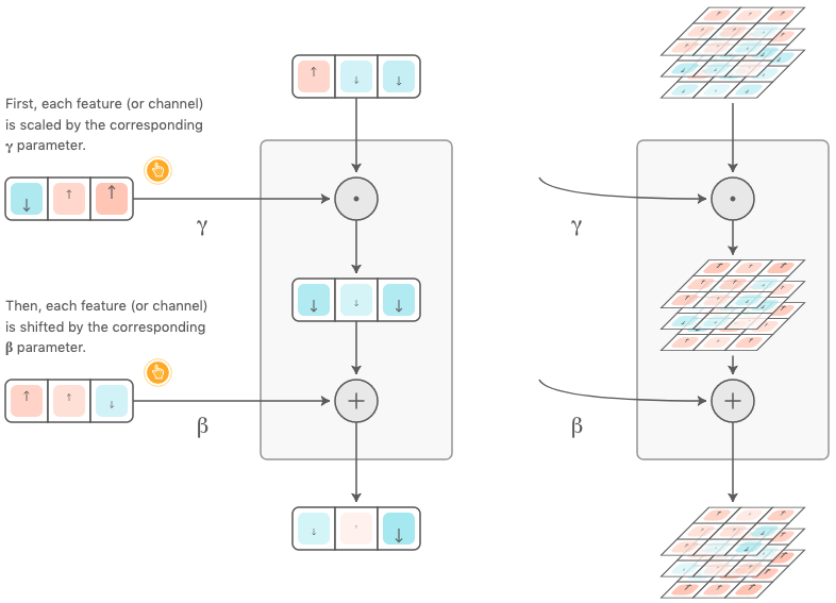
The learned parameters modulate a neural network's activations F_i, where i refers to a feature or feature map, via a feature-wise affine transformation:

$$FiLM(F_i|\gamma_i, \beta_i) = \gamma_i F_i + \beta_i. \tag{2.6}$$

As noted in [Dumoulin et al., 2018], concatenation-based conditioning can be reformulated as conditional biasing. Therefore, the FiLM conditioning can be considered as a generalized form of both concatenation-based and multiplication-based feature transformation techniques as it provides the shifting bias as the concatenation conditioning and the scaling coefficients as the multiplicative conditioning.



(a) Feature-wise Linear Modulation (FiLM) and a FiLM-ed network.



(b) FiLM, modulating activations of a a convolutional layer.

Figure 2.5: Feature-wise Linear Modulation (FiLM). Illustration from [Dumoulin et al., 2018].

Chapter 3

Audio-visual machine learning: tasks, approaches and challenges

3.1. Introduction

Machine learning has evolved very fast during the last two decades. The progress in supervised machine learning, especially with the wide adoption of deep learning techniques, is overwhelmingly impressive. Some examples include automatic speech recognition that advanced from an early attempt to recognize a limited set of utterances [Bahl et al., 1983] to end-to-end noise-robust automatic speech transcription [Chan et al., 2016, Chung et al., 2017]. Similarly, in image recognition the gained difference between early techniques and modern achievements is as between handwritten black-and-white digit recognition [LeCun et al., 1989] and 1000-category image categorization where algorithms surpass an ordinary human performance level [Beyer et al., 2020]. Problems, that were previously unthinkable, such as realistic image [Karras et al., 2019] and music [Dhariwal et al., 2020] generation, have been successfully addressed.

There are several principal causes for that outbreak, which include the availability of diverse and large-scale datasets, the scalability of existing machine learning methods together with advances in cloud and accelerated computing and algorithmic enhancements that allow for a better

generalization ability.

The most impressive results have been achieved for problems that only operate on a single modality of data but multimodal machine learning was indicated as one of the areas where the next breakthrough in world understanding is expected [Jordan and Mitchell, 2015], and numerous approaches have recently emerged. For example, different deep learning architectures have been proposed for audio-visual speech recognition [Ngiam et al., 2011, Huang and Kingsbury, 2013, Hu et al., 2016], audio-visual emotion recognition [Kim et al., 2013, Zhang et al., 2016, Xu et al., 2016, Pang and Ngo, 2015], cross-modal representation learning [Aytar et al., 2016] or image classification and retrieval using images and text [Srivastava and Salakhutdinov, 2012, Sohn et al., 2014].

In this chapter we aim to provide an overview of audio-visual machine learning, especially focusing on audio-visual music information retrieval (AV MIR). We begin with providing a historical perspective of the field of audio-visual machine learning and different types of problems which were generally of interest to the community. Next, we provide an overview of AV tasks and approaches in the field of MIR and indicate the tasks of interest. This is followed by a state-of-the-art review of unimodal approaches for the selected tasks. Lastly, we discuss methods that allow the transition from a single modality to a multimodal approach, such as representation learning, joint training and data fusion techniques.

3.2. Historical perspective and problems

AV speech processing. A relatively long history of audio-visual signal processing originates in one practical problem of automatic speech recognition (ASR). The fact that recognizing speech is especially challenging in a noisy environment has driven research in audio-visual ASR for more than three decades [Finn and Montgomery, 1988]. Likewise, the first attempt to integrate audio and visual information for ASR with feed-forward neural networks goes back to late 1980s [Yuhas et al., 1989], already making use of the backpropagation algorithm.

The interest in AV methods in speech processing has been strong and steady since then, with various AV approaches proposed for speech recognition, enhancement and separation, mostly based on probabilistic graphical models such as Gaussian Mixture Models (GMMs) [Hershey et al., 2004], Hidden Markov Models (HMMs) [Dupont and Luettin, 2000], and Dynamic Bayesian Networks (DBNs) [Nefian et al., 2002].

In recent years, the use of neural networks for audio-visual speech processing has received undivided attention from researchers. Representative studies include AV speaker-independent speech separation [Ephrat et al., 2018], lip reading [Chung et al., 2017, Michelsanti et al., 2020] and even more advanced end-to-end AV speech recognition [Ngiam et al., 2011, Afouras et al., 2018].

AV source localization and separation. Among other research areas in audio-visual signal processing, generic source localization and separation has received significant attention. The primary motivation for this field is that not only audio and visual information are related, but often we can see or imagine the source of origin for every particular sound. In the physical world, they have a causal relation, so that in many cases a sound is produced by an object with a certain visual appearance. In the literature, we can commonly find the term *audio-visual object (AVO)* [Llagostera Casanovas et al., 2010, Parekh et al., 2019b] which emphasize causal relations between audio and visual data, contrasting it from just simultaneously happening events. In addition, the study of some misattribution effects (i.e. ventriloquism) has shown that people tend to relate audio and visual events if they happen simultaneously.

Having this in mind, a correlation-based approach for source localisation was proposed as early as in 2000 [Hershey and Movellan, 2000]. It consisted in calculating intensity changes in audio and video and computing correlations between audio and every pixel in a sequence of frames. The authors showed that the method can successfully identify the speaking person at every time frame in videos of two people speaking in turns.

Concurrently, another system that constructs a two-dimensional spatial likelihood function for sound-based localization and vision-based localization was proposed [Aarabi and Zaky, 2001]. The multimodal integration

was modeled as a weighted linear combination of results from individual modalities. It is worth noting that, while being precise, the method relies on multi-camera and multi-microphone setups for the underlying subsystems.

As a continuation of the research line of a single-view synchronized audio and video, Kidron et al. presented a method that detects pixels associated with a sound source while filtering out other dynamic pixels [Kidron et al., 2005]. The method uses a refined version of canonical correlation analysis (CCA) and, in contrast to previous studies, which mostly focus on speech applications, it can handle different types of sounding sources, not only people speaking but also musical instruments being played. The authors also discuss the *chorus ambiguity* phenomenon when several people sing in synchrony, and in this particular case they accept the detection of any of the faces as a successful result. The main concern raised by the authors is the extreme *locality* of the pixel regions associated with an audio event which they overcome by introducing a sparsity constraint. That work was further extended in [Barzelay and Schechner, 2007], incorporating temporal information for matching visual and audio onsets.

As for the source separation, another generic approach was proposed and tested for speech and musical instrument sounds [Llagostera Casanovas et al., 2010]. First, the authors decompose audio and video signals into two sets of sparse atoms, and compute correlation scores between energy peaks in audio and video atoms, identifying connected sources. Next, spectral GMMs are constructed on segments where only one source is active. Lastly, the trained GMMs are used to separate the mixture.

The field of visually assisted source separation and source localization was notably uplifted with the breakout of deep learning techniques [Ephrat et al., 2018, Owens and Efros, 2018, Lu et al., 2019, Parekh et al., 2019b, Xu et al., 2019], in particular, with explicit focus on musical data [Zhao et al., 2018, Zhao et al., 2019, Gao et al., 2018, Gao and Grauman, 2019, Xu et al., 2019].

AV classification. Historically, general-purpose audio-visual classification was driven by a practical problem of annotation of large-scale

video collections (such as TV shows, movies, news programs) and a corresponding retrieval task. Following the prevalent direction in audio-visual speech processing, many early models were based on HMMs and GMMs [Nam et al., 1998, Jinqiao Wang et al., 2006]. The representative cases include violence detection in movies and drama (for example, gunshots, blood and dynamic activity) in [Nam et al., 1998], where a Gaussian model was fitted with a set of low-level features extracted independently from each modality. Later, a a grammar-based taxonomy for different video shots and a template-matching system were proposed to classify a video segment into a particular category [Carrive et al., 2000]. In this work, the authors take into account both visual and auditory cues while constructing the templates (for example, they suggest using a jungle detection for news), but do not analyse any particular low-level features.

In [Zhang and Kuo, 2001], an observation was made that even if visual shot boundaries may imply the presence of two shots, the audio information can contradict, indicating that two shots are within the same performance and the segments should be treated jointly. That motivated the authors to propose a system for audio-visual content classification based on audio data analysis only. Their statistical rule-based method operates on a set of low-level audio features and successfully distinguishes between speech, music, songs, silence and environmental sounds in video programs, allowing automatic segmentation and indexing.

As in the last decade the amount of audio-visual content generated by users of various online platforms keeps increasing, and the variety of the content is increasing too, the focus of audio-visual classification research has shifted from TV-oriented problems to more general-purpose video classification. The whole area of audio-visual learning has gotten a significant boost. Along this line of research there are works focused on representation learning with further applications in classification, action recognition and source localization [Aytar et al., 2016, Arandjelovic and Zisserman, 2017, Arandjelovic and Zisserman, 2018, Senocak et al., 2019, Korbar et al., 2018, Gao et al., 2019, Liu et al., 2019, Parekh et al., 2019a]. Most of them combine features from two-stream networks (with one tower processing the audio modality and another

one processing the visual modality) either by concatenating them or by having an additional attention module. Some of them employ time synchrony for the samples of the same video [Owens and Efros, 2018, Korbar et al., 2018, Arandjelovic and Zisserman, 2018], while others learn to extract features by identifying if the audio sample corresponds to a given visual data [Senocak et al., 2019, Korbar et al., 2018, Aytar et al., 2016, Arandjelovic and Zisserman, 2018]. More recent work also focuses on the usage of audio for distilling redundant visual information to reduce computational costs [Gao et al., 2019].

Due to increased variety and volume of data, it is beyond the bounds of possibility to annotate them manually and use supervised methods. Therefore, unsupervised audio-visual classification became of special importance, including self-supervision [Aytar et al., 2016, Patrick et al., 2020] and labelling via clustering [Asano et al., 2020].

3.3. Audio-visual MIR tasks

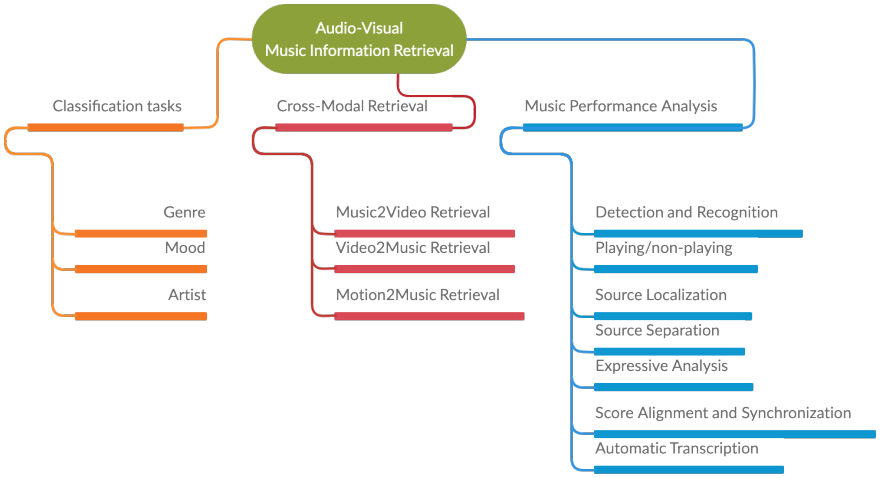


Figure 3.1: A taxonomy of common audio-visual tasks and research problems in music information retrieval.

In Section 3.2 we discussed the progress in general-purpose audio-visual signal processing, in particular, in the speech domain. The music domain has its challenges and particular qualities. For example, from the technical acoustic perspective, musical signals have wider frequency range, more clear pitch, and pronounced rhythm. Different musical signals often overlap in time and frequency, which makes it more challenging. From the visual perspective, the sources of musical signals are much more varied: there are five families of musical instruments in Hornbostel-Sachs classification [Von Hornbostel and Sachs, 1914], and an extensive list of musical instruments keeps growing as people create new ones.

Recently, the discussion of relevant problems and the number of relevant research in the field of AV MIR have grown quite notably, resulting in two significant outcomes such as a recent tutorial in audio-visual music processing [Li et al., 2019] and an overview of audio-visual methods for musical performance analysis [Duan et al., 2019]. In this section, we aim to provide an overview of research problems addressed by the MIR community, proposing a tentative taxonomy of already existing tasks (see Figure 3.1) and extending previous reviews with recent developments. It is worth noting that a substantial part of audio-visual MIR research is focused on the analysis of musical performances, and this is also the main topic of this thesis. In this section, we first discuss the subtasks in this field (in general and with more details for musical instrument recognition and source separation), and later continue with a summary of research in music video analysis, mostly seen as classification and cross-modal retrieval problems.

3.3.1. Audio-visual analysis of musical performances

Audio-visual detection and recognition

As it is true for a generic case, the primary motivation of using both audio and video sources for musical instrument recognition is based on two pillars: the data complementarity and the human ability to aggregate multi-sensory data [Cao et al., 2019]. Another important point is that not all music performance videos are perfectly clean and well-segmented.

Most of them (especially those available online) are noisy and sometimes contradictory, therefore the use of multiple sources of information for feature extraction can reinforce the initial decision and make recognition more robust.

For this reason, several audio-visual recognition methods have been proposed in a deep fully-supervised framework [Slizovskaia et al., 2017, Liu et al., 2019], deep multiple instance learning framework for joint instrument recognition, localization and separation [Parekh et al., 2019b], and deep weakly-supervised framework [Arandjelovic and Zisserman, 2017, Liu et al., 2019] for recognition and activity detection.

In one way or another, all aforementioned work use multi-stream audio-visual neural networks, trained to minimize either categorical cross-entropy on the predicted labels [Slizovskaia et al., 2017, Parekh et al., 2019b, Liu et al., 2019], or solving an audio-visual correspondence task [Arandjelovic and Zisserman, 2017] with the following use of the learnt concepts for audio classification.

Audio-visual playing/non-playing task

When it comes to complex scenes, such as orchestra performances, the task of detecting playing and non-playing activity of each instrument can become very difficult, given that it is common to have several parts played by the same or different instruments or a group of instruments. The task of detecting the playing activity in video recordings was referenced in [Bazzica et al., 2014] as an auxiliary step for the score alignment problem. In the follow-up work by the same authors [Bazzica et al., 2016], an integral multimodal method was developed for playing/non-playing activity detection in symphonic music videos.

To overcome the intra-class variability issues, mentioned in the previous work, and the lack of annotated data, a weakly-supervised approach for playing activity detection was proposed in [Liu et al., 2019]. The authors train a latent movement model for 9 different musical instruments, using video-level instrument labels as a helper target function for audio-based, image-based and optical-flow based subnetworks.

Audio-visual source localization

The task of source localization as a primary problem has been of interest to researchers for a long time [Kidron et al., 2005, Barzelay and Schechner, 2007] in both speech and music context.

More recent research which focuses solely on chamber musical performances [Li et al., 2017] explores the association of musical scores with their spatio-temporal visual locations in video recordings. First, the authors perform audio-score alignment based on chroma features and Dynamic Time Warping, therefore automatically obtaining video-score alignment. Next, they use optical flow to compute bow strokes motion velocities and correlate them with audio onsets. The further video analysis consists in fitting a GMM for player detection and computing a histogram of motion magnitudes for fine-grained localisation of a high-motion region.

While not conducting a dedicated study on source localization, several methods that use CNNs in a self-supervised or weakly-supervised setup report that a coarse localization map of a sounding source can be obtained as a by-product of training audio-visual networks [Arandjelovic and Zisserman, 2018, Liu et al., 2019].

Audio-visual source separation

Audio-visual source separation methods in the context of the music performance domain have received a lot of attention recently, being addressed as a multi-modal matrix decomposition problem [Parekh et al., 2017], a self-supervised problem under the encoder-decoder framework [Zhao et al., 2018], or as a weakly-supervised problem making use of both CNNs and matrix decomposition [Gao et al., 2018, Parekh et al., 2019b].

Parekh et al. look for sparse motion patterns which are similar to audio activation matrices obtained with Non-negative Matrix Factorization (NMF) [Parekh et al., 2017]. In particular, from the visual modality, the authors compute frame-wise average magnitude velocities of clustered motion trajectories. Then, a linear transformation which transforms the motion velocity matrix into the spectral activation matrix is used to constrain the non-negative least square cost function together with a sparsity constraint.

Both NMF and the audio-motion transformation are jointly optimized. The results show a noticeable drop in signal-to-distortion ratio (SDR) while going from Duos to Quartets (from 7.14dB to 0.67dB for the best method while using soft masks for reconstruction). As [Li et al., 2017], the proposed method has troubles separating sounds of the same instrument while addressing this problem for the first time. Interestingly, the authors only focus on the motion component of videos ignoring other visual characteristics such as shape, color, and texture.

Deep learning methods have been widely adopted for AV source separation [Zhao et al., 2018, Zhao et al., 2019, Gao et al., 2018, Gao and Grauman, 2019, Xu et al., 2019]. Starting with capturing only visual appearance features [Zhao et al., 2018, Gao et al., 2018, Gao and Grauman, 2019, Xu et al., 2019] there is a shift towards capturing and integrating motion data and analyzing videos at a higher frame rate [Parekh et al., 2017, Zhao et al., 2019, Gan et al., 2020]. As a primarily audio source separation network, most of them employ the U-Net architecture and operate on a 2D STFT representation of the audio signal.

Notably, most of the approaches listed above are focused on separating only two sounding sources, while many musical performances have more instruments playing in synchrony. Two of the antecedent works in audio-visual source separation explore approaches which can be applied for estimating multiple sources [Gao and Grauman, 2019, Xu et al., 2019], separating one source at a time. However, they have only been trained on *artificial* mixtures of up to 4 sources and *real* mixtures of 2 sources. The separation enhancement scheme proposed in [Xu et al., 2019] consists in extracting one source at a time from a residual audio mixture while considering maximum visual energy at every step, which follows the idea proposed in [Kavalerov et al., 2019]. Authors train the network with mixtures of 2 and 3 instruments, and test it on mixtures of up to 5 instruments.

Audio-visual expressive analysis

As there are several studies on music perception indicating the importance of visual perception in overall evaluation of expressiveness [Platz and Kopiez, 2012] and quality [Griffiths and Reay, 2018] of musical performances, audio-visual expressive analysis probably will be one of the important future research directions in MIR. So far, individual studies on expressive analysis of different musical instruments have been conducted, such as bassoon and saxophone [Dahl and Friberg, 2007], violin [Visentini et al., 2011, Zijl and Luck, 2013], and piano [Thompson and Luck, 2012]. Moreover, a multi-modal dataset and a corresponding analysis of ensemble expressive performance in string quartets was published [Marchini et al., 2014].

However, given that the task is highly subjective and the use of unsupervised techniques is not straightforward, to the best of our knowledge, the problem has not been addressed with deep learning methods yet.

Audio-visual alignment, score alignment, synchronization

The tasks of synchronization and score alignment are common in MIR domain and can be solved, for example, using Dynamic Time Warping (DTW) [Wang et al., 2015]. Nevertheless, there are several use-cases, where audio-visual techniques can be used. Some examples include synchronization of multi-camera recordings of orchestras, synchronization in a computer-aided performances (for example between a musician and a robotized accompaniment, or highlighting fingering in the educational context), and synchronization in distributed musical performances conducted via a web-based video streaming service or in a post-processing stage.

In the context of music education, Shaffer and Pletzer propose the usage of an audio-visual playback, consisting in a reference audio and fingering visualization, being played at an adaptive speed and therefore taking into consideration the actual speed of student's performance [Shaffer and Pletzer, 2009].

An audio-visual method for multi-camera video recordings synchronization was proposed in [Shrestha et al., 2010]. The approach is based

on detecting and matching audio and video features, such as flashes, audio onsets and audio fingerprints. The proposed method was tested, in particularly, for concert recordings and a piano recording.

An audio-visual score synchronization method was proposed in [Bazzica et al., 2014], exploring the robustness of playing/non-playing (P/NP) instrument-wise labels extracted from audio and video independently. The P/NP scores from corresponding modalities are synchronized with each other using the DTW algorith, and then used for the final score alignment. Similarly, a video to score alignment via audio-visual onsets was used in [Li et al., 2017] for associating sound tracks to players in chamber music performance videos.

In the context of computer-aided performances, Lim et al. detect audio onsets and flutist's movements to help a robot to perform in synchrony with a musician [Lim et al., 2010]. In the same direction, an audio-visual score following system was proposed in [Maezawa and Yamamoto, 2016] for inter-musician coordinating in ensemble performances, where one of the "musicians" is a robot playing an accompaniment part. The method uses an HMM model where prior probabilities are computed from musical scores and posteriors are obtained analysing CQT and changes in Optical Flow.

Finally, in the context of networked musical performances, two audio-visual synchronization systems were proposed, with a special focus on performance assessment [Humphrey and Gryner, 2015] and on facilitation of the rehearsal process [Bell, 2018].

Audio-visual transcription

Audio-visual and visual transcription of solo performances is another subfield of AV MIR. The transcription problem is a meaningful task by itself, especially for noisy sounds and multi-instrumental performances, and it has direct applications in music teaching and synthesis as well.

Each musical instrument has a very particular way of playing it though, so no generic method has been proposed till now. However, as in the case with expressive analysis, a number of experiments has been carried out

for different types of instruments. The individual studies include visual violin transcription [Wang et al., 2007], AV methods for transcription of drum strokes [Gillet and Richard, 2005, Marenco et al., 2015, Bhalerao et al., 2020], solo guitar [Goldstein and Moses, 2018], solo clarinet [Zinemanas et al., 2017], and piano transcription [Gorodnichy and Yogeswaran, 2006].

There is still room for improvement, and as it has been shown that semi-supervised audio-visual deep learning models can perform well in piano transcription [Koepke et al., 2020], we can expect more generic methods in the near future, especially for high-resource musical instruments, such as piano, violin, drums, and guitar.

3.3.2. Audio-visual analysis of music videos

Audio-visual classification

Speaking of audio-visual recordings in the music domain, so far we have been explicitly focusing on videos of instrumental performances. However, *music videos* also have a significant share in all video recordings, and are of interest to the MIR community as well. This type of content conveys various kinds of information, and the most practical problem, that researchers have been faced with, is audio-visual classification.

The field shares reliable methods with the general-purpose video classification field as discussed in Section 3.2 and Subsection 3.3.1. Thus, an audio-visual approach from [Nanni et al., 2017] combines acoustic features and texture features extracted from spectrogram images for general-purpose audio classification.

Notwithstanding, different taxonomies for individual aspects that can be estimated from music videos have been proposed. Among them, we would like to highlight:

- **multimodal genre classification** [Oramas et al., 2018], [Schindler, 2019];
- **multimodal mood classification** [Sasaki et al., 2015];

- **multimodal artist classification** [Schindler and Rauber, 2015].

Audio-visual cross-modal retrieval and generation

In this subsection, we are going to list several audio-visual problems that are relevant not only to the previously mentioned music videos but also to *dance music videos*. It has been noted [Gillet and Richard, 2006, Tsuchida et al., 2019b] that analysis of the musical structure of this type of content can be helpful for structuring motions and shots in videos, and vice versa. Following the general trend, there is a shift from the convenient analysis methods to deep learning methods. Thus, onset changes and instrumentation changes were employed in an early audio-visual approach for structuring and segmenting music videos [Gillet and Richard, 2006], while deep learning methods became prevalent in more recent studies [Tsuchida et al., 2019b, Su et al., 2020].

Without discussing every method in details, we would like to outline several practical audio-visual cross-modal problems that can be found in the literature in the context of music and dance music videos:

- **creating soundtracks for silent videos** [Su et al., 2020];
- **audio-visual music recommendation systems** [Sasaki et al., 2015];
- **music-triggered dance generation** [Tsuchida et al., 2019b, Lee et al., 2019, Zhuang et al., 2020]
- **dance music retrieval from motion** [Barleycorn, 2019, Tsuchida et al., 2019a].

3.4. Unimodal approaches for classification and source separation

3.4.1. Classification of musical instruments

Audio-based classification

Various tasks in MIR have been addressed with deep learning methods. Among them, we find approaches for musical onset detection [Schluter and Bock, 2014], musical instrument recognition [Han et al., 2016, Lostanlen and Cella, 2016], automatic music transcription [Sigtia et al., 2016], acoustic event detection [Espi et al., 2015, Salamon and Bello, 2017], automatic tagging [Choi et al., 2016], audio source separation [Chandna et al., 2017] and various classification tasks [Pons and Serra, 2017, Hershey et al., 2016].

Research in deep learning MIR has advanced significantly and a number of classification and detection methods has been proposed recently [Pons and Serra, 2019, Fonseca et al., 2019]. Although there are some end-to-end methods working with raw audio [van den Oord et al., 2016, Aytar et al., 2016], they require huge data collections and a lot of time to train. The most common approaches first transform audio data into two-dimensional image-like representations (e.g. Short-Time Fourier Transform (STFT) spectrogram [Espi et al., 2015, Chandna et al., 2017], log mel-spectrogram [Hershey et al., 2016, Choi et al., 2016, Han et al., 2016, Schluter and Bock, 2014] or Constant-Q spectrogram [Lostanlen and Cella, 2016, Sigtia et al., 2016]) and then train various CNN architectures. Besides, most of the architectures are either shallow, consist of only straight layer connections, or exploit squared filter shapes, which came up directly from image processing. In Chapter 4, we explore a few enhancements over traditional models, such as separable convolutions [Chollet, 2016] and partially task-specific filter shapes [Pons and Serra, 2017].

Video-based classification

The breakthrough in pattern recognition on static images was largely due to its impressive feature learning ability. The computer vision community has been struggled for decades to find a way to avoid handcrafted features for solving large-scale video analysis tasks in a unique non-specific way [Tran et al., 2015, Karpathy et al., 2014].

Over the last years, most of the best solutions in action recognition [Simonyan and Zisserman, 2014, Feichtenhofer et al., 2016, Ng et al., 2015, Tran et al., 2015, Caba Heilbron et al., 2015], scene recognition [Karpathy et al., 2014, Abu-El-Haija et al., 2016] and general multi-label video classification [Karpathy et al., 2014, Abu-El-Haija et al., 2016] tasks exploit either deep neural networks on raw spatio-temporal data [Simonyan and Zisserman, 2014, Feichtenhofer et al., 2016, Karpathy et al., 2014, Tran et al., 2015] or combine them with motion features such as improved Dense Trajectories (including HOG, HOF and MBH) [Caba Heilbron et al., 2015, Tran et al., 2015] and Optical Flow images [Ng et al., 2015]. The most straightforward way to incorporate temporal information in video CNNs is to switch from 2D convolutions to 3D convolutions [Tran et al., 2015, Karpathy et al., 2014], although it leads to difficulties in the choice of parameters such as the optimal shape for the filters, the frame-rate for analysis or the clip size, to name a few.

Several alternative methods have been recently proposed, such as two-stream CNNs [Simonyan and Zisserman, 2014, Feichtenhofer et al., 2016], which use single-frame architecture for spatial modeling and precomputed multi-frame optical flow images for temporal modeling, while aggregating information at the prediction stage [Simonyan and Zisserman, 2014] or at several layers of the network [Feichtenhofer et al., 2016]. The approach in [Ng et al., 2015] examines different feature-pooling methods on CNN architectures with up to 120 frames as well as the capability of Long Short-Term Memory networks to catch temporal information. Although this approach provides good results, its computational performance is far from satisfactory. A good compromise between accuracy and speed

for large-scale video classification has been proposed by several teams of researchers [Karpathy et al., 2014, Abu-El-Haija et al., 2016]. They build systems upon frame-level spatial features and exploit average pooling and Deep Bag of Frame (DBoF) pooling for clip-level and video-level predictions. In Chapter 4, we primarily make use of the image classification architectures for extracting appearance features, such as Inception v3 architecture [Szegedy et al., 2016] and ResNet-50 [He et al., 2016b].

3.4.2. Source separation

Audio-based source separation

For many years, a general approach for solving an audio source separation problem would include one of the matrix-factorization algorithms. Independent Component Analysis (ICA) [Hyvärinen and Oja, 2000] and Non-negative Matrix Factorization (NMF) [Virtanen, 2007] are two common techniques used for source separation.

With the recent achievements in machine learning, researchers have started to adopt deep neural network paradigms to address the source separation problem. Since CNNs have been proven to be successful in image processing, raw audio data is often converted to 2D spectrogram images for analysis. The image data is then fed to a convolutional autoencoder which generates a set of masks that can be used to recover sound sources using inverse Short Time Fourier Transform (iSTFT) [Jansson et al., 2017, Chandna et al., 2017, Uhlich et al., 2017].

In Chapter 5, we aim to continue researching on deep learning methods for the source separation problem. Furthermore, we focus on improving the results by experimenting with less conventional approaches. On one hand, we work not only with STFT representation, but directly with raw waveforms as well. This approach is an active research area [Stoller et al., 2018b, Lluis et al., 2018] and gives us an additional advantage of preserving the phase information unlike other CNNs which only use the magnitude of STFT [Chandna et al., 2017, Uhlich et al., 2017]. On another hand, we want to enhance our results through conditioning with

instrument labels and appearance features extracted from video data. This type of guidance has been shown to have a good impact on the source separation performance. Thus, in [Parekh et al., 2017], the authors use visual guidance for improving source separation quality. Additionally, in a concurrent work [Seetharaman et al., 2019], the authors explore a similar idea of class-conditioning over the joint embedded space, but unlike us, they use an auxiliary network to model parameters of a GMM for the final source separation, and they take spectrograms as an input of the model.

In this thesis we experiment with two models based on the U-Net [Ronneberger et al., 2015] architecture, a convolutional encoder-decoder network developed for image segmentation. The U-Net approach has been adapted already for singing voice separation in [Jansson et al., 2017], where this model applies 2D convolutions and works with spectrograms. We make use of the vanilla U-Net with STFT input and masks estimation as well as adapting the Wave-U-Net model [Stoller et al., 2018b]. Instead of doing a 2D convolution, Wave-U-Net performs series of 1D convolutions, downsampling and upsampling with skip connections on a raw waveform signal. This approach was presented at SiSEC evaluation campaign [Stöter et al., 2018] and demonstrated competitive performance.

The input to this network is a single channel audio mix, and the desired output is the separated K channels of individual audio sources, where K is the number of sources present in the audio mix. An interesting aspect of the Wave-U-Net is that it avoids implicit zero paddings in the downsampling layers, and it performs linear interpolation as opposed to de-convolution. This means that our dimension size is not preserved, and our output results will actually become a lot shorter compared to our input. However, by doing this we can better preserve temporal continuity and avoid audio artifacts in the results.

In the experiments with the vanilla 2D U-Net, we utilize masks estimation approach. The mask estimation step has always been an essential component of model-based source separation algorithms [Carabias-Orti et al., 2013, Miron et al., 2016, Parekh et al., 2017, Carabias-Orti et al., 2011, Ozerov and Févotte, 2009, Virtanen, 2007]. Consecutively, the masking-based approach for training neural

networks has received a lot of attention recently and has been very successful in Single-Channel Source Separation (SCSS) [Chandna et al., 2017, Wisdom et al., 2019, Jansson et al., 2017]. While being consistent in the estimation objective, many authors propose additional schemes and techniques with the aim of raising the separation performance. Thus, the work reported in [Wisdom et al., 2019] shows an improvement of 0.7 dB in scale-invariant signal-to-distortion ratio (SI-SDR) metric [Le Roux et al., 2019] by integrating mixture-consistency and STFT consistency constraints into the training pipeline. Despite the fact that most of the existing work estimates binary or ratio masks, the estimation of STFT magnitude values has also been used in practice [Doire and Okubadejo, 2019] together with loss function computation in time-frequency [Stöter et al., 2019] or time domain [Kavalerov et al., 2019] while internally estimating the masks.

It is worth noting that the set of methods which has been successfully used in source separation is very diverse, and the optimal choice of an architecture remains an open research question. Some examples include LSTMs [Luo and Mesgarani, 2018] and BLSTMs [Uhlich et al., 2017, Stöter et al., 2019], fully-connected architectures [Grais et al., 2016], U-Nets [Jansson et al., 2017, Doire and Okubadejo, 2019], GANs ([Stoller et al., 2018a] and [Choi et al., 2017]), as well as combinations of the above proposed by [Uhlich et al., 2017, Kavalerov et al., 2019]. Some research works suggest the estimation of each source separately with a dedicated network [Chandna et al., 2017, Stöter et al., 2019], while other approaches employ one-to-many encoder-decoder networks with a shared encoder and one decoder per source [Doire and Okubadejo, 2019]. Overall, the use of an individual network for each source seems to provide a better performance but it comes at the cost of increased training time.

There have been diverse proposals for loss functions, which include L_2-distance [Chandna et al., 2017, Uhlich et al., 2017], and L_1-distance [Jansson et al., 2017, Doire and Okubadejo, 2019] on estimated spectrograms, L_2-distance on ratio and binary masks [Grais et al., 2016], L_1-distance on ratio masks [Gao and Grauman, 2019], binary cross entropy on binary masks [Zhao et al., 2018, Zhao et al., 2019], as well as nega-

tive SI-SDR [Le Roux et al., 2019, Luo and Mesgarani, 2018] and SNR [Kavalerov et al., 2019] as objective functions.

3.5. From unimodal to multimodal: techniques

Multimodal machine learning aims to build models "that can process and relate information from multiple modalities" [Baltrušaitis et al., 2018]. As it has been discussed in Section 3.2, researches in signal processing community have devoted a lot of attention to multimodal methods for automatic analysis of audio-visual recordings, and a number techniques has been developed. In this section we will mostly discuss methods for audio-visual representation learning and fusion techniques. For getting a broader perspective on multi-modal audio-visual methods an interested reader is referred to classical and recent surveys in the field [Maragos et al., 2008, Atrey et al., 2010, Katsaggelos et al., 2015, Baltrušaitis et al., 2018].

3.5.1. Representation learning

Representation learning is a set of techniques used in machine learning that help to discover a compact feature representation for a given data modality with respect to a given task.

As it has been already discussed in Section 2.3, two common types of data representation in almost any data domain are the handcrafted feature representation and learned representation. The learned representations have become especially popular with the advance of end-to-end deep learning methods, where a specialized data representation is learnt concurrently with solving the task of interest while minimizing the loss function. If a learned representation obtained this way while solving a supervised task is robust enough, it can be later used for solving other similar problems in the data domain.

As not all data can be manually annotated and used in the supervised manner, a field of self-supervised representation learning is of a particular

interest. In self-supervised representation learning models, the goal is to solve a *surrogate task* that does not need an explicit supervision, or need it at the minimum level. Since the initial review was published [Bengio et al., 2013], representation learning has been widely adapted in natural language processing, computer vision, reinforcement learning and quite everywhere [Weng, 2019]. In each domain, researches use inductive biases present in the data, to construct good surrogate problems with minimal supervision. Some examples include next frame prediction and solving a jigsaw puzzles in computer vision domain, contrastive predictive coding (CPC) in speech domain [Oord et al., 2018], and transposition-invariance of CQT representation in music domain [Yesiler et al., 2020].

In audio-visual machine learning, a number of representation learning research has been proposed utilizing the synchrony between audio and video modalities and predicting whether given audio and video segments are temporally aligned [Owens and Efros, 2018]. As training on the synchronization problem has been reported tricky, a simpler task of audio-visual correspondence is of a common use as well [Arandjelovic and Zisserman, 2017, Arandjelovic and Zisserman, 2018].

Different objective functions are exploited in audio-visual deep representation learning such as cross-entropy [Arandjelovic and Zisserman, 2017, Arandjelovic and Zisserman, 2018], KL-divergence [Aytar et al., 2016, Gao et al., 2019], contrastive loss [Korbar et al., 2018] and triplet loss [Senocak et al., 2019]. Distinctively, Korbar et al. [Korbar et al., 2018] also use *curriculum learning* by first training the network with easy examples (correspondence is defined as being sampled from the same video) and then with hard and superhard examples (correspondence is defined as time-synchrony with/without time shift within the same video).

In addition, the idea of visually-guided *co-separation* has been proposed in [Gao and Grauman, 2019]. The method consists in guiding source separation by integrating visual features of a detected musical instrument at the bottleneck of the primary U-Net, while the training is done using mix-and-separate approach with a combination of separation and consistency losses. The latter is defined as a cross-entropy loss between ground truth instrument labels and the predictions obtained with an

additional classifier on the preliminary separated sources.

3.5.2. Fusion techniques

In multimodal machine learning, data fusion approaches are commonly classified as (1) *early fusion* [Hu et al., 2016, Pang and Ngo, 2015, Ngiam et al., 2011], where the network learns hidden representation from concatenated multimodal input; (2) *late fusion* [Xu et al., 2016, Pang and Ngo, 2015, Ngiam et al., 2011], where the networks for all data sources are optimized separately and the learned representations are then combined to model the joint distribution of multiple modalities; and (3) *hybrid fusion* [Feichtenhofer et al., 2016, Karpathy et al., 2014], where the network may have multiple fusion layers and optimize several learning representation simultaneously.

Various integration techniques have been studied in last decades. For example, one of the early works in audio-visual speech recognition [Yuhas et al., 1989], proposes to use a weighted linear combination of audio and visual low-level features. Other representative techniques include audio-visual early feature fusion [Smaragdis and Casey, 2003], multimodal matrix factorization [Žitnik and Zupan, 2014], fusion via a product of affinity kernels, learned individually for the audio and the video data [Dov et al., 2017].

It is still an open problem, which fusion technique is better for a particular task. In [Smaragdis and Casey, 2003], the authors present an approach using early fusion technique. They perform an independent component analysis for dimensionality reduction in audio (STFT) and video (sequence of frames) simultaneously, concatenating a vector of frequencies and a video frame, reshaped into the vector, at the moment t. However, the method has a common issue of all early fusion approaches, namely, it imposes the synchrony in frame rates between audio and video sources, and the approach only works well with static objects and scenes.

Similar uncertainty of how the information is (or should be) integrated is inherent in perception studies. We still do not know how human fuse multisensory information, and contradictory evidence has been reported

in favor of early [Schwartz et al., 2002] and late [Cao et al., 2019] fusion techniques. Moreover, a recent perceptual study in audio-visual data integration shows that if two stimuli do not coincide to each other, the only one source would be used, either from visual or audio cortex [Cao et al., 2019].

In recently proposed methods for audio-visual source separation, several late fusion techniques have been used to combine the data obtained from different modalities, such as late fusion [Parekh et al., 2019b], conditioning at the bottleneck via tile-and-multiply [Gao and Grauman, 2019], concatenation [Ephrat et al., 2018], attention mechanism [Zhao et al., 2018, Zhao et al., 2019], and FiLM conditioning [Dumoulin et al., 2018, Zhao et al., 2019].

Unlike previous studies, in the present work in Chapter 5 we analyse different ways to combine audio and visual information and extend prior work for multiple and unknown in advance number of sources.

Multimodal fusion via FiLM conditioning.

In the previous section we reviewed an existing research line in source separation which combines information from visual and audio modality. It can be reformulated as audio source separation *conditioned* on visual information. We observe that, while there are several strategies of data fusion (i.e. concatenation or co-processing), another possibility is to modulate activations of a primary audio network by a context vector extracted from another modality, which is known as **F**eature-w**i**se **L**inear **M**odulation (**FiLM**) [Dumoulin et al., 2018]. The conceptual idea of FiLM conditioning is simple: it takes a set of learned features and scales and shifts them accordingly to a context vector. Scaling and shifting parameters (γ, β) are learned based on an input context vector $\mathbf{c}$ by an arbitrary function f which is called FiLM-generator:

$$(\gamma, \beta) = f(\mathbf{c}). \tag{3.1}$$

The learned parameters modulate a neural network's activations F_i, where i refers to a feature or feature map, via a feature-wise affine transformation:

$$FiLM(F_i|\gamma_i, \beta_i) = \gamma_i F_i + \beta_i. \tag{3.2}$$

Other studies consider *weak conditioning* in source separation using only labels of target sources [Meseguer-Brocal and Peeters, 2019, Slizovskaia et al., 2019] in contrast to *strong conditioning* where the context vector could be available frame-wise [Tzinis et al., 2019, Schulze-Forster et al., 2019]. The employed weak label conditioning techniques include FiLM [Meseguer-Brocal and Peeters, 2019] and tile-and-multiply [Slizovskaia et al., 2019]. For strong conditioning, a binary vocal activity vector and vocals magnitude vector have been used for singing voice separation with attention mechanism [Schulze-Forster et al., 2019].

Later, the idea has been explored in the context of universal source separation with conditioning on classification embeddings [Tzinis et al., 2019]. First, the method extracts the context embeddings with the classification network, then upsamples and normalizes them, which is followed by conditioning of the primary source separation network either by concatenation with network's activations or gating the activations by the embeddings. Another work goes along this line and train a source separation model based solely on weak labels [Pishdadian et al., 2019]. The method consists in training a classifier network and using the classification loss (with an additional constraint for the estimated sources to sum to the mixture) as the objective function for separation.

We find various strategies to integrate side information, and different modules of the network being conditioned. However, most of the studies inject the context vector at the bottleneck of encoder-decoder architecture with a rare exception of early fusion in [Tzinis et al., 2019]. The same authors [Tzinis et al., 2019] report that integration of the context vector at every layer of the primary network leads to overfitting.

3.6. Conclusion

In this chapter, we discuss the historical perspective of audio-visual machine learning and how different audio-visual methods were first developed for speech processing.

Next, we talk about the growing area of AV MIR, contribute with a tentative taxonomy of the tasks that have been already addressed by the community and list representative examples of relevant research work. We continue with a more detailed survey for the tasks of interest, providing a short overview for classification and source separation problems in a unimodal setup and identifying the individual components that we used in the proposed multimodal methods.

We finish the chapter discussing different approaches for representation learning and fusion techniques which help to develop advanced multimodal algorithms by facilitating the training process, reducing the need for annotated data and making use of inductive biases which are present in the multimodal data.

For the musical instrument classification task, we review relevant audio-only approaches, focusing on the shift from hand-crafted features to deep learning methods. One limitation which we identified in the literature is that the task has not been approached in the audio-visual context, although the research in the adjacent fields suggest that the robustness of the recognition can be improved. We address this problem in Chapter 4, proposing a novel multimodal CNN architecture which brings together the power of computer vision and machine listening. Based on the reviewed work in multimodal aggregation, we have chosen to use the late fusion technique for data aggregation, which was motivated by three particular aspects: (1) previously reported results [Ngiam et al., 2011, Karpathy et al., 2014], (2) the need to catch fairly high-level concepts on top of low-level data (which is to recognize musical instruments based on a set of pixels and audio signal), and (3) perception studies that indicate that multimodal object recognition operates on high-level unimodal representations [Cao et al., 2019].

For the sound source separation problem, we found several weaknesses in the methods proposed in the literature. In particular, even though

deep learning approaches have grown in the popularity and have shown outstanding performance, most of the conventional approaches are trained to solve rather narrow separation tasks, including a limited number of sources. In Chapter 5 not only we propose several extensions of the review models that can operate on multiple sources, but also employ the usage of extra modalities to guide the separation process. Besides, we experiment with both end-to-end and STFT-based architectures. The primary motivation behind the first type of architecture is the possibility to avoid phase loss at the reconstruction step, while the second type is less computationally expensive and allows us to conduct more experiments to find the optimal data fusion strategy for the task.

Chapter 4

Audio-visual music instrument classification

4.1. Introduction

Humans recognize a musical instrument by combining multiple perception modalities. For example, we can distinguish a violin from a cello by its timbre, size, bow movements and relative position of the instrument with respect to the performer's body. Although the task is fairly easy for humans to perform, combining multimodal information is not trivial for machine learning algorithms.

Musical instrument recognition is a well-known problem in the MIR field. State-of-the-art methods are based on the combination of audio feature extraction (representative of the time-frequency distribution of the signal), automatic classification methods and context information on the music material under analysis. Nowadays, these algorithms provide good accuracy in recognizing musical instruments from monophonic audio recordings (i.e. single instrument playing), although the performance depends on the number of instruments and size of the audio collection used for training [Herrera-Boyer et al., 2003]. This performance significantly drops in polyphonic music scenarios (i.e. more than one instrument playing), where it is easier to recognize instruments if they are predominant in

the audio signal [Bosch et al., 2012].

Nevertheless, current approaches are based on the analysis of good-quality audio material and fail for real-world scenarios such as the one addressed here. Moreover, it's typical to find the presence of the sound of the instruments. In contrast, in this thesis, our problem is to recognize the physical presence of the instruments by either sound or visual component. With this method, we hope to advance the field of indexing music videos in large-scale collections.

User-generated videos are widely found on social networks to share users' own musical performances. They may contain multiple instruments, different types of noise, blur, or compression artifacts. In addition, they are varied in terms of recording conditions and quality [Slizovskaia et al., 2016]. Yet, despite these downsides, collections of user-generated videos are a rich source of knowledge. While the most important information comes from audio, visual content also plays an important role in detecting musical instruments in videos. Thus, different taxonomies for musical instruments rely on audio characteristics as well as on visual characteristics (such as a keyboard, wood, brass, bowed string). In this work, we explore the relationship between audio and visual cues and take advantage of complementary information provided by the nature of the task.

In Section 3.4.1 we reviewed audio-based and visual-based approaches for musical instrument classification available at the time the study was conducted. Among the surveyed architectures, we determine our interest in STFT-based convolutional methods for audio processing, sequential frame-based CNNs for video processing and the hybrid fusion method for multimodal coupling.

We train and evaluate a multimodal CNN architecture on two large-scale video datasets: YouTube-8M [Abu-El-Haija et al., 2016] and FCVID [Jiang et al., 2018] which contain more than 60000 and 5000 musical performance videos with musical instruments, respectively. The proposed architectures demonstrate state-of-the-art results in audio and video object recognition, provide additional robustness to missing modalities, and remain computationally cheap to train. In addition, our approach meets the

standards of reproducible research.

Our contributions include: (1) a novel multimodal CNN architecture for audio-visual musical instruments recognition which outperforms unimodal state-of-the-art techniques with the largest musical performance video datasets used in the literature at the time of the study; (2) evaluation of a few recent and popular audio-only and general-purpose CNN architectures in the context of user-generated musical performance videos; (3) both FCVID and YouTube-8M datasets have been constructed for visual concept recognition, this notwithstanding, we show in a set of experiments that audio information plays a crucial role in the categorization of music videos and can significantly improve recognition performance over visual input.

The findings of the study described in this chapter were published as the following stand-alone publications:

- **"Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies"**. O. Slizovskaia, E. Gómez and G. Haro. In *Proceedings of 13th Sound and Music Computing Conference (SMC). 2016.*

- **"Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture"**. O. Slizovskaia, E. Gómez and G. Haro. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR). 2017.*

- **"Correspondence between audio and visual deep models for musical instrument detection in video recordings"**. O. Slizovskaia, E. Gómez and G. Haro. In *The 18th International Society for Music Information Retrieval Conference (ISMIR17), Late-breaking/demo session (LBD). 2017.*

- **"A Case Study of Deep-Learned Activations via Hand-Crafted Audio Features"**. O. Slizovskaia, E. Gómez and G. Haro. In *The 2018 Joint Workshop on Machine Learning for Music, The Federated Artificial Intelligence Meeting (FAIM), Joint workshop program of ICML, IJCAI/ECAI, and AAMAS. 2018.*

4.2. Proposed method

In this section, we describe our models for the task of multimodal musical instrument classification. The schematic illustration of the model is represented in Figure 4.1. The model is a two-stream CNN: (1) the audio subnetwork takes an STFT representation of the audio signal and learn the latent representation of it, (2) similarly, the visual subnetwork takes a sequence of synchronized video frames and process them to get the latent video representation, (3) two data representations then concatenated and jointly processed via a small classification subnetwork to obtain the final predictions.

4.2.1. Visual-based recognition

Recent works [Ng et al., 2015, Varol et al., 2017] report that spatio-temporal features can be better captured with long clips, while for short clips frame-level features have a greater impact on video object recognition performance [Karpathy et al., 2014]. Considering the fact that learning over long clips is a very time-consuming process, we follow the approach from [Abu-El-Haija et al., 2016] and extract frame-level features from videos.

For detecting instruments in static video frames, we experiment with Inception v3 architecture [Szegedy et al., 2016] since it's one of the most prominent and successful ones and it has been shown to provide a notable generalization ability in various tasks [Szegedy et al., 2016, Hershey et al., 2016]. We explore the influence of the total number of frames selected from the videos at the training phase. Moreover, we study the impact of fine-tuning the model over an independent set of images of musical instruments. The pretraining details are provided in Section 4.3.3.

4.2.2. Audio-based recognition

For audio feature representation learning, we have chosen the model from [Han et al., 2016] (we refer to it later as *Han et al. 2016*) as a baseline. This model has shown a superior performance on the task of predominant musical instrument classification comparing to the conventional feature-based machine learning methods. *Han et al. 2016* is a classical deep CNN architecture with 8 convolutional layers stacked in a sequence, followed by one fully connected layer. Max-pooling and dropout layers are placed after every second convolutional layer. All convolutional filters have a shape of 3×3, which is similar to popular CNNs used in computer vision.

We also experiment with a modified model from [Choi et al., 2016] (we refer to it later as *Choi et al. 2016*) with a final classification softmax layer instead of gated recurrent unit layers. This architecture follows the idea of stacking convolutional layers as well, but has a larger receptive field and exploits more advanced activation function and batch normalization [Ioffe and Szegedy, 2015], one of the effective regularization techniques.

In addition, we explore a recent Xception [Chollet, 2016] architecture for audio-based instrument recognition. We modify the input layer so that the receptive field is the same as in [Choi et al., 2016], and employ rectangular filters of size 48×3 at the first layer for better capturing the timbral characteristics of musical instruments. To reflect the changes of the input layer, we set the number of filters for separable convolutions equal to 768. The description of the network input is provided in Section 4.3.2.

4.2.3. Multimodal recognition

In this work, we investigate multimodal fusion strategy, so we use audio and video for both training and evaluation. Although the most direct approach for multimodal learning would be to train a model over concatenated audio-visual input (and thereby to fully integrate the modalities and learn a joint feature representation), earlier work in [Ngiam et al., 2011] demonstrates that there are almost no cross-modal connections in the resulting architecture. Moreover, such an approach would limit us to a small number of hidden layers, which is not desirable. Thereby, following

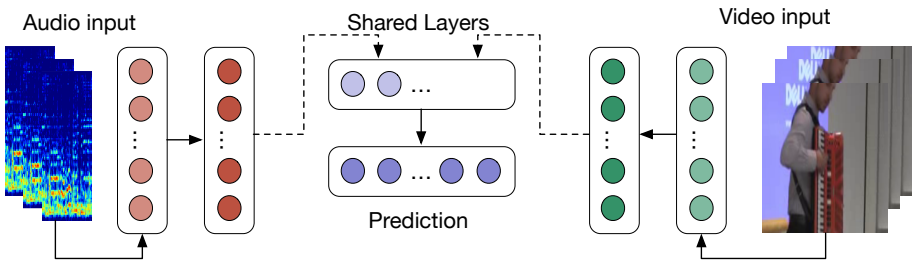


Figure 4.1: Schematic representation of our multimodal CNN architecture for musical instrument recognition.

the literature [Zhang et al., 2016, Xu et al., 2016, Ngiam et al., 2011], we individually train audio and video representation models and then exploit learned features from the last layers of the networks to train and evaluate the joint model as shown in Figure 4.1. Since the specific parameters for the audio and visual networks change for each experiment, we comment on the architecture of the late fusion model. The input layer of the model takes a concatenated feature vector of size $(k + 1, n)$, where k is the number of video frames (plus one vector of the audio features), and n corresponds to the penultimate layer size in the audio and visual networks. The model consists of two fully-connected layers (each layer contains 1024 neurons and ReLU activation function) preceding the batch normalization, and a softmax prediction layer.

4.2.4. Implementation details

Our approach is implemented with Keras [Chollet et al., 2015] and TensorFlow [Abadi et al., 2016]. We found that the best optimization strategy for video models consists of a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.0001 and a momentum of 0.9. We halved the learning rate every 5 epochs. We set the batch size to 64 and an early stopping criterion to 5 epochs for our visual-based experiments. For audio architectures, we use the Adam [Kingma and Ba, 2014]

optimizer with various batch sizes and 10 epochs for an early stopping criterion. All experiments are conducted on a single NVIDIA Titan X 12GB GPU. The code, extracted features, pre-trained models, and experimental results are available online[1].

4.3. Experiments and Results

4.3.1. Datasets

FCVID: The Fudan-Columbia Video Dataset (FCVID) [Jiang et al., 2018] contains videos, labels, several pre-computed descriptors and a category hierarchy. For our task, we consider a subcategory of the FCVID dataset namely *Musical performance with instruments* which contains 12 different classes including popular instruments, chamber music, rock band and orchestral performances. The subset contains 5154 videos with a total duration of almost 260 hours. All videos in the dataset were manually annotated by a team of 20 people (at least 3 annotations per video). Unfortunately, we could not find any information about human performance rate and agreement rate for this dataset.

YouTube-8M: The YouTube-8M Dataset [Abu-El-Haija et al., 2016] is a recently released large-scale video benchmark that consists of about 8 million YouTube videos corresponding with 4800 visual entities. The vocabulary for the dataset was created by humans, while the labels for individual videos were automatically obtained. To evaluate our task we check the dataset entities and select those that match musical instruments. We gather a dataset containing 235k videos of 46 classes. At the same time, we found that the resulting dataset contains a number of fine-grained categories, represented by only a few videos, while the top 3 categories form 75% of the dataset. To be able to compare our results to the FCVID dataset and to avoid problems related to dataset granularity and high imbalance, we reduce the number of categories to 13 and adjust the class distribution by undersampling the top 3 classes. The final dataset then

[1] http://github.com/Veleslavia/ICMR2017

Property	FCVID	YouTube-8M
Total number of categories	12	13 (46)
Total number of videos	5,154	60,862 (235,260)
Total video duration	259.84 hr	4,152.09 hr
Mean video duration	3.03 min	4.09 min
Videos per category (mean/std)	429 / 101	4,677 / 6,445

Table 4.1: Statistics of the musical instrument sub-datasets extracted from the FCVID [Jiang et al., 2018] and YouTube-8M [Abu-El-Haija et al., 2016] datasets. Numbers in parentheses correspond to sub-dataset statistics before undersampling.

contains more than 60k videos with a total duration of about 4k hours, making it the largest musical instrument recognition dataset at the time of the study. It is also worth mentioning that the original vocabulary for the dataset contains only visual entities and has been built with an emphasis on the ease of visual object recognition. The average human performance reported in [Abu-El-Haija et al., 2016] is 78.8% in precision and 14.5% in recall. For our experiments, we sample videos from both datasets to train, validation, and test splits with ratios of 70%, 15%, and 15% respectively. The details about the datasets can be found in Table 4.1.

4.3.2. Data Preprocessing

First, we separate audio and visual data and preprocess them individually. For audio, we convert the stereo input to mono by averaging the left and right channels and downsample it. We then compute two different one-channel log-mel-spectrogram representations following the models proposed in [Han et al., 2016, Choi et al., 2016]. The model [Han et al., 2016] (*Han et al. 2016*) has an input size of 128×43 (128 mel-frequency bins and 43 time frames) which corresponds to approximately 3 seconds of audio converted by STFT with a Hann window of size 1024 samples and a hop size of 512 samples. The model [Choi et al., 2016]

(*Choi et al. 2016*) has an input size of 96×1366 (96 mel-frequency bins and 1366 time frames, respectively) which corresponds to approximately 30 seconds of audio converted using an STFT with a window size of 512 samples and a hop size of 256 samples. For all the experiments we select 30 seconds from each video: for model *Han et al. 2016* we select 10 segments by 3 seconds, uniformly distributed in original audio (and average predictions over 10 segments); for the model *Choi et al. 2016* we investigate two segmentation strategies: central cropping (30 seconds from the middle of the audio, *CC*) and uniformly cropped segments (10 segments by 3 seconds, *UC*). To obtain a proper input for our visual model we take frames from videos with 1 fps frame rate, then resize every frame to size $256 \times 256 \times 3$, make a central crop with size $224 \times 224 \times 3$ and apply random horizontal flipping. In the experiments in which different numbers of frames are evaluated, we randomly select k frames for every video.

For both datasets, we only have one label per video, so we assign a video-level label to every selected frame.

4.3.3. Experimental setup

Metrics: For experimental evaluation we use three standard information retrieval metrics: accuracy (Hit@1, the success rate at top-1 prediction), top-3 accuracy (Hit@3, the success rate at top-3 predictions), and F1-measure (the harmonic mean of precision and recall).

Pre-training of Inception v3 model: Since it has been proven that pre-training helps to improve generalization ability and reduce training time [Erhan et al., 2010], we initialize the Inception v3 model with the model weights trained from ImageNet [Deng et al., 2009] and fine-tune the model on a subset of musical instrument images as described in [Slizovskaia et al., 2016].

4.3.4. Results

Visual-only classification results: Table 4.2 provides a summary of the visual-based musical recognition experiments. We observe that using

Dataset	FMs	PT	Steps	Time	Hit@1	Hit@3	F1
FCVID	20	No	32K	19h	42.30	64.53	43.16
FCVID	30	No	16K	11h	65.39	81.75	67.29
FCVID	30	Yes	16K	11h	68.77	84.26	70.33
FCVID	50	No	24K	22h	67.47	83.21	69.38
FCVID	50	Yes	21K	19h	**69.39**	**84.32**	**71.23**
FCVID	100	No	43K	98h	68.56	83.97	70.42
FCVID	100	Yes	36K	84h	67.76	83.50	69.16
YT-8M	10	No	58K	82h	61.15	78.45	52.19
YT-8M	20	Yes	57K	92h	**70.07**	**84.20**	**71.09**

Table 4.2: Comparison of clip-level performance for visual instrument classification model trained on different numbers of frames (FMs) with or without pre-training (PT) on ImageNet musical instruments. All rows use the same Inception v3 architecture.

pre-training and increasing the number of frames for training from 20 to 50 provides a significant improvement to the performance of the classifier ($F1 = 71.23$, FCVID) vs the baseline method ($F1 = 43.16$, FCVID). However, further increase of the number of frames to 100 does not yield higher performance. Our experiments also demonstrate noticeable success in using a pre-trained model compared to one with random initialization. The combination of two aspects (pretraining and increased number of frames) also demonstrates noticeable performance improvement on the YouTube-8M dataset (from $F1 = 61.15$ to $F1 = 70.07$). At the same time, increasing the number of frames results in a longer training process (from 22 to 84 hours for the FCVID dataset and from 82 to 92 hours for the YouTube-8M dataset), while use of the pre-trained model decreases training time (from 22 to 19 hours for the 50-frames model on the FCVID dataset).

Audio-only classification results: Results for audio-based musical instrument recognition are presented in Table 4.3. We observe that the highest accuracy is obtained by *(Choi et al. 2016)* for both datasets (79.81

Method	#Params	Dataset	Hit@1	Hit@3	F1
[Han et al., 2016]	1.5M	FCVID	64.13	76.82	53.64
[Choi et al., 2016] + CC	2.4M	FCVID	77.73	92.05	77.18
[Choi et al., 2016] + UC	2.4M	FCVID	**79.81**	**96.09**	78.71
[Chollet, 2016] + UC	9.6M	FCVID	78.69	94.44	**79.35**
[Han et al., 2016]	1.5M	YT-8M	59.37	70.87	56.50
[Choi et al., 2016] + UC	2.4M	YT-8M	**83.58**	94.23	**84.26**
[Chollet, 2016] + UC	9.6M	YT-8M	83.53	**94.69**	84.16

Table 4.3: Clip-level performance of different audio architectures and frame selection methods trained and evaluated on the FCVID (top) and YouTube-8M datasets (bottom).

for FCVID and 83.58 for YouTube-8M), although the results are similar to the ones obtained using the Xception architecture (78.69 for FCVID and 83.53 for YouTube-8M). For the FCVID dataset we experimented with central cropped (CC) and uniformly cropped (UC) segments for *(Choi et al. 2016)* architecture. Since we determined that the UC segments provide additional robustness, we use them throughout all the remaining experiments.

In addition, we observe that the classification results are significantly higher than the ones obtained using *(Han et al. 2016)* and that audio-based classification significantly outperforms video-based classification ($F1 = 79.35$ for audio vs $F1 = 71.23$ for video, FCVID; and $F1 = 84.26$ for audio vs $F1 = 71.09$ for video, YouTube-8M).

Multimodal classification results: Results for the combination of audio and video models are shown in Table 4.4. We observe that the highest accuracy of the audio-visual approach for FCVID is obtained using the Xception architecture ($Hit@1 = 88.28$), and the results are slightly lower for Choi ($Hit@1 = 86.97$). These results are noticeably better than the ones obtained by audio-only architectures for the FCVID dataset and significantly higher than the ones obtained using video-only architectures. For the YouTube-8M dataset, we observe that the classification performance

Method	Dataset	Hit@1	Hit@3	F1
[Chollet, 2016] / 50 frames	FCVID	**88.28**	**97.00**	**88.27**
[Choi et al., 2016] / 50 frames	FCVID	86.97	96.09	87.25
[Chollet, 2016] / 20 frames	YT-8M	82.64	91.37	78.95
[Choi et al., 2016] / 20 frames	YT-8M	**84.01**	**93.41**	**84.69**

Table 4.4: Overall performance of the proposed multimodal neural network for Choi [Choi et al., 2016] and Xception [Chollet, 2016] feature representations.

of our multimodal method is 13% higher than the visual-only method. Compared to the audio-only approach, our combined method demonstrates similar results.

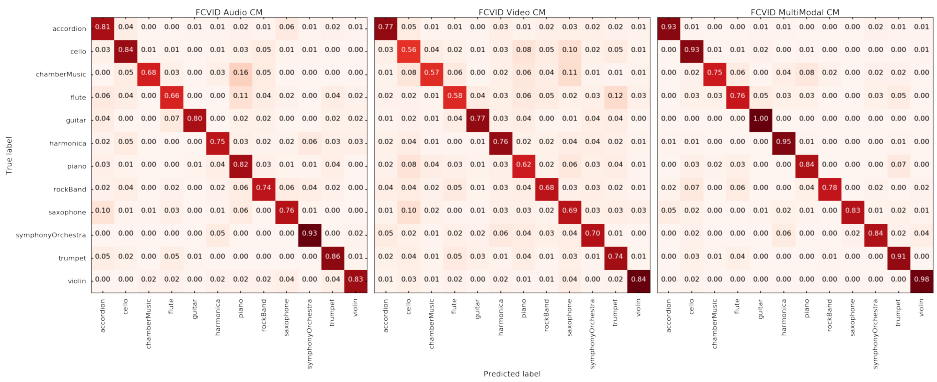


Figure 4.2: Comparison of confusion matrices for FCVID dataset. From left to right: audio-only recognition, video-only-recognition, multimodal recognition.

Confusion matrices: Figure 4.2 shows the confusion matrices obtained for the FCVID dataset using the three proposed approaches: audio-only, video-only, and multimodal. As shown the confusion is significantly reduced in the proposed multimodal approach vs the alternative methods, especially in the cases with harmonica (where the percentage of correct

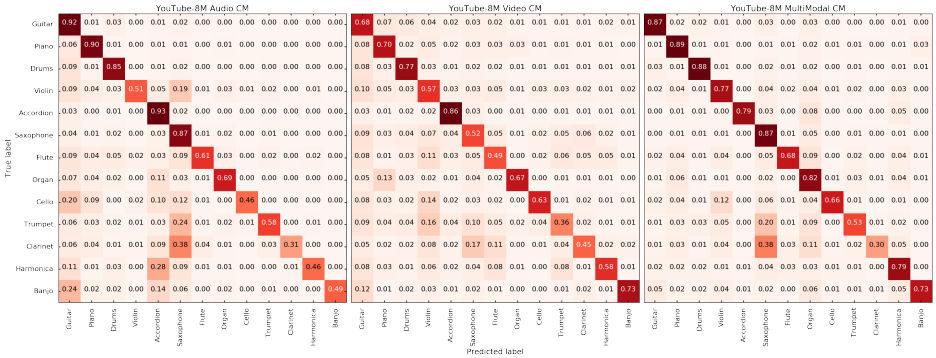


Figure 4.3: Comparison of confusion matrices for YouTube-8M dataset. From left to right: audio-only recognition, video-only-recognition, multimodal recognition.

predictions increases from 75-76% to 99%), and violin, accordion, guitar, and chamber music (with respectively, 11%, 10%, 10%, 10% of increase with respect to the audio alone and 10%, 14%, 13%, 21% of increase with respect to the video alone).

Figure 4.3 shows the confusion matrices obtained for the YouTube-8M dataset. We notice several significant differences with comparison to the FCVID dataset. The first is that the classification performance varies drastically between both categories and approaches. We believe that this is related to high imbalance of the categories and substantial diversity of videos. Also, for the visual data, we notice a number of annotation errors, so that the video sequence does not contain the target instrument while being annotated to the certain category.

To test this assumption we carry out a simple experiment on human recognition performance. Given a video from the YouTube-8M dataset, we ask non-expert humans to label it with one of the considered categories. When multiple instruments are present, we ask them to choose the predominant one. The total number of evaluated videos is 547, evaluated by 20 different people without specific musical training. We determine the human performance rate for our task to be 86.00 in precision, 85.00 in recall, and 85.00 in F1-measure. Those results are comparable to our multimodal

results ($F1 = 85.00$ vs $F1 = 84.69$). That allows us to conclude that the task (and dataset) is difficult to solve, even for humans.

The feedback from our participants also contains claims that the instruments are often not present in videos from the YouTube-8M dataset, or they might be only present in the audio stream. Despite the fact that it is much easier and faster to recognize the instrument by its shape, they say that if the instrument is not present on the frame, it is still possible to recognize it from the audio.

4.4. Case study on interpretability

The explainability of CNNs is a particularly challenging task in all areas of application, and it is notably under-researched in the music and audio domain. In this work, we approach explainability by exploiting the knowledge we have on hand-crafted audio features. We follow the problem of musical instrument recognition and experiment with one of the audio networks that demonstrated the best results in the previous study. Additionally, we expand the set of architectures with newest methods. We compute the similarity between a set of traditional audio features and representations learned by CNNs. We also propose a technique for measuring the similarity between activation maps and audio features, which are typically presented in the form of a matrix, such as a chromagram or spectrogram.

We observe that some neurons' activations correspond to well-known classical audio features. In particular, for shallow layers, we found similarities between activations and harmonic and percussive components of the spectrum. For deeper layers, we compare chromagrams with high-level activation maps as well as loudness and onset rate with deep-learned embeddings.

4.4.1. Motivation

In this section, we focus on *feature analysis* in the music domain using computer vision methods. Our goal is to find similar patterns between the features (activations and activation maps) learned by a network and hand-crafted audio features, which are well understood in the literature. For that purpose, we analyze features from a dataset of user-generated recordings of different musical instrument performances.

For feature attribution understanding, there are two major directions: (1) perturbation based algorithms, such as LIME [Ribeiro et al., 2016], Axiomatic Attribution [Sundararajan et al., 2017] or Saliency Analysis [Montavon et al., 2017], and (2) gradient-based algorithms such as Guided Backpropagation [Simonyan et al., 2013, Montavon et al., 2017], Class-Activation Mapping (CAM) [Zhou et al., 2016], and Network Dissection [Bau et al., 2017]. In the music domain, the SoundLIME [Mishra et al., 2017] algorithm has been adapted from the original LIME. However, in most cases, the above techniques can be limitedly applied to spectrograms because, unlike a typical image, two dimensions of a spectrogram represent different qualities: time and frequency.

Therefore, manual feature exploration remains popular. One could create a playlist which corresponds to a particular neuron and make a decision about this neuron's 'specialization' by listening to the playlist. This approach was proposed by [Dieleman, 2014] and it provides valuable insights. However, it is not scalable because it requires an expert to listen to the playlist and guess the rationale behind.

Also, we can take advantage of a number of well-established mid-level audio features that have been proposed and studied in the MIR literature [Schedl et al., 2014]. We know that CNNs in computer vision learn edges in the first layer and more complex concepts in subsequent layers. We hypothesize that audio-based CNNs can occasionally learn some of the hand-crafted features in a similar manner. We try to identify those features in pre-trained neural networks.

4.4.2. Methodology

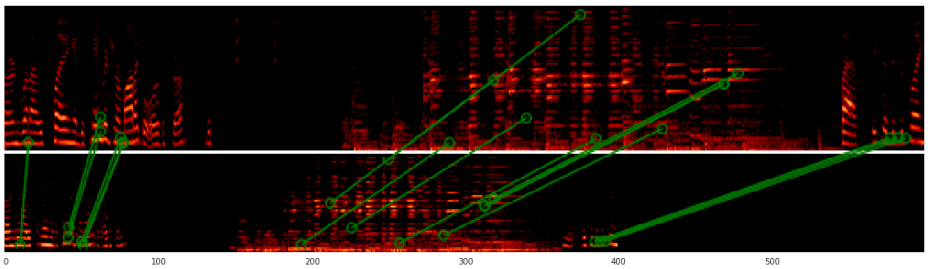
Hand-crafted audio features. We focus our study in a compact set of mid-level features related to different musical facets: onset rate, loudness and Harmonic Pitch Class Profile (HPCP) computed by `Essentia` [Bogdanov et al., 2013], and Harmonic/Percussive Sound Separation (HPSS) computed by `librosa` [McFee et al., 2015].

Network Architectures. We explore three state-of-the-art VGG-style architectures: CNN AudioTagger (CNN-AT) [Choi et al., 2016], that yielded the best performance in our previous study, VGGish [Hershey et al., 2016], and Musically Motivated CNN (MM-CNN) [Pons and Serra, 2017]. All three receive mel-spectrum as the input, consist of blocks of convolutional and max-pooling layers, and dense layers.

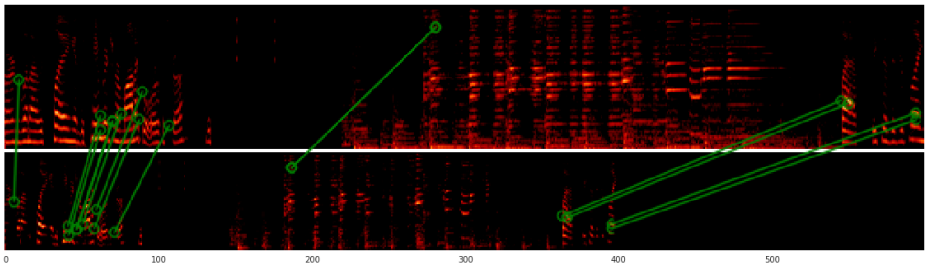
The differences between architectures and their initializations include filters' shape (squared filters in CNN-AT and VGGish, and rectangular filters in MM-CNN), activation function and pre-training settings. We trained CNN-AT and MM-CNN on a subset of the FCVID [Jiang et al., 2018] dataset. VGGish is initialized with weights provided by the authors. This network has been trained on a large-scale AudioSet dataset [Gemmeke et al., 2017] and potentially has stronger discriminative ability.

Similarity measures: individual activations. For high-level embeddings of a network, we consider each activation as an individual feature and compare them with onset rate and mean loudness. We consider two similarity metrics: (1) Pearson Correlation Coefficient and (2) Euclidean distance over the normalized vectors.

Similarity measures: activation maps. Activations of convolutional layers have a form of a matrix. They are slightly offset from the original input spectrum due to the padding, and proportionally scaled to the input because of max pooling. To some extent, we can think of them as pseudo-spectrograms or as filtered and aggregated spectrograms. In order to compare those activations with HPSS or HPCP, we need a method for fuzzy matrix comparison which is scale- and shift-invariant. We propose a



(a) Top: original log-mel-spectrum. Bottom: logharmonic component of HPSS, scaled. SIFT matches are connected.



(b) Top: original log-mel-spectrum. Bottom: logpercussive component of HPSS, scaled. SIFT matches are connected.

Figure 4.4: An example of SIFT matching for scaled harmonic (4.4a) and percussive (4.4b) parts of HPSS and shifted spectrum.

visual-inspired similarity metric based on Scale-Invariant Feature Transform (SIFT) [Lowe, 1999] descriptors. SIFT descriptors are among the most recognized features in computer vision and a reasonable choice for similarity measurement [Hua et al., 2012].

To compute similarity between a feature map and an activation map we compute SIFT descriptors and matches between descriptors. An example of matching is shown in Figure 4.4. Each match is characterized by the matched descriptor indexes and a matching distance.

4.4.3. Results

High-level embeddings vs. onset rate and loudness. We explored three high-level activation layers of the VGGish model: an embedding layer with 128 neurons and two fully-connected layers with 4096 neurons each. For the embedding layer, we found statistically significant correlations for both onset rate and loudness, with some examples of the corresponding features shown in Figure 4.5. In the first fully-connected layer we discovered that neuron #1964 has an outstanding correlation with loudness (with correlation coefficient $r = 0.76$). For CNN-AT we found that activation #259 corresponds to onset rate.

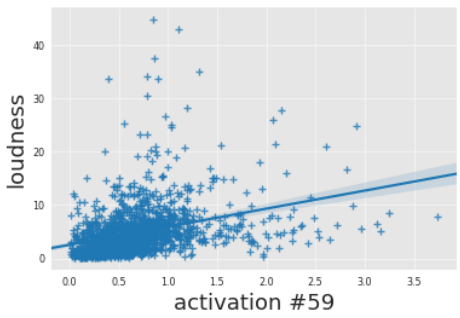
Low-level feature correspondences. We found a number of interesting activation maps which look similar to the HPSS decomposition in the first convolutional layer of the VGGish network. The histograms of similarity metrics with respect to activation maps are depicted in Figure 4.6. More examples can be also found in supplementary materials.[2]

The second convolutional layer of the VGGish network does not have a strong correspondence to the HPSS decomposition even though some linear combinations of activation maps could be similar. We assume that this behavior is caused by the fact that the filters of the second layer are more specialized.

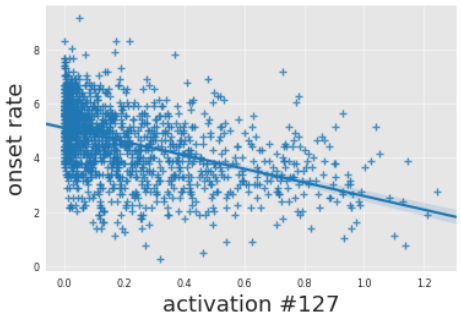
For the CNN-AT network we examine the second convolutional layer and we observe that similarity metric histograms for the HPSS decomposition are not consistent, which might be related to a higher false matching rate between decompositions and activation maps. Nevertheless, we would like to demonstrate an example of SIFT matches between HPSS and activation maps of the second layer of the CNN-AT network in Figure 4.7. Even for the activation maps of the second layer, SIFT descriptors remain reliable and capable of capturing keypoint features.

Finally, the first layers of the MM-CNN architecture represent strongly filtered spectrograms, so we presume that the tall rectangular filters of this architecture are similar to band-pass filters.

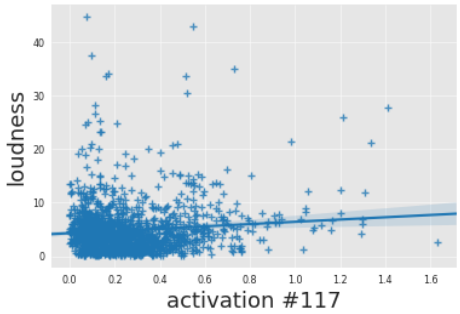
[2]Supplementary materials (high resolution figures, code and more examples) can be accessed at `https://goo.gl/jM3jZM`.



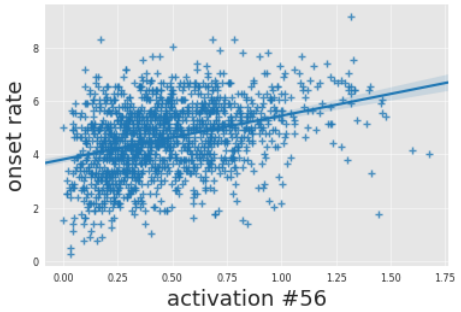
(a) Loudness/Activation #59.



(b) Onset rate/Activation #127.

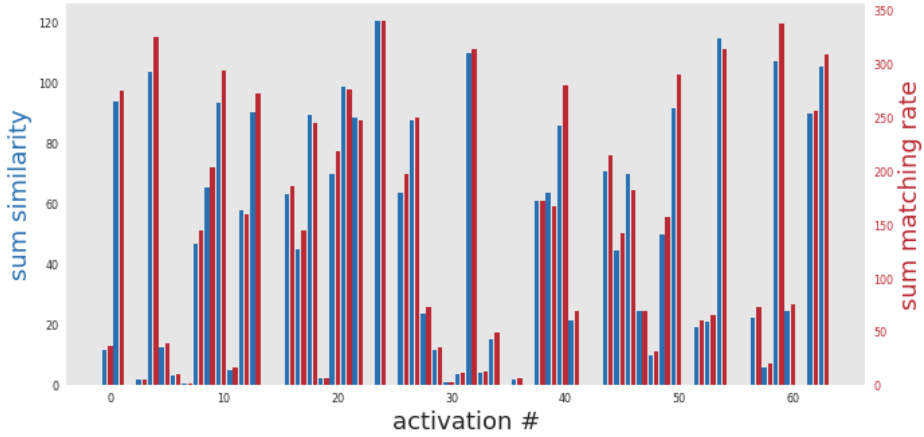


(c) Loudness/Activation #117.

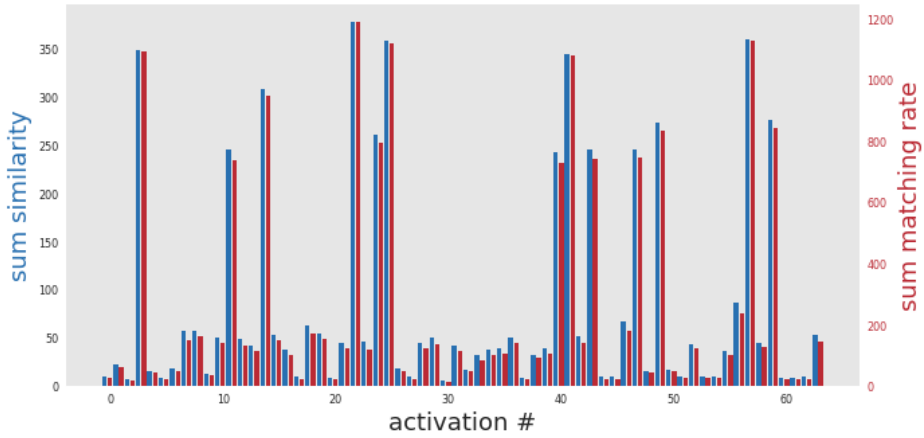


(d) Onset rate/Activation #56.

Figure 4.5: An example of correspondences between VGGish embeddings and mid-level audio features: 4.5a and 4.5b are correlation-based correspondences, 4.5c and 4.5d are $L2$-distance based correspondences.

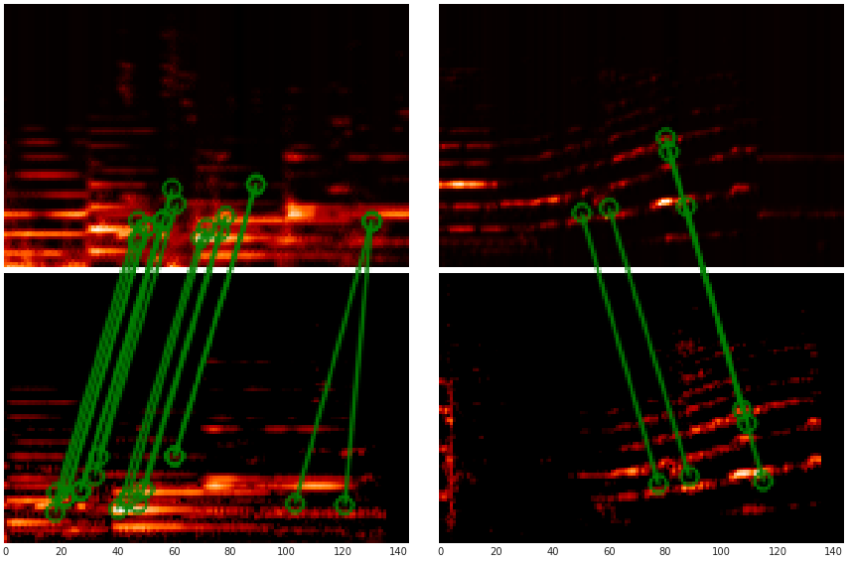


(a) Similarity metric histograms between harmonic component of HPSS and activation maps of the first convolutional layer.



(b) Similarity metric histograms between percussive component of HPSS and activation maps of the first convolutional layer.

Figure 4.6: Histograms of similarity metrics for activation maps of the first convolutional layer of VGGish network.



(a) Top: h-HPSS.
Bottom: activation map #25.

(b) Top: p-HPSS.
Bottom: activation map #57.

Figure 4.7: An example of correspondences between HPSS and activation maps of the second layer of CNN-AT network.

In addition, while investigating groups of activations and activation maps, we have also made the following observation:

- The networks have many redundant neurons and high cross-correlations between deep-learned features, so those features can be removed.
- (E,Re)LU activations have an additional advantage as they enforce filtering and produce sparse pseudo-spectrograms.

4.5. Conclusion

To summarize, we make several contributions in this chapter. First, we introduce a multimodal method for musical instrument recognition in user-generated videos. Second, we show the case when visual object recognition can be enhanced by adding audio information. Third, we evaluate several baseline convolutional neural network architectures for audio classification. Fourth, we investigate the influence of the number of frames used for image-based object recognition in video and the influence of using a pre-trained model. We evaluate our method on a heterogeneous large-scale dataset of user-generated videos so that it can be used with different datasets and scenarios.

Our results demonstrate that both modalities are important to obtain better performance. In addition, we show that the audio-only network and our multimodal approach perform remarkably close to the human performance rate for the musical instrument subset of the automatically annotated YouTube-8M dataset. This illustrates the fact that people may not only determine the video concept based on visual cues but also on the auditory ones. Moreover, the considered audio-only models clearly outperform the video-only models and the multimodal network performs better than those based on a single modality in one of the considered datasets, illustrating the advantage of multiple modalities.

Even if the models we investigate are complex and allow one to construct features in a very different way than traditional methods, the correspondences between hand-crafted features and activations provide insights

for better understanding of the internal representations of CNNs. We believe that the proposed methodology can be applied to identify important neurons in other tasks and architectures.

Chapter 5

Audio-visual music source separation

In this chapter, we focus on Single Channel Source Separation (SCSS). This task is usually solved considering only the audio modality, but in this work, we explore the effects of integrating two additional kinds of context data, namely instrument labels and their visual properties.

5.1. Introduction

The goal of music source separation is to extract the mixture of audio sources into their individually separated source tracks. Undoubtedly, this is a challenging problem to solve and many attempts have been made to estimate the source signals as closely as possible from the observation of the mixture signals. The most common cases may vary with respect to the target task (such as singing voice [Rafii et al., 2018, Jansson et al., 2017] or multi-instrument source separation [Miron et al., 2016, Chandna et al., 2017, Han and Raphael, 2010]), use of additional information (blind [Chandna et al., 2017, Jansson et al., 2017] or informed source separation [Miron et al., 2016, Carabias-Orti et al., 2013, Han and Raphael, 2010]), and the number of channels used for reconstruction (monau-

ral [Chandna et al., 2017] or multi-channel [Miron et al., 2016, Carabias-Orti et al., 2013] source separation).

There are many challenging aspects related to audio source separation. Most importantly, accurate separation with minimal distortion is desired. Supplementary information such as the number of sources present in the mix, musical notes in the form of MIDI or sheet music can be helpful but is not widely available in most cases. However, information such as the source instrument labels can be easily found from video recordings of musical performances readily available on the web. Therefore, it seems reasonable to learn to integrate the instrument label and corresponding visual information into the source separation pipeline. At the same time, many sophisticated score- and timbre-informed methods have been proposed in the literature already [Rafii et al., 2018]. We admire the idea of simplifying those frameworks, which became possible only recently with the advent of end-to-end deep neural networks.

We work with audio-visual recordings of musical ensembles with several families of instruments that can be commonly found in a symphonic orchestra such as strings, woodwinds and brass instruments; that is, mostly chamber music. Source separation with such a setup is known to be an incredibly challenging task and attempts to solve it have employed multi-channel score-informed methods [Miron et al., 2016] or timbre-informed methods [Carabias-Orti et al., 2013]. It is worth emphasizing that the above studies operate on multi-channel recordings and no clear ground truth was available. Besides, once a musical piece has been recorded, there is no simple way to unmix it.

The problem has several origins of complexity, to mention a few:

- The instruments within a family could be quite similar to one another;
- The number of sources in the mixture is unknown in advance;
- There is a high overlap in time and frequency between sources.

Even for instruments which have essentially different timbres, tone colors, and practical techniques, such as clarinet and viola, some musicians may

mimic a sound of one while playing another [Lee, 2004].

As for combining different modalities of information, for many years the key technical problem was the huge gap (both in dimensions and content) between representations of the modalities [Kidron et al., 2005]. One of the common approaches consisted of feature construction followed by dimensionality reduction [Kidron et al., 2005, Hershey and Movellan, 2000]. With the advent of deep learning techniques, the problem of the dimensionality mismatch can be considered to be solved, while a proper way of fusing different data representations remains an issue.

Another limitation of previous works is that the evaluation was done in somewhat unrealistic settings: typically, mixes of only two sources are considered and instruments from the same family are rarely present. In contrast to [Zhao et al., 2018], we added viola and double bass to the string instruments, and trombone to the brass instruments, increasing the overall variety of timbres. Besides performing the source separation, our method (in non-conditioned settings and while conditioned by visual information) associates the outputs with the different types of instruments, implying the presence of that instrument in the mix.

In this chapter, we study how to separate musical recordings of small ensembles (from duets to quintets) into individual audio tracks. We propose an extension of the Wave-U-Net [Stoller et al., 2018b], an end-to-end convolutional encoder-decoder model with skip connections, which supports a non-fixed number of sources and takes advantage of instrument labels in assisting source separation. This work also explores conditioning techniques at different levels of a primary U-Net source separation network.

We are not the first ones to propose Conditioned-U-Net for source separation or audio-visual source separation [Gao and Grauman, 2019, Zhao et al., 2018, Zhao et al., 2019, Korbar et al., 2018]. However, unlike prior approaches that were trained with an arbitrary choice of additional data integration, we conduct a thorough study identifying the optimal type of conditioning and comparing possible conditioning strategies with two types of context data: the presence or absence of instruments in the

mixture and the video stream data. Another notable contribution of our approach is that training is done by employing a curriculum learning strategy on mixtures of up to 7 sources, and evaluation is carried out on real-world mixtures from the URMP [Li et al., 2018a] dataset which has up to 4 different instruments per piece, often from the same family. The complexity of the task allows the present approach to be used as a baseline for future research. In order to facilitate that, the present study is reproducible as we provide pretrained models, code, data and all the training parameters. The supplementary materials and examples are available at `https://veleslavia.github.io/conditioned-u-net/`.

5.2. Conditioned Wave-U-Net

5.2.1. Multi-Source Extension

The challenge with the original Wave-U-Net model is that it can only support a predefined number of input sources (2 and 4 sources in the original settings), limiting its application to only the specific group of instruments on which it was trained. We aim to build a more flexible model that can support a dynamic number of input sources and, therefore, be more suitable for separating classical music recordings. In classical music, the number of instruments playing in an ensemble may vary a lot but the instruments themselves are often known in advance. Here we don't tackle the problem of separating different parts played by the same instrument (like violin1 vs violin2) but rather try to separate a sound track played by the same instrument (violin1+violin2 vs viola). Therefore, we can fix a maximum number of output sources to the number of all different instruments that are present in the dataset. This is still not a true dynamic model since the number of sources must be specified in advance. Thus, in order to have a more general model we fix the number of sources to a reasonable large number.

For the sources that are not available in the mix, the model is trained with silent audio as a substitute. Therefore, the model outputs all possible sources and is forced to associate each output with a certain instrument

and output silence for the sources that are not present in the mix. Note that at the training time we implicitly specify which source should be aligned with a particular instrument, but it is not needed at the inference time. We can instead use an energy threshold for extracting the sources of interest. We will refer to this model as *Exp-Wave-U-Net*.

5.2.2. Label Conditioning

In order to enhance the source separation results, we propose a conditioned label-informed Wave-U-Net model (*CExp-Wave-U-Net*). In particular, we use a binary vector whose size is the maximum number of sources considered. Each position of the vector is associated with a certain instrument: 1 indicates that the instrument is being played and 0 indicates either a non present instrument or a silent instrument (non-playing).

Conditioning is a term used to describe the process of fusing information of one medium into the context of another medium. In case of Wave-U-Net, there are three locations where the use of conditioning is appropriate and corresponds to different fusion strategies:

- for early fusion, the conditioning can be applied to the top layer of the encoder, before downsampling;
- for middle fusion, we can integrate label information at the bottleneck of the Wave-U-Net;
- for late fusion, we can aggregate labels with audio output of the last decoder layer (after upsampling).

Moreover, there is a possibility of using several conditioning mechanisms (as described in [Dumoulin et al., 2018]) such as

- concatenation-based conditioning;
- conditional biasing (additive bias);
- conditional scaling (multiplicative bias).

In this work, we experiment with multiplicative conditioning using instrument labels at the bottleneck of the Wave-U-Net model. Therefore, the overall idea is to cancel out the unwanted sources at the most compressed part of Wave-U-Net while emphasizing the sources of interest. Although the early fusion approach can be more abundant as it allows one to integrate more information from the very beginning, we use multiplicative middle fusion because it provides a reasonable trade-off between expressiveness of the network and memory and computational costs. At the same time, we leave additive bias and concatenation-based conditioning for further investigation.

5.2.3. Experimental setup

Dataset

As described earlier, the model takes the input in the form of a mix of the output sources where each source is either an instrumental track or a silent audio track for instruments not present in the mix. Instrument labels can be included optionally. We took advantage of the University of Rochester Musical Performance Dataset (URMP) [Li et al., 2018a] which consists of 44 pieces (11 duets, 12 trios, 14 quartets and 7 quintets) played by 13 different instruments (see Figure 5.1). We used 33 pieces for training and validation, and 11 pieces for testing.

Baseline

For the evaluation, we compare two proposed models with a Timbre-Informed NMF method from [Carabias-Orti et al., 2013]. In this method, the authors first learn a timbre model for each note of each instrument, then apply these trained templates as the basis functions in NMF factorization procedure. Note that the timbre templates are trained with RWC [Goto, 2004], a dataset which consists of recordings of individual notes for different instruments. Unlike our approach, Timbre-Informed NMF requires specifying the timbre models for each piece at the inference time. We used learned timbre models for all instruments except saxophone.

5.2.4. Implementation details

Our implementation is available online[1] and is based on the original Wave-U-Net code[2]. We improved both input and training pipelines compared to the original work. The input pipeline is implemented as a TensorFlow Dataset and now supports parallel distributed reading. The training pipeline is re-implemented via a high-level TensorFlow Estimator API and supports both local and distributed training. Our implementation also supports a half-precision floating-point format, which allows us to increase both training speed and batch size without loss of quality.

We train the model on a single Google Cloud TPU instance for 200k steps which takes approximately 23 hours. The best results are achieved using an Adam optimizer with an initial learning rate of 1e-4. The aforementioned modifications together with the use of TPU allowed us to speed up the training process by 24.8 times (35.3 times for the half-precision case) compared to a single GPU training.

5.2.5. Results

We perform a quantitative evaluation of the model performance using standard metrics for blind source separation: *Source to Distortion Ratio* (SDR), *Source to Inference Ratio* (SIR), and *Source to Artifacts Ratio* (SIR) [Vincent et al., 2006].

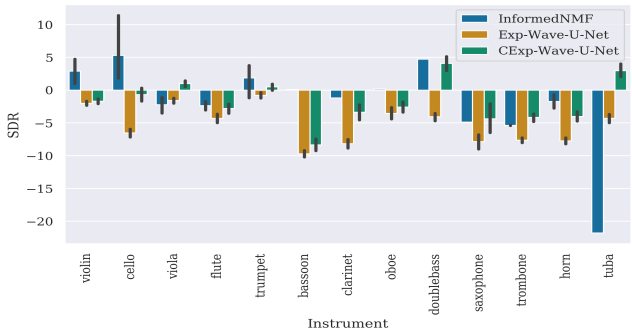
[1] `https://github.com/Veleslavia/vimss`
[2] `https://github.com/f90/Wave-U-Net`

Method	**SDR**	**SIR**	**SAR**
InformedNMF	**-0.16**	1.42	9.31
Exp-Wave-U-Net	-4.12	-3.06	**12.18**
CExp-Wave-U-Net	-1.37	**2.16**	6.36

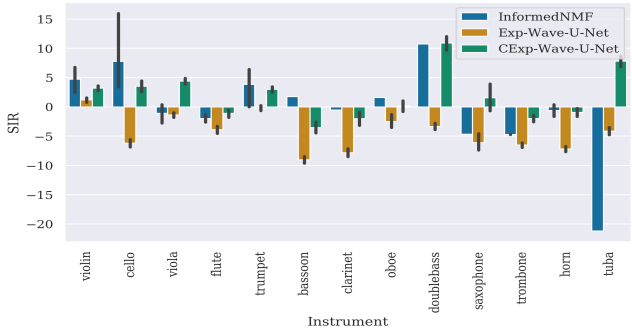
Table 5.1: URMP [Li et al., 2018a] dataset: SDR, SIR and SAR for different methods averaged over the testing set. Best values are shown in bold. Exp-Wave-U-Net stands for an extension of Wave-U-Net with multiple output sources, CExp-Wave-U-Net stands for a version of Exp-Wave-U-Net conditioned by labels of the instruments.

Model	**nSources**	**SDR**	**SIR**	**SAR**
InformedNMF	2	3.08	4.98	10.55
	3	0.07	1.69	9.01
	4	-3.84	-2.62	8.65
Exp-Wave-U-Net	2	-0.42	1.75	10.98
	3	-3.85	-2.74	11.97
	4	-5.90	-5.33	12.87
CExp-Wave-U-Net	2	-0.16	4.62	7.48
	3	-0.68	2.88	5.91
	4	-2.56	0.44	6.35

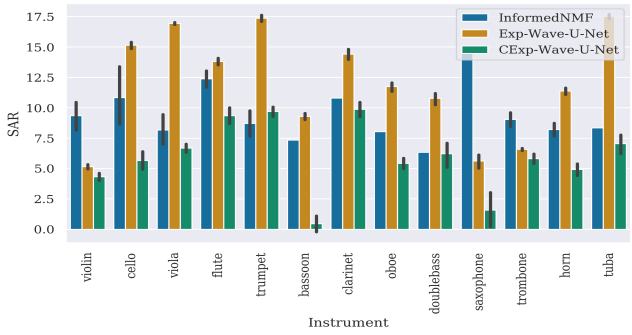
Table 5.2: SDR, SIR and SAR for different methods averaged with respect to the number of sources in the mix.



(a) SDR (dB)



(b) SIR (dB)



(c) SAR (dB)

Figure 5.1: Results in terms of SDR, SIR, and SAR for each instrument in the testing set of the URMP [Li et al., 2018a] dataset.

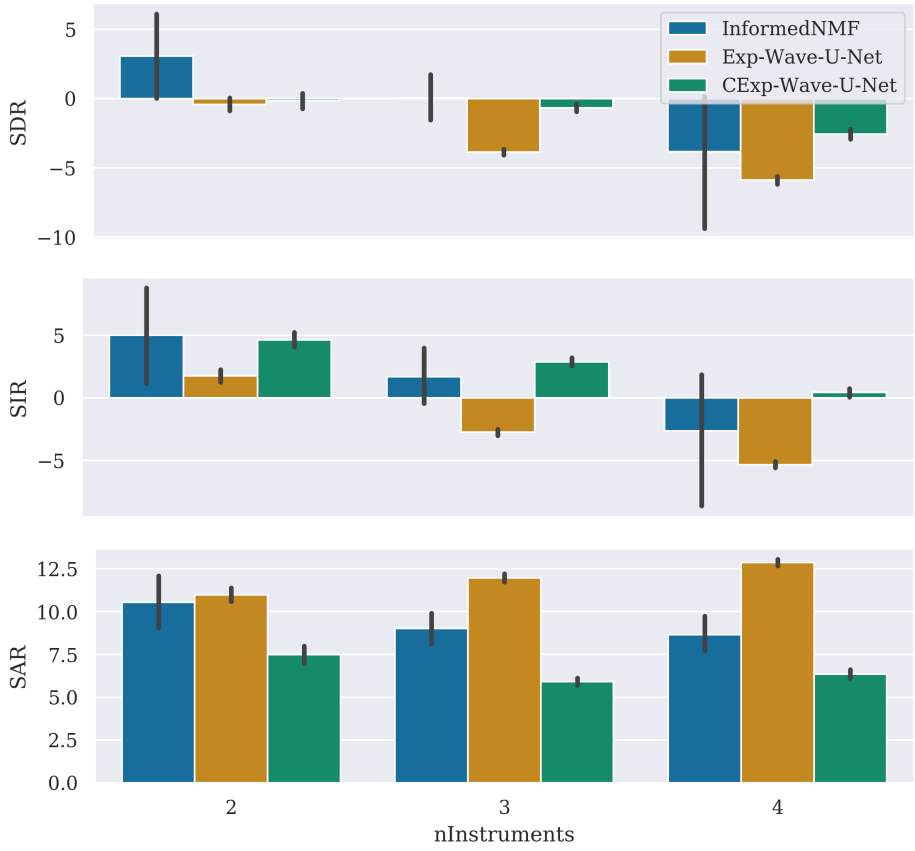


Figure 5.2: Results in terms of SDR, SIR, and SAR averaged and reported by the number of instruments in the testing set of the URMP [Li et al., 2018a] dataset.

Table 5.1 shows the average values of the metrics over all pieces and instruments in the dataset. We can see that there is no single winner, but each method seems to be better with respect to one of the metrics. For example, InformedNMF baseline outperforms both deep models in terms of SDR while it is inferior to Exp-Wave-U-Net in terms of SIR and to CExp-Wave-U-Net in terms of SIR. Note that we can't directly compare our results with Wave-U-Net because it would require training from 3 to 11

different models, while for Exp-Wave-U-Net we just train a single model for all instruments and numbers of sources.

Next, we analyze the separation performance in depth for each instrument. Figure 5.1 summarizes the results for each model and metric. We can see that the baseline approach (InformedNMF) performs reasonably well in terms of SDR and SIR for all instruments except for trombone and tuba. Exp-Wave-U-Net performs worse in SDR and SIR for all instruments but consistently outperforms the baseline and CExp-Wave-U-Net in SAR apart from violin, trombone and saxophone. CExp-Wave-U-Net performs as well as the other two in SDR and SIR (and achieves the best results for tuba, doublebass, saxophone and viola) but consistently worse in SAR.

At last, we report the separation results averaged with respect to the number of sources in the input mix in Table 5.2. It is interesting to note that the performance of all methods decreases as the number of sources increases. However, it is more interesting that the performance of CExp-Wave-U-Net does not drop as much as in the case of InformedNMF and Exp-Wave-U-Net. In absolute values (see Table 5.2), SDR for CExp-Wave-U-Net decreases from -0.16 dB to -2.56 dB while for the model without conditioning those values are -0.42 dB to -5.90 dB, and from 3.08 dB to -3.84 dB for the NMF baseline. A similar pattern persists for SIR. From these results, we could anticipate that the conditioned model is more suitable for multi-instrument source separation.

We would like to mention that despite their widespread use, the standard metrics are unable to estimate how well the model can discard unwanted sources (they are undefined if the ground truth is silence). Nonetheless, we would like to provide samples of separated sources which should be discarded[3]. We notice that both conditioned and unconditioned versions of Exp-Wave-U-Net systematically output quieter sources for the absent instruments than InformedNFM, initialized by all possible timbre templates.

Some qualitative results for original and expanded[4] Wave-U-Net can be also found online.

[3] https://goo.gl/e18F41

[4] https://youtu.be/mGfhgLt1Ds4, https://youtu.be/mVqIMXoSDqE

5.3. Conditioned U-Net

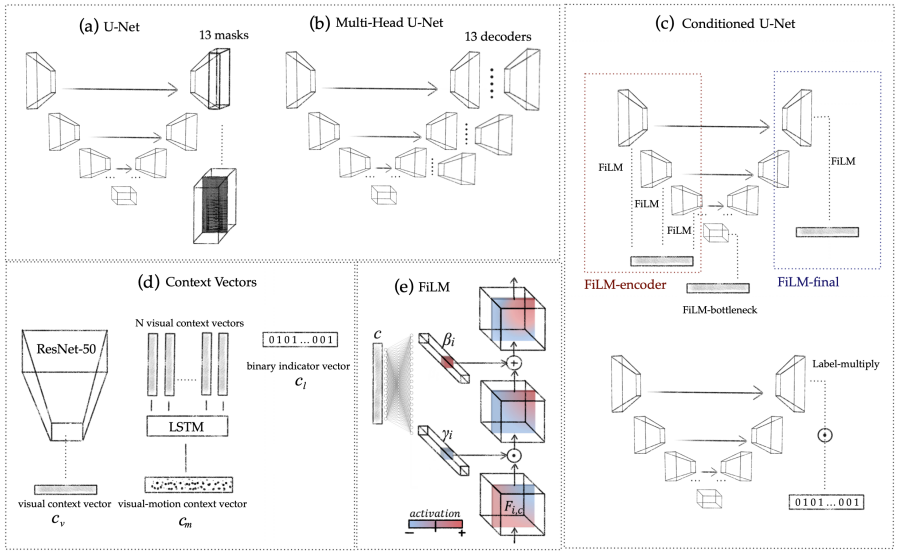


Figure 5.3: Summary of architectures, methods and context information used in the experiments. There are two baselines for the source separation architecture: (a) U-Net which outputs 13 masks at the last upconvolutional layer, (b) Multi-Head U-Net with one shared encoder and 13 specialized decoders which output one mask each. (c) There are several choices for U-Net conditioning: three types of FiLM conditioning and multiplicative conditioning of the output masks. (d) We use three possible types of context information for conditioning: (1) static visual context vector (which is a feature vector obtained at the last convolutional layer of ImageNet-pretrained ResNet-50), (2) visual-motion context vector obtained as the output of an LSTM trained on N visual context vectors from consecutive video frames, and (3) binary indicator vector which encodes which instruments are present in the mix. (e) We outline the FiLM method in subfigure (e) as in [Dumoulin et al., 2018].

In this work, we study the effect of integrating two types of context information, namely labels and visual context, at different locations of the network, while keeping the architecture fixed and simple.

We use a mix-and-separate approach for training, such that every mixture is generated on the fly and, therefore, unique. To create a mixture, we take the following steps: (1) we sample an arbitrary subset of instruments; (2) we subsequently pick a random segment from one of the audios of that instrument category; and (3) we sum time-domain values of the segments and clip them to the $[-1, 1]$ range. Given a magnitude spectrogram of the mixture, our network learns to predict K real-valued masks $\hat{M}_i$, one mask per potential instrument present in the mixture (we use $K = 13$ different instruments in our experiments, see Figure 5.3(a) U-Net). Each output mask is associated with a certain kind of instrument, and their order is fixed to reduce the source permutation effect.

Additionally, we employ a curriculum learning strategy for training, gradually increasing the number of sources in the mixture. Consequently, the predictions of the network are *sparse*, meaning that many sources should be silent (and many masks are all zeros) as only a subset of instruments is present in the mix.

5.3.1. U-Net and Multi-Head U-Net baselines

As the focus of this work is on studying the effect of different types of conditioning, we leave for future research the analysis of different source separation networks and adopt two simple U-Net versions as the baseline architectures, given that U-Net has been extensively used in source separation and has demonstrated good performance [Jansson et al., 2017, Doire and Okubadejo, 2019, Zhao et al., 2018, Zhao et al., 2019, Gao and Grauman, 2019].

U-Net [Ronneberger et al., 2015] is an encoder-decoder architecture with *skip connections* such that activations of every i^{th} layer of the encoder are concatenated with activations of $N - i^{th}$ layer of the decoder, which can be considered as a light form of conditioning by itself. Following [Zhao et al., 2018, Zhao et al., 2019], we have chosen one of the

architectures they propose and set the number of layers to $N = 6$. We employ two variants of the architecture, namely: (a) a baseline U-Net architecture as pictured in Figure 5.3(a) which outputs 13 masks after the last upconvolutional layer, and (b) Multi-Head U-Net (MHU-Net) [Doire and Okubadejo, 2019] as pictured in Figure 5.3(b) which has a single shared encoder and 13 decoders, where each dedicated decoder yields a mask for its corresponding instrument.

Audio is resampled at 11025 Hz before preprocessing. We use the Hann window and STFT is computed for every segment of approx. 6 seconds (65535 audio samples) with a window size of 1022 (this value is taken for compatibility with [Zhao et al., 2018]) and a hop size of 256, which results in a matrix of 512×256 STFT bins. Those parameters are taken from [Zhao et al., 2018] and some of them have been proven to work well, e.g. the window size of about 23ms goes well with the best performance window size of 25ms in [Kavalerov et al., 2019] for universal sound separation. Next, we study a few preprocessing strategies over the STFT representation, including linear and log-sampled frequency scales for STFT, as well as log-scale and dB-scale with normalization for STFT magnitude values as discussed in Section 5.3.5.

The choice of the loss functions is dependent on the type of the mask. For binary masks at each time-frequency bin i we compute binary cross entropy (BCE) loss:

$$\mathcal{L}^{bce} = -\sum_{i=1}^{|E|} (w\, y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)), \tag{5.1}$$

where y_i and $\hat{y}_i$ represent ground truth and predicted mask values, $|E|$ is the total number of points in the mask, and w is a positive weight which is used to compensate for the class imbalance in the mask values.

For ratio masks we employ smooth L_1 loss which is defined as:

$$\ell_i^{smooth} = \begin{cases} 0.5(y_i - \hat{y}_i)^2, & \text{if } |y_i - \hat{y}_i| < 1 \\ |y_i - \hat{y}_i| - 0.5, & \text{otherwise} \end{cases} \tag{5.2}$$

$$\mathcal{L}^{smooth} = \sum_{i=1}^{|E|} \ell_i^{smooth}, \tag{5.3}$$

where $|y_i - \hat{y}_i|$ refers to the distance between ground truth and predicted mask values.

Finally, the total loss is the sum of the 13 individual BCE losses (5.1) in the case of binary masks, or the smooth L_1 losses (5.3) in the case of ratio masks.

5.3.2. Conditioned U-Net

In this section we describe the conditioning strategies and the types of context data which we use in our Conditioned U-Net architecture (Figure 5.3(c, d)).

Weak label conditioning

We study *weak conditioning* for source separation which means that instrument labels are available at the level of individual recordings. They indicate the presence or absence of each instrument in the mix, which is encoded in a binary indicator vector $\mathbf{c}_l \in \{0, 1\}^K$ where K is the total number of instrument classes considered.

Then, we use $\mathbf{c}_l$ as a conditioning context vector and compare three types of FiLM conditioning: introduced (1) at the bottleneck, (2) at all encoder layers, and (3) at the final decoder layer as indicated in Figure 5.3(c). More formally, for each layer j we have activations or embeddings: $\boldsymbol{a}^{(j)}$, and the conditioning is as follows:

$$(\gamma_j, \beta_j) = f_j(\boldsymbol{c}_l) \tag{5.4}$$

$$\hat{\boldsymbol{a}}^{(j)} = \gamma_j \boldsymbol{a}^{(j)} + \beta_j \tag{5.5}$$

Furthermore, we explore simple multiplicative conditioning with the binary indicator vector:

$$\hat{M}_i = \boldsymbol{c}_l[i]\tilde{M}_i, \tag{5.6}$$

where $\mathbf{c}_l[i]$ is the i^{th} component of the context vector and $\tilde{M}_i$ is the i^{th} preliminary mask as predicted by (MH)U-Net.

Visual conditioning

In the case of visually-informed source separation, we consider both static characteristics and motion-aware conditioning. Nonetheless, we would like to note that learning temporal information from videos is a challenging task which is still under research. Therefore, visually-informed methods mostly use a single frame for conditioning [Zhao et al., 2018, Gao and Grauman, 2019, Gao et al., 2018], with some exception of dense trajectories [Li et al., 2017, Parekh et al., 2017], and deep-learned dense trajectories [Zhao et al., 2019].

Like [Parekh et al., 2017, Gao and Grauman, 2019], we assume that rough spatial location of each source is given (e.g. it can be obtained by a segmentation or human detection algorithm). Keeping this assumption in mind, we use uncropped frames from individual videos for training and evaluation. In a real life scenario (e.g. for testing) we use a bounding box around every player.

For visual context conditioning, we take a single video frame corresponding to the beginning of the audio source sample. We use a pretrained ResNet-50 [He et al., 2016a] to extract a visual feature vector of size 2048 for every present source, then concatenate them, obtaining a visual context vector $\mathbf{c}_v$ of size $K' \times 2048$ where K' is the maximum number of sources in the mixture. The context vector for the unavailable sources is set to all zeros. As for the case of weak label conditioning, we compare three alternatives for the FiLM conditioning (see Figure5.3(c)).

For visual-motion conditioning, we first extract visual feature vectors with the pretrained ResNet-50 at a fixed frame-rate within a selected segment. We then pass the obtained sequence of vectors through a small uni-directional LSTM network as in [Gao et al., 2019], with the aim to capture motion characteristics while keeping visual information. We take the last LSTM hidden state of size 1024 for every sequence and concatenate the obtained features resulting in a motion context vector $\mathbf{c}_m$ of size $K' \times 1024$. Due to the large computational cost, and the results of the ablation study (Section 5.3.5), we only report this approach with FiLM conditioning at the bottleneck of the audio U-Net.

5.3.3. Experimental setup

In what follows, we thoroughly evaluate the proposed method on various setups. In particular, we compare the different conditioned networks with respect to several performance metrics.

Dataset

In our experiments we use two multimodal datasets of musical performances: the Solos dataset [Montesinos et al., 2020], that we recently introduced, for training and evaluation, and the URMP dataset [Li et al., 2018a] for testing.

The original URMP dataset consists of 44 arrangements (of which 11 are duets, 12 are trios, 14 are quartets, and 7 are quintets). Each source track was recorded separately with an external coordination, and the final mixes were assembled afterwards. The instrumentation is a typical one for chamber and orchestral music, and includes such families of instruments as strings (violin, viola, cello and double bass), woodwinds (flute, oboe, clarinet, bassoon, saxophone), and brass (trumpet, horn, trombone, tuba). The dataset is constructed to reflect the complexity of the musical world where the same instrument within a section can appear more than once.

As we only tackle the problem of separating sources of different instruments, we mix source tracks of the same instrument within the same piece and consider the resulting mix as a single source. For example, for a string quartet (which consists of 2 violins, a viola and a cello), we join two source tracks of violin, which results in a corresponding "trio" where two violins are considered as a single source. Also, we remove four pieces (02_Sonata_vn_vn, 04_Allegro_fl_fl, 05_Entertainer_tpt_tpt, 06_Entertainer_sax_sax) from the dataset as they are duets of the same instrument and thus there would be nothing to separate. After this preprocessing, we were left with 12 duets, 20 trios and 8 quartets in the final set.

The Solos dataset consists of 755 YouTube videos of solo musical performances of the same 13 instruments categories as the URMP dataset. It has a total duration of about 66 hours. A major part of the dataset

consists of audition performances which ensures, together with manual and semi-automatic checking, a good quality of audio and video. The dataset is positioned as a tool to facilitate the training via the *mix-and-separate* strategy while being complementary to the URMP dataset. The latter allows proper evaluation on real-world mixtures.

Metrics

Several studies indicate that widely-adopted source separation metrics such as signal to distortion ratio (SDR), signal to inference ratio (SIR), and signal to artifacts ratio (SIR) [Vincent et al., 2006] do not always agree with human perception [Le Roux et al., 2019, Kilgour et al., 2019, Zhao et al., 2018, Gao and Grauman, 2019]. Recently, *scale-invariant* and *scale-dependent* SDR (SI-SDR, SD-SDR) metrics have been proposed [Le Roux et al., 2019] in order to tackle this issue.

Unlike previous works [Gao and Grauman, 2019, Xu et al., 2019, Zhao et al., 2018, Zhao et al., 2019], our method produces *sparse* outputs since many predicted sources are expected to be silent. However, all the above metrics are ill-defined for silent sources and targets. To address this issue, we also compute cumulative *predicted energy at silence* (PES) and *energy at predicted silence* (EPS) as proposed in [Schulze-Forster et al., 2019]. For SI-SDR and SD-SDR larger values indicate better performance, while for PES and EPS smaller values indicate better performance. For numerical stability of the log function, in our implementation we add a small constant $\epsilon = 10^{-9}$ which results in the lower boundary of the metrics being -80 dB.

5.3.4. Implementation details

Our U-Net is composed of 6 blocks in the encoder and 6 blocks in the decoder. Each encoder block consists of a convolutional layer followed by batch normalization with an optional conditioning layer (for FiLM-encoder conditioning), and ReLU non-linearity. A decoder block consists of a bilinear upsampling layer, a convolutional layer, batch normalization,

ReLU non-linearity, and a dropout layer.

The network is trained for 500k iterations with a batch size of 16, Adam optimizer, and an initial learning rate of 10^{-5} which is halved after 25k iterations with no improvement on the validation set.

We opted for a curriculum learning strategy. It consists of starting the training with only mixtures of 2 sources, and gradually increasing the maximum number of sources to 7. The increment is carried out if validation loss does not decrease for 10k iterations.

For training and evaluation we utilize the *mix-and-separate* procedure by creating artificial mixtures from individual videos of Solos. Every training sample has an arbitrary number of sources with an upper bound of the maximum number of sources at the current curriculum stage. For testing, we use real mixtures from the URMP dataset.

5.3.5. Baseline ablation study

In preparation for conditioned source separation analysis and to define the optimal hyperparameters of the baseline U-Net architecture as described in Section 5.3.1, we conduct a series of ablation experiments. We examine the following set of hyperparameters: (1) linear vs. log frequency scale for the STFT representation, (2) binary vs. ratio masks estimation, (3) data augmentation with normally-distributed noise, (4) log vs. dB-normalized scale for the STFT values, (5) the use of curriculum learning, and (6) the effectiveness of Multi-Head U-Net vs. vanilla U-Net.

Our final baseline configuration is a model that takes dB-normalized and log-frequency scaled STFT as input. It has a single decoder and predicts binary masks. We have opted out of augmenting the input with normally-distributed noise and have used curriculum learning.

Method	ID	SI-SDR ↑	SD-SDR ↑	PES ↓	SDR ↑	SIR ↑	SAR ↑
IBM	U	11.4 ± 5.9	11.2 ± 6.4	n/a	10.31 ± 4.42	17.47 ± 5.54	11.84 ± 4.30
input mix	L	-3.7 ± 5.7	-3.7 ± 5.7	18.2 ± 4.2	-3.48 ± 4.82	-3.20 ± 4.95	18.10 ± 11.21
log-scale STFT	1	$\mathbf{-12.5 \pm 21.0}$	$\mathbf{-18.1 \pm 28.2}$	$\boldsymbol{-47.9 \pm 30.2}$	$\mathbf{-4.35 \pm 8.48}$	$\mathbf{-0.18 \pm 7.34}$	$\mathbf{5.03 \pm 8.55}$
linear-scale STFT	2	-15.9 ± 20.4	-24.1 ± 28.8	-33.7 ± 24.5	-5.86 ± 9.05	-0.68 ± 8.33	3.16 ± 8.65
binary masks	3	-10.7 ± 19.9	-14.9 ± 24.7	$\mathbf{-41.6 \pm 33.3}$	-3.09 ± 8.64	1.12 ± 8.68	4.89 ± 6.78
ratio masks	4	$\mathbf{-2.3 \pm 7.3}$	$\mathbf{-10.8 \pm 12.7}$	-11.9 ± 8.4	$\mathbf{0.52 \pm 6.60}$	$\mathbf{3.54 \pm 8.48}$	$\mathbf{8.06 \pm 3.52}$
w/o noise	5	$\mathbf{-10.6 \pm 20.1}$	$\mathbf{-14.9 \pm 25.6}$	$\mathbf{-42.1 \pm 32.5}$	$\mathbf{-3.12 \pm 8.84}$	$\mathbf{1.28 \pm 8.41}$	4.99 ± 7.60
w/ noise	6	-17.2 ± 24.5	-22.2 ± 28.4	-32.5 ± 34.7	-6.34 ± 11.56	-1.99 ± 10.42	4.99 ± 8.88
log-value STFT	7	-19.1 ± 25.5	-26.5 ± 31.8	$\mathbf{-47.3 \pm 31.4}$	-6.57 ± 11.00	0.95 ± 9.43	0.98 ± 10.30
dB-normalized STFT	3	$\mathbf{-10.7 \pm 19.9}$	$\mathbf{-14.9 \pm 24.7}$	-41.6 ± 33.3	$\mathbf{-3.09 \pm 8.64}$	$\mathbf{1.12 \pm 8.68}$	$\mathbf{4.89 \pm 6.78}$
no curriculum	8	-17.2 ± 24.8	-21.9 ± 28.2	-33.6 ± 34.5	-6.37 ± 12.10	-1.93 ± 10.95	4.57 ± 8.54
curriculum	5	$\boldsymbol{-10.6 \pm 20.1}$	$\mathbf{-14.9 \pm 25.6}$	$\mathbf{-42.1 \pm 32.5}$	$\mathbf{-3.12 \pm 8.84}$	$\boldsymbol{1.28 \pm 8.41}$	$\mathbf{4.99 \pm 7.60}$
U-Net	9	-12.3 ± 19.3	-17.9 ± 26.5	$\mathbf{-44.0 \pm 27.8}$	-4.19 ± 8.06	-0.36 ± 7.45	$\boldsymbol{5.32 \pm 7.87}$
MHU-Net	3	$\mathbf{-10.7 \pm 19.9}$	$\boldsymbol{-14.9 \pm 24.7}$	-41.6 ± 33.3	$\boldsymbol{-3.09 \pm 8.64}$	$\mathbf{1.12 \pm 8.68}$	4.89 ± 6.78

Table 5.3: Ablation studies results for the URMP dataset. The first two rows indicate two possible baselines: ideal binary masks (IBM, U stands for the upper bound), and the usage of the input mixture as a predicted source (input mix, L stands for the lower bound). Note that the SAR metric is ambiguous and is reported for consistency only. Within each pair of the ablation experiments we highlight the best results in **bold**. The most important results for binary mask estimations are *italicized*.

5.3.6. Results

Ablation studies

We report the metrics obtained by our baseline models in the ablation study in Table 5.3, and the full list of hyperparameters is given in Appendix A.1. The experiments can be matched by the experiment ID. We also provide the metrics for two baselines, the upper bound separation quality (U) with ideal binary masks (IBM), and the mixture metrics (L) which reproduce the input mixture at every possible output source.

Although the results for a multi-decoder architecture (with experiment ids: 3-8) have a higher separation quality, they double the required computational cost. Therefore, we have opted out of training the MHU-Net architecture. Table 5.3 shows that ratio masks (Exp. 4), when compared to binary masks (Exp. 3), give higher (SI-/SD-)SDR but perform much worse in terms of PES. In particular, the increment in SI-SDR is 8.4dB, in SDR it is 3.6dB, while the drop in PES is 29.7dB. In practice, we noticed that, while training with ratio masks, the to-be-silent output sources eventually happen to be an original mixture with a lowered volume. Therefore, in all following experiments we predict binary masks. Further study on combining the binary and soft masks as in [Grais et al., 2016] may help solve this issue. We also observed that augmenting input data with normally-distributed noise does not improve separation performance and thus other more advanced techniques are needed.

Figure 5.4 shows the performance measured by SI-SDR, SD-SDR and PES in Exp. 4 for each instrument in the URMP dataset. The results emphasize the fluctuations between the instruments. We can see that for the case of bassoon, tuba, horn, and viola the mean SI-SDR is about -6.5dB which is quite poor. In contrast, for some string instruments such as cello, double bass and violin, the SI-SDR is higher (with the maximum mean value of 1.8dB for cello). There is also the special case of the saxophone whose performance metrics are good on average, but whose standard deviation is the highest among all the instruments.

Overall, the ablation studies indicate that different aspects of the separation quality measured by the standard metrics can be enhanced by applying

different learning strategies. Notably, the curriculum learning technique helps to improve overall separation quality for all the metrics measured. The next significant improvement of 3.4dB in SI-SDR is obtained by changing the frequency scale of the STFT representation, followed by the multi-decoder U-Net architecture (1.6dB improvement in SI-SDR) and dB-normalized STFT values (8.4dB improvement in SI-SDR), which improve (SI-/SD-)SDR but worsen PES (-3.4dB and -5.7dB decrease, respectively).

Conditioning on labels

We further study weak label conditioning of the single-decoder U-Net model. We provide results for linear-frequency scale and log-frequency scale STFT inputs and four conditioning schemes as described in Section 5.3.2. The summary of weak label conditioned source separation results is shown in Table 5.4.

We observe that the best performance in terms of (SI-/SD-)SDR is obtained with multiplicative conditioning of the output masks, but that this also leads to high PES, even worse than in the case of ratio masks in the ablation study. Explicitly, the label-multiply conditioning method achieves -2.8dB and -3.0dB of SI-SDR for the linear-frequency scale (Exp. 12) and log-frequency scale (Exp. 16), respectively. However, it yields 7.4dB and 8.9dB for this scales in PES.

Within FiLM conditioning experiments, we note that the FiLM-bottleneck conditioning undoubtedly outperforms the FiLM-encoder and FiLM-final types of conditioning by a mean margin of 1.8dB in SI-SDR and 0.7dB in SDR. We found that FiLM-encoder and FiLM-final conditioning may lead to overfitting and worsen the results w.r.t. non-conditioned U-Net, while FiLM-bottleneck conditioning coherently improves the results in all tested settings.

Although the log-scale STFT input outperforms the linear-scale STFT input for the cases of no conditioning or FiLM-encoder conditioning, there is no significant difference for FiLM-bottleneck and label-multiply conditioning, and there is a drop in the performance for FiLM-final conditioning.

Figure 5.5 shows scatter plots of input SI-SDR versus improvement in

SI-SDR for each segment in the URMP dataset. Subfigure (a) demonstrates results for the model of Exp. 12 with multiplicative label conditioning from Table 5.4. Subfigure (b) displays the upper bound results obtained with ideal binary masks. The figure indicates the potential upper bound separation performance that can be achieved with this dataset.

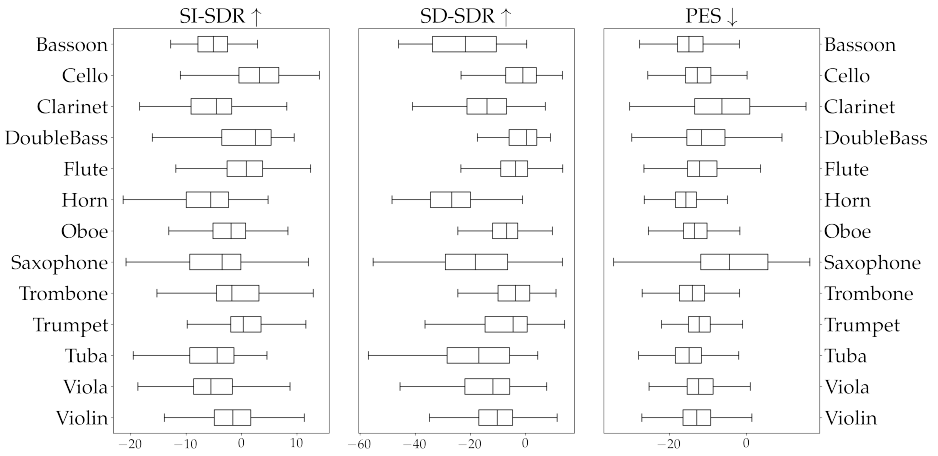
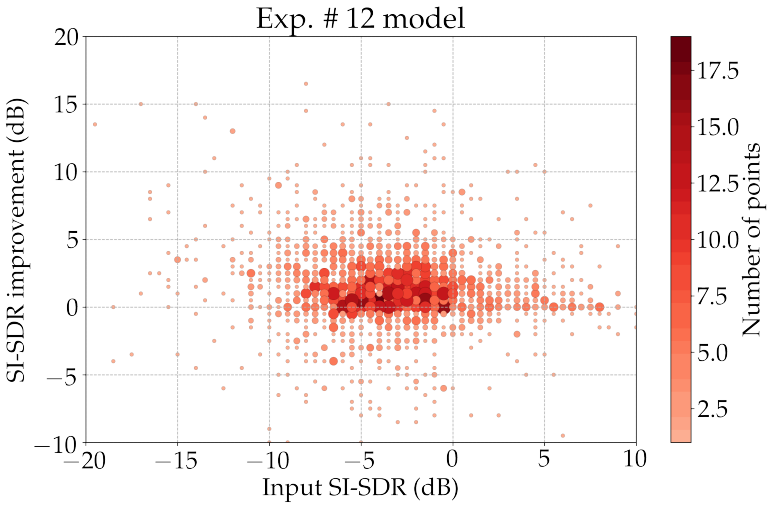
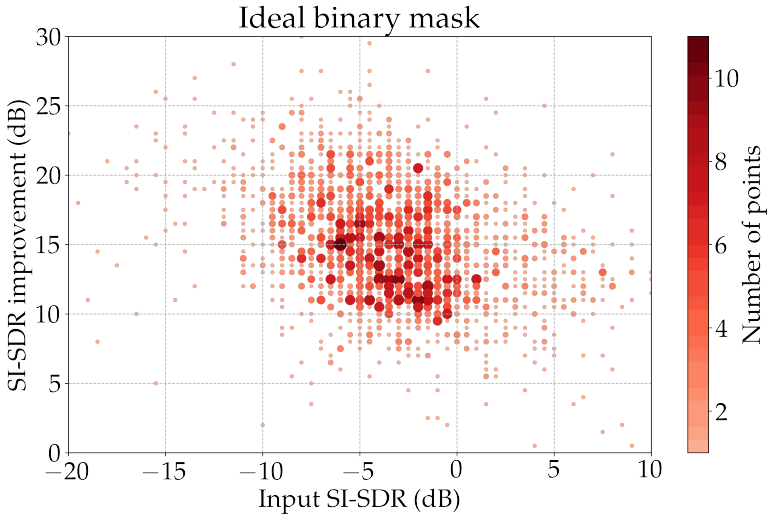


Figure 5.4: Exp. 4 per-instrument boxplots for the URMP dataset. Note that the x axis scale limits vary from metric to metric. The principal reason for the difference between SI-SDR and SD-SDR is that SD-SDR accounts for volume changes.



(a) Input SI-SDR vs. Exp. 12 SI-SDR improvement



(b) Input SI-SDR vs. ideal binary masks SI-SDR improvement

Figure 5.5: Input SI-SDR vs. SI-SDR improvement scatter plots. (a) Results of label-multiply conditioned U-Net with linear-scale STFT and b) oracle binary masking. The darkness and the size of points is proportional to the number of overlapping points.

Method	ID	SI-SDR ↑	SD-SDR ↑	PES ↓	SDR ↑	SIR ↑	SAR ↑
linear-scale STFT							
w/o conditioning	2	-15.9 ± 20.4	-24.1 ± 28.8	-33.7 ± 24.5	-5.86 ± 9.05	-0.68 ± 8.33	3.16 ± 8.65
FiLM-encoder	10	-14.5 ± 19.7	-21.9 ± 27.7	$\mathbf{-34.5 \pm 25.7}$	-5.19 ± 8.43	$\mathbf{-0.25 \pm 7.86}$	3.56 ± 8.27
FiLM-bottleneck	9	$\mathbf{-12.3 \pm 19.3}$	$\mathbf{-17.9 \pm 26.5}$	$\mathbf{-44.0 \pm 27.8}$	$\mathbf{-4.19 \pm 8.06}$	-0.36 ± 7.45	5.32 ± 7.87
FiLM-final	11	-13.3 ± 21.2	-18.3 ± 27.4	-31.4 ± 34.1	-5.12 ± 8.84	-0.95 ± 7.91	$\mathbf{6.50 \pm 10.76}$
Label-multiply	12	$\mathbf{-2.8 \pm 8.6}$	$\mathbf{-3.2 \pm 9.1}$	7.4 ± 8.9	$\mathbf{-1.46 \pm 5.47}$	$\mathbf{0.00 \pm 5.97}$	$\mathbf{9.34 \pm 4.33}$
log-scale STFT							
w/o conditioning	1	-12.5 ± 21.0	-18.1 ± 28.2	$\mathbf{-47.9 \pm 30.2}$	-4.35 ± 8.48	-0.18 ± 7.34	5.03 ± 8.55
FiLM-encoder	13	-14.0 ± 20.6	-21.3 ± 28.8	-37.2 ± 28.4	-4.86 ± 8.67	$\mathbf{-0.08 \pm 8.06}$	3.98 ± 8.24
FiLM-bottleneck	14	$\mathbf{-12.4 \pm 19.9}$	$\mathbf{-17.7 \pm 26.5}$	$\mathbf{-45.6 \pm 30.1}$	$\mathbf{-4.26 \pm 8.24}$	-0.57 ± 7.47	5.35 ± 7.73
FiLM-final	15	-14.5 ± 22.6	-20.2 ± 28.6	-35.9 ± 36.3	-4.89 ± 9.24	-0.79 ± 8.59	$\mathbf{6.13 \pm 9.64}$
Label-multiply	16	$\mathbf{-3.0 \pm 10.3}$	$\mathbf{-3.3 \pm 10.7}$	8.9 ± 7.4	$\mathbf{-1.48 \pm 5.85}$	$\mathbf{-0.14 \pm 6.40}$	$\mathbf{9.90 \pm 4.54}$

Table 5.4: Conditioned U-Net with Labels (URMP metrics). Two sets of experiments are conducted, with linear-frequency scale STFT as input, and log-frequency scale STFT as input. The most relevant results are highlighted in **bold**.

Method	# frames	SI-SDR ↑	SD-SDR ↑	PES ↓	SDR ↑	SIR ↑	SAR ↑
w/o conditioning	0	-12.5 ± 21.0	-18.1 ± 28.2	$\mathbf{-47.9 \pm 30.2}$	-4.35 ± 8.48	-0.18 ± 7.34	5.03 ± 8.55
FiLM-encoder	1	-14.5 ± 20.4	-22.4 ± 28.8	-37.4 ± 28.2	-4.98 ± 8.58	-0.19 ± 8.02	3.96 ± 8.34
FiLM-bottleneck	1	-12.0 ± 20.2	-17.1 ± 26.6	-44.9 ± 29.5	-4.20 ± 8.54	-0.38 ± 7.54	5.41 ± 8.19
FiLM-bottleneck-ft	1	$\mathbf{-10.5 \pm 19.6}$	$\mathbf{-15.2 \pm 26.6}$	-46.8 ± 30.0	$\mathbf{-3.65 \pm 8.33}$	$\mathbf{0.20 \pm 7.14}$	5.65 ± 8.33
FiLM-bottleneck	15	-12.2 ± 19.0	-17.8 ± 26.1	-37.6 ± 29.8	-4.94 ± 8.60	-0.41 ± 8.24	4.48 ± 8.30
FiLM-bottleneck	50	-14.6 ± 21.1	-21.5 ± 28.7	-38.8 ± 30.1	-5.25 ± 8.74	-0.81 ± 7.91	4.39 ± 8.48
FiLM-final	1	-13.3 ± 21.3	-18.6 ± 27.7	-39.0 ± 35.0	-4.96 ± 9.40	-0.81 ± 8.19	$\mathbf{5.79 \pm 9.66}$
SoP-unet7 [Zhao et al., 2018]	3	-18.69 ± 8.97	-21.09 ± 9.36	n/a	-3.76 ± 4.00	-1.45 ± 4.68	7.56 ± 3.13
SoP-unet7-ft [Zhao et al., 2018]	3	-17.48 ± 8.50	-20.25 ± 9.29	n/a	$\mathbf{-2.57 \pm 4.99}$	$\mathbf{0.47 \pm 6.43}$	6.89 ± 2.48
SoP-unet5-Solos	3	$\mathbf{-16.97 \pm 8.61}$	$\mathbf{-18.69 \pm 8.86}$	n/a	-2.92 ± 4.64	-1.67 ± 5.34	$\mathbf{11.07 \pm 6.87}$

Table 5.5: Results for Visually Conditioned U-Net experiments with different types of conditioning and different numbers of frames used (evaluated on the URMP dataset). We also show the results of the Sound-of-Pixels model [Zhao et al., 2018] for (1) the released pre-trained architecture (SoP-unet7), (2) the original architecture finetuned on the Solos dataset (SoP-unet7-ft), (3) and trained from scratch on the Solos dataset (SoP-unet5-Solos). The most important results are highlighted in **bold**.

Conditioning on visual information

We compare the visually conditioned U-Net with its corresponding non-conditioned and label conditioned baselines.

Table 5.5 shows the performance of the single-frame visually conditioned U-Net given the same FiLM locations as in the label conditioning case. It also indicates the results of conditioning by visual-motion context vector learned from 15 and 50 frames per segment (with the frame rate set to 2.5 fps and 8.3 fps respectively). Lastly, we report the results for the Sound-of-Pixels (SoP) [Zhao et al., 2018] method. SoP-unet7 stands for the original method trained on the Music dataset published in [Zhao et al., 2018]. We used the officially provided weights and evaluated the model on the URMP dataset. SoP-unet7-ft refers to the version that was fine-tuned on the Solos dataset. SoP-unet5-Solos accounts for a model with 5 blocks in U-Net and which is trained from scratch. In all SoP networks both visual and audio networks are trained simultaneously, while in our conditioning experiments the visual network is frozen in all experiments except for FiLM-BOTTLENECK-FT.

The results show that the visually conditioned U-Net, analogously to the label conditioned U-Net, outperforms its non-conditioned baseline only for the case of FiLM-bottleneck conditioning, whereas FiLM-encoder and FiLM-final methods result in a performance drop up to 2dB in SI-SDR. FiLM-bottleneck single-frame conditioning slightly outperforms its hypothetical label conditioned upper bound, Exp. 14 from Table 5.4, and FiLM-bottleneck-ft outperforms the baseline by a margin of 0.8dB. Additionally, in the experiments where both audio and visual subnetworks are trained, FiLM-bottleneck-ft architecture outperforms SoP-unet7 trained on Music in both SDR and SIR. However, it performs worse in SDR when compared to SoP-unet5-Solos trained on Solos while still performing better in SIR. Clearly, SoP-unet7-ft trained on Music and fine-tuned on Solos performs the best, in terms of SDR, SIR and SIR, out of all visually-conditioned networks which indicate that the performance can still be improved by employing datasets which are bigger and of better quality. The experiments with the visual-motion context vector indicate the need

for better motion representation as the results show the performance drop w.r.t. single-frame visual conditioning.

Unsuccessful attempts

We would like to report several strategies that did not improve source separation performance in our experiments.

In one of the experiments, we used L_2 loss while directly predicting spectrogram values instead of using the masking-based approach. However, the network failed to converge. We hypothesize that this behaviour accounts for the higher complexity of the spectrograms and the sparsity of the outputs.

We also unsuccessfully attempted to employ multi-task learning in order to further regularize the embedding space. In these experiments we jointly optimized classification and separation losses trying to predict which instruments are present in the mixture using the bottleneck U-Net features as an input for a small classifier consisting of a single fully-connected layer. While generally converging, the classification and separation performance were lower than the results of stand-alone models.

5.3.7. Discussion

In our experiments we observe that the use of external information generally improves separation performance. FiLM-encoder conditioning leads to overfitting and only improves SIR. FiLM-final conditioning improves SAR but not the rest of the metrics. FiLM-bottleneck and Label-multiply conditioning improve over all their corresponding baselines in all the metrics except PES, and the same behavior is observed while predicting ratio masks and using Multi-Head U-Net.

From the results we observe that U-Net conditioned on the visual context vector improves over the unconditioned versions in terms of (SI-/SD-)SDR but performs worse in terms of PES and SIR. A possible explanation for this observation may have to do with the capacity of the network to learn playing/non-playing activity from the visual information.

However, it may still have difficulties separating musical instruments from the same family (such as viola and violin) which may result in more interferences and mispredictions when both are present in the mixture, which is a common case for the URMP dataset.

By inspecting the results obtained by the Sound-of-Pixels method, we highlight the importance of taking the source separation problem into the real-world scenario, as the method was previously tested in mix-and-separate settings and the reported results had an average SDR of 8dB. Our results demonstrate the demand for the testing on the real mixtures rather than using the mix-and-separate approach. Notably, even 5-blocks Sound-of-Pixels trained on Solos performs better than 7-blocks Sound-of-Pixels trained on Music. Joined fine-tuning of the original Sound-of-Pixels model allows one to improve the quality of source separation for 1.2dB in SDR which also indicates the need to enlarge the datasets and enhance their quality.

Following [Zhao et al., 2018], we confirm that directly integrating visual information from multiple frames in a form of visual features worsens separation results. Although from the literature we know that source separation can benefit from integrating motion information [Zhao et al., 2019, Parekh et al., 2017, Li et al., 2017], we would like to note that all aforementioned methods use complex pre-processing in order to extract reliable motion features, which brings attention to the problem of closing the gap between motion and audio representations.

Another fact that should be noted is that all sources of information should be correctly combined, preserving synchrony between them. While for single-frame visual and weak label information it is not so important, for temporal data such as motion, pitch, and musical scores it may become a crucial aspect for successful conditioning. Therefore, a different baseline source separation architecture, such as an RNN-based network, may improve the current results due to its sequential nature which better preserves time-domain information.

Taking into consideration the above-mentioned observation, we can note that the U-Net architecture may be a limitation of our study, and the results may be different for other baseline architectures.

Given that the best results in terms of different metrics are achieved using different setups (e.g. binary and ratio masks), we would like to emphasize that a further enhancement can be obtained by having the best of both worlds as has been proposed in [Grais et al., 2016]. Finally, we would like to note the opportunity to surpass the current performance by employing additional constraints for the loss functions as in [Gao and Grauman, 2019] and [Wisdom et al., 2019], or weighting the loss values of the masks with the magnitude values of the mixture [Zhao et al., 2018, Gao and Grauman, 2019] as it may help to avoid treating every time frequency bin equally and focus attention on the areas where most of the energy is concentrated.

5.4. Conclusion

We tackle a problem of Single Channel Source Separation for multi-instrument polyphonic music conditioned on external data. In this work we have shown that the use of extra information such as (1) binary vectors indicating the presence or absence of musical instruments in the mix and (2) visual feature vectors extracted from corresponding video frames improve separation performance.

In this chapter we have proposed and explored two extensions of the Wave-U-Net architecture in the context of source separation of ensemble recordings with unknown numbers of input sources. We have shown that both Exp-Wave-U-Net and CExp-Wave-U-Net perform fairly well in comparison to the InformedNMF model despite being trained on just 33 audio mixes. We observed that CExp-Wave-U-Net outperforms the baseline approach when the number of input sources is bigger than 2. Moreover, we observed that Exp-Wave-U-Net produces a quieter output for the non-present instruments.

We also show that different types of conditioning have different effects w.r.t. the performance metrics. We have conducted a thorough study of FiLM-conditioning introduced at three possible locations of the primary source separation U-Net model. We have demonstrated that the best results

can be obtained with FiLM-bottleneck conditioning and with multiplicative label conditioning on the predicted masks.

The results shown in the present work indicate that the real-case scenario such as chamber quartets source separation is challenging and there is still a significant performance gap of about 13dB between the state-of-the-art separation methods and ideal binary masks.

Potential improvements could include modifying the U-Net architecture, combining binary and soft masks to obtain a good balance between SDR and PES. We observe that visual guidance seems to be a prominent direction of research because in this case not only do we not need to have manually annotated instrument labels but we can also obtain additional information of the playing and non-playing state of each instrument by analyzing the corresponding video stream. This can be especially useful for resolving ambiguity and interference between two instruments of the same kind. Therefore, another possibility for future research could be integrating an advanced motion analysis network and employing audio-motion synchrony for conditioning the network, as well as conditioning on musical scores.

Chapter 6

Conclusions

The primary focus of this thesis is enhancing MIR techniques with the use of extra information and modern deep learning methods. This work lies at the intersection of MIR, AV Signal Processing, and Multimodal Deep Learning. From the MIR perspective, we address two well-known tasks: musical instrument classification and sound source separation. From the AV Signal Processing perspective, we explore modern ways to understand audio and visual data and construct abstractions from it. Finally, from Multimodal Deep Learning, we examine different ways of aggregating available data modalities, defining an optimal data fusion approach for each task.

6.1. Overview and contributions

The present thesis deals with an automatic analysis of musical performance videos using mainly deep learning methods and audio and visual components of musical performances. The principal research objective of our work was designing machine learning methods for musical performance video analysis that enable the use of the audio and visual data. In what follows, we summarize the content of the thesis and contributions of the preceding chapters.

- In Chapter 1, we discuss the motivation behind the studies in audio-visual music information retrieval, the complexity of the field, the multifaceted nature of music data, the potential of AV MIR, and related challenges. In particular, we outline two practical MIR tasks that can benefit from the use of multi-modal data: musical instrument classification and sound source separation. We also define the practical domains that can benefit from audio-visual analysis of musical data. Also, we put emphasis on data-related challenges and processing-related challenges, which are common in AV MIR.

- In Chapter 2, we set the stage for our research providing the necessary background information. We give a formal definition of the classification and source separation tasks, providing information about the deep learning framework and methods that we used in the subsequent studies. We also discuss the non-trivial topics of data representation and data fusion, which both form part of the challenges mentioned in Chapter 1.

- Chapter 3 presents a comprehensive review of the historical perspectives on audio-visual signal processing, common tasks and methods. That is followed by a more specific review of AV MIR tasks, with a special focus on classification and source separation methods, representation learning, and techniques for multimodal data aggregation.

- Chapter 4 presents a study on audio-visual instrument classification and explores a concatenation-based late-fusion technique for multimodal analysis. The results highlight the effectiveness of the proposed method compared with audio-only and video-only classification. Additionally, we have conducted an analysis on the interpretability of features learned by the proposed audio architecture, showing relations between hand-crafted features and learned representations.

- In Chapter 5, we introduce a series of novel source separation methods conditioned on instrument labels and visual data in the context of classical chamber music. We demonstrated that the inte-

gration of extra modalities improves separation performance. We conducted a detailed study on different data fusion methods, including concatenation-based, multiplicative and FiLM conditioning.

In the following, we describe our conclusions with respect to the research questions formulated in Section 1.2.

RQ1. Which MIR tasks can benefit from audio-visual analysis? What is a potential improvement that we can obtain?

In Section 3.3 we provide a detailed overview of existing and prominent directions in AV MIR. We feature various audio-visual musical performance analysis tasks in Section 3.3.1, focusing on their practical applications. In our experiments, we have proven the benefit of using audio and visual information in the task of instrument classification and source separation on which we have focused. Thanks to modern audio-visual deep learning techniques, development of a number of multi-task learning methods, addressing several problems at the same time, and making use of complementarity of audio and visual information is possible. In addition, we outline a prospective research area of self-supervised audio-visual MIR, with an example problem being automatic audio-visual onset detection via joint learning of co-occurrence of audio and motion events.

RQ2. Can we extract meaningful and effective audio-visual features useful for MIR?

The learned representations for the music instrument classification task have been studied. Their effectiveness is demonstrated by superior performance in the tasks of interest when compared to single modality approaches. Joint audio-visual features can be obtained after combining learned representations from different domains. It is important to maintain cross-domain synchrony for time-sensitive tasks, and, as shown in the literature [Owens and Efros, 2018, Korbar et al., 2018, Parekh et al., 2019b, Zhao et al., 2019], we can use that synchrony to learn discriminative features from audio and visual components. In Chapter 4 we conduct an interpretability study, showing the correspondences between learned and hand-crafted audio features, which indirectly verifies the meaningfulness of the learned representation.

RQ3. How can audio-visual strategies and features help us to better understand underlying multimodal relationships? How much impact does visual information have on musical performance analysis?

In Chapter 3 we detail methods that use analysis of music data for the better structural decomposition of video data, and vice versa, showing the relationships between these modalities. In our experiments in Chapter 4 and Chapter 5, we demonstrate that audio-visual methods usually perform better than unimodal (usually, audio-based) techniques. Thus, for classification, we achieved an 8.47% improvement in Hit@1 compared to the audio-only approach. Similarly, for source separation, we improved the performance of the algorithm by 2dB in SI-SDR by integrating extra visual cues. Additionally, the presence of extra modalities helps us to better understand underlying confusions of a black-box machine learning algorithm, when decisions from individual modalities can be tracked separately, as is the case for musical instrument classification.

RQ4. What are optimal strategies for multimodal fusion? How sensitive are different machine learning methods to incorrect data fusion techniques and the missing modalities problem?

In the present thesis, we mostly focus on late and hybrid fusion methods. We propose a methodology for data aggregation for the tasks of AV classification of musical instruments and AV source separation. We experiment with different fusion and conditioning techniques in the context of conditioned end-to-end and STFT-based source separation, using extra modalities of instrument labels and video frames. Our results demonstrate superiority of the late and hybrid fusion methods over the early fusion techniques.

6.2. Limitations and future research directions

The conducted studies and the overall progress in multimodal data processing have opened a number of prominent research directions with potential improvements both in the performance of studied methods and

the field of AV MIR in general.

One of the limitations of this research study is that we adopt pretrained models for extracting visual appearance features, making use of availability of pretrained models from the computer vision field. The approach makes joint training stable and non-degenerate, meaning that both streams conveyed useful information. However, if we would like to extend our work to learning and integrating, for example, motion patterns, we would probably need to train a motion network from scratch jointly with an audio network. A recent study in joint multimodal learning [Wang et al., 2020] reports that such joint training can result in inferior performance because the useful information flows through the predominant modality. To overcome that issue, they propose a Gradient-Blending method based on an overfitting-to-generalization ratio (OGR). It serves as an extra regularization in the multimodal training process and has shown to result in improved stability and better performance in the audio-visual classification task.

Another possible way to improve the quality of the proposed methods is in the use of unsupervised and self-supervised representation learning techniques as discussed in Section 2.3 and Section 3.5.1, which can be especially beneficial for end-to-end methods. Such methods can also help to overcome the shortage of datasets dedicated to a specific task.

A better understanding of biases in the audio-visual data and the source separation problem can result in the integration of auxiliary losses for the source reconstruction. Different techniques for advanced motion analysis in musicians' movements, that can include more precise fingering analysis and the usage of skeleton data as in [Gan et al., 2020], may also result in enhanced performance.

Finally, the current work can be extended by integrating more modalities into the multimodal analysis, such as, for example, visual musical scores or corresponding MIDI files.

6.3. List of contributions

6.3.1. Scientific publications

The scientific contributions of the thesis can be represented with the following list of publications, organized by research problems addressed in this work:

Musical Instrument Recognition

- **"Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies"**. O. Slizovskaia, E. Gómez and G. Haro. In *Proceedings of 13th Sound and Music Computing Conference (SMC). 2016.*

- **"Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture"**. O. Slizovskaia, E. Gómez and G. Haro. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR). 2017.*

- **"Timbre analysis of music audio signals with convolutional neural networks".** J. Pons, O. Slizovskaia, R, Gong, E Gómez and X. Serra. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO). 2017.*

Interpretability of Audio-Visual Deep Learning Models

- **"Correspondence between audio and visual deep models for musical instrument detection in video recordings"**. O. Slizovskaia, E. Gómez and G. Haro. In *The 18th International Society for Music Information Retrieval Conference (ISMIR17), Late-breaking/demo session (LBD). 2017.*

- **"A Case Study of Deep-Learned Activations via Hand-Crafted Audio Features"**. O. Slizovskaia, E. Gómez and G. Haro. In *The*

2018 Joint Workshop on Machine Learning for Music, The Federated Artificial Intelligence Meeting (FAIM), Joint workshop program of ICML, IJCAI/ECAI, and AAMAS. 2018.

Source Separation in Music Performance Videos

- **"End-to-end sound source separation conditioned on instrument labels".** O. Slizovskaia, L. Kim, G. Haro and E Gómez. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019.*

- **"Conditioned Source Separation for Music Instrument Performances".** O.Slizovskaia, G. Haro and E. Gómez. Under review for *The IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

- **"Solos: A Dataset for Audio-Visual Music Analysis".** J.F. Montesinos, O. Slizovskaia and G. Haro. Submitted to *IEEE 22nd International Workshop on Multimedia Signal Processing, 2020.*

Contributed Work on Relevant Problems

- **"Vocoder-Based Speech Synthesis from Silent Videos".** D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.H. Tan and J. Jensen. Submitted to *The 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020). 2020.*

- **"Input complexity and out-of-distribution detection with likelihood-based generative models".** J. Serrà, D. Álvarez, V. Gómez, O Slizovskaia, J.F. Núñez, and J. Luque. In *Proceedings of the 8th International Conference on Learning Representations (ICLR). 2020.*

- **"Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks".** E. Fonseca,

R. Gong, D. Bogdanov, O. Slizovskaia, E. Gómez and X. Serra. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). 2017.*

- **"Acoustic scene classification by fusing LightGBM and VGG-net multichannel predictions".** R. Gong, E. Fonseca, D. Bogdanov, O. Slizovskaia, E. Gómez and X. Serra. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). 2017.*

6.3.2. Tools and assets

Aiming to facilitate the reproducibility of the proposed methods, we published a set of tools and assets for the problems of instrument classification and source separation:

Assets related to musical instrument classification

- Open code for experiments published in [Slizovskaia et al., 2016]: `https://github.com/Veleslavia/SMC2016`
- Open code for experiments published in [Slizovskaia et al., 2017]: `https://github.com/Veleslavia/ICMR2017`
- Open code for experiments published in [Pons et al., 2017]: `https://github.com/Veleslavia/EUSIPCO2017`

Assets related to sound source separation

- Two open implementations for end-to-end sound source separation published in [Slizovskaia et al., 2019]:
  `https://github.com/Veleslavia/vimss`
  `https://github.com/Veleslavia/vimss_torch`
- Open code and a project page for experiments on conditioned source separation published in [Slizovskaia et al., 2020] (under

preparation):
https://github.com/veleslavia/conditioned-u-net
https://veleslavia.github.io/conditioned-u-net

- Open dataset for audio-visual music analysis published in [Montesinos et al., 2020]:
https://www.juanmontesinos.com/Solos/

Bibliography

[Aarabi and Zaky, 2001] Aarabi, P. and Zaky, S. (2001). Robust sound localization using multi-source audiovisual information fusion. *Information Fusion*, 2(3):209 – 223.

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., and others (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[Abu-El-Haija et al., 2016] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint*.

[Afouras et al., 2018] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Aggarwal and Ryoo, 2011] Aggarwal, J. and Ryoo, M. (2011). Human activity analysis: A review. *ACM Comput. Surv.*, 43(3).

[Alías et al., 2016] Alías, F., Socoró, J. C., and Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 6(5):143.

[Arandjelovic and Zisserman, 2017] Arandjelovic, R. and Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617.

[Arandjelovic and Zisserman, 2018] Arandjelovic, R. and Zisserman, A. (2018). Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451.

[Asano et al., 2020] Asano, Y. M., Patrick, M., Rupprecht, C., and Vedaldi, A. (2020). Labelling unlabelled videos from scratch with multi-modal self-supervision.

[Atrey et al., 2010] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

[Aytar et al., 2016] Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

[Bahl et al., 1983] Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

[Baltrušaitis et al., 2018] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

[Barleycorn, 2019] Barleycorn, W. (2019). Machine learning in dance - cross-modal music suggestion based on expressive dance movement queries. Master's thesis, Pompeu Fabra University.

[Barzelay and Schechner, 2007] Barzelay, Z. and Schechner, Y. Y. (2007). Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[Bau et al., 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Csail, A. T. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549.

[Bazzica et al., 2014] Bazzica, A., Liem, C. C., and Hanjalic, A. (2014). Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music. In *ISMIR*, pages 201–206.

[Bazzica et al., 2016] Bazzica, A., Liem, C. C., and Hanjalic, A. (2016). On detecting the playing/non-playing activity of musicians in symphonic music videos. *Computer Vision and Image Understanding*, 144:188–204.

[Bell, 2018] Bell, J. (2018). Audiovisual scores and parts synchronized over the web. In *The International Conference on Technologies for Music Notation and Representation (TENOR)*.

[Bello et al., 2004] Bello, J. P., Duxbury, C., Davies, M., and Sandler, M. (2004). On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[Beyer et al., 2020] Beyer, L., HÃ©naff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. (2020). Are we done with imagenet?

[Bhalerao et al., 2020] Bhalerao, R. H., Kshirsagar, V., and Raval, M. (2020). Finger tracking based tabla syllable transcription. In Palaiahnakote, S., Sanniti di Baja, G., Wang, L., and Yan, W. Q., editors, *Asian Conference on Pattern Recognition*, pages 569–579. Springer International Publishing.

[Bogdanov et al., 2013] Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., Serra, X., et al. (2013). Essentia: An audio analysis

library for music information retrieval. In *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 493–498. International Society for Music Information Retrieval (ISMIR).

[Bosch et al., 2012] Bosch, J., Janer, J., Fuhrmann, F., and Herrera, P. (2012). A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, pages 559–564, Porto, Portugal.

[Caba Heilbron et al., 2015] Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

[Cao et al., 2019] Cao, Y., Summerfield, C., Park, H., Giordano, B. L., and Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, 102(5):1076–1087.

[Carabias-Orti et al., 2013] Carabias-Orti, J. J., Cobos, M., Vera-Candeas, P., and Rodríguez-Serrano, F. J. (2013). Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings. *EURASIP Journal on Advances in Signal Processing*, 2013(1):184.

[Carabias-Orti et al., 2011] Carabias-Orti, J. J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., and Canadas-Quesada, F. J. (2011). Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1144–1158.

[Carrive et al., 2000] Carrive, J., Pachet, F., and Ronfard, R. (2000). Clavis-a temporal reasoning system for classification of audiovisual sequences. In *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAOAt*.

[Chan et al., 2016] Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

[Chandna et al., 2017] Chandna, P., Miron, M., Janer, J., and Gómez, E. (2017). Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. In *13th International Conference on Latent Variable Analysis and Signal Separation (LVA ICA2017)*.

[Choi et al., 2017] Choi, H., Lee, J.-h., and Lee, K. (2017). Singing voice separation using generative adversarial networks,. In *ML4Audio Workshop, 31st Conf. Neural Information Processing Systems (NIPS 2017)*.

[Choi et al., 2016] Choi, K., Fazekas, G., and Sandler, M. (2016). Automatic Tagging Using Deep Convolutional Neural Networks. In *International Society of Music Information Retrieval Conference, New York, USA*. ISMIR.

[Chollet, 2016] Chollet, F. (2016). Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357*.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. https://keras.io.

[Chung et al., 2017] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE.

[Dahl and Friberg, 2007] Dahl, S. and Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception*, 24(5):433–454.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image

Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Dhariwal et al., 2020] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

[Dieleman, 2014] Dieleman, S. (2014). Recommending music on Spotify with deep learning.

[Doire and Okubadejo, 2019] Doire, C. S. J. and Okubadejo, O. (2019). Interleaved Multitask Learning for Audio Source Separation with Independent Databases. *arXiv e-prints*, page arXiv:1908.05182.

[Domínguez, 2015] Domínguez, A. (2015). A history of the convolution operation. *IEEE Pulse Retrospectroscope*, 6(1):38–49.

[Dov et al., 2017] Dov, D., Talmon, R., and Cohen, I. (2017). Multimodal kernel method for activity detection of sound sources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1322–1334.

[Duan et al., 2019] Duan, Z., Essid, S., Liem, C. C. S., Richard, G., and Sharma, G. (2019). Audiovisual analysis of music performances: Overview of an emerging field. *IEEE Signal Processing Magazine*, 36(1):63–73.

[Dumoulin et al., 2018] Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H. d., Courville, A., and Bengio, Y. (2018). Feature-wise transformations. *Distill*. https://distill.pub/2018/feature-wise-transformations.

[Dupont and Luettin, 2000] Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.

[Ephrat et al., 2018] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual

model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4):112.

[Erhan et al., 2010] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why Does Unsupervised Pre-training Help Deep Learning? *J. Mach. Learn. Res.*, 11:625–660.

[Espi et al., 2015] Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (2015). Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):26.

[Essid and Richard, 2012] Essid, S. and Richard, G. (2012). Fusion of Multimodal Information in Music Content Analysis. In Müller, M., Goto, M., and Schedl, M., editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 37–52. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

[Fang et al., 2018] Fang, J. T., Chang, Y. R., and Chang, P. C. (2018). Deep learning of chroma representation for cover song identification in compression domain. *Multidimensional Systems and Signal Processing*, 29(3):887–902.

[Feichtenhofer et al., 2016] Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941.

[Finn and Montgomery, 1988] Finn, K. E. and Montgomery, A. A. (1988). Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164.

[Fonseca et al., 2019] Fonseca, E., Plakal, M., Ellis, D. P., Font, F., Favory, X., and Serra, X. (2019). Learning sound event classifiers from web audio with noisy labels. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE.

[Gan et al., 2020] Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. (2020). Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487.

[Gao et al., 2018] Gao, R., Feris, R., and Grauman, K. (2018). Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53.

[Gao and Grauman, 2019] Gao, R. and Grauman, K. (2019). Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3879–3888.

[Gao et al., 2019] Gao, R., Oh, T.-H., Grauman, K., and Torresani, L. (2019). Listen to look: Action recognition by previewing audio. *arXiv preprint arXiv:1912.04487*.

[Gemmeke et al., 2017] Gemmeke, J., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[Gillet and Richard, 2005] Gillet, O. and Richard, G. (2005). Automatic transcription of drum sequences using audiovisual features. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3. IEEE.

[Gillet and Richard, 2006] Gillet, O. and Richard, G. (2006). Comparing audio and video segmentations for music videos indexing. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V.

[Goldstein and Moses, 2018] Goldstein, S. and Moses, Y. (2018). Guitar music transcription from silent video. In *British Machine Vision Conference*, pages 309–321. BMVA Press.

[Gorodnichy and Yogeswaran, 2006] Gorodnichy, D. O. and Yogeswaran, A. (2006). Detection and tracking of pianist hands and fingers. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 63–63.

[Goto, 2004] Goto, M. (2004). Development of the rwc music database. In *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, pages 553–556.

[Grais et al., 2016] Grais, E. M., Roma, G., Simpson, A. J., and Plumbley, M. (2016). Combining mask estimates for single channel audio source separation using deep neural networks. *Interspeech2016 Proceedings*.

[Griffiths and Reay, 2018] Griffiths, N. K. and Reay, J. L. (2018). The relative importance of aural and visual information in the evaluation of western canon music performance by musicians and nonmusicians. *Music Perception*, 35(3):364–375.

[Han et al., 2016] Han, Y., Kim, J., and Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *arXiv preprint arXiv:1605.09507*.

[Han and Raphael, 2010] Han, Y. and Raphael, C. (2010). Informed source separation of orchestra and soloist. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 315–320.

[He et al., 2016a] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[He et al., 2016b] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Herrera-Boyer et al., 2003] Herrera-Boyer, P., Peeters, G., and Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research*, 32(1):3–21.

[Hershey et al., 2004] Hershey, J., Attias, H., Jojic, N., and Kristjansson, T. (2004). Audio-visual graphical models for speech processing. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–649. IEEE.

[Hershey and Movellan, 2000] Hershey, J. R. and Movellan, J. R. (2000). Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819.

[Hershey et al., 2016] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K. (2016). CNN Architectures for Large-Scale Audio Classification. *arXiv preprint arXiv:1609.09430*.

[Hori et al., 2017] Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., and Sumi, K. (2017). Attention-Based Multimodal Fusion for Video Description. In *Conference on Computer Vision and Pattern Recognition*, pages 4193–4202.

[Hu et al., 2016] Hu, D., Li, X., and others (2016). Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582.

[Hua et al., 2012] Hua, S., Chen, G., Wei, H., and Jiang, Q. (2012). Similarity measure for image resizing using SIFT feature. *EURASIP Journal on Image and Video Processing*, 2012(1):6.

[Huang and Kingsbury, 2013] Huang, J. and Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7596–7599. IEEE.

[Humphrey and Gryner, 2015] Humphrey, S. and Gryner, F. (United States Patent, US20150046824A1, 2015). Synchronized display and

performance mapping of musical performances submitted from remote locations.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 448–456.

[Jansson et al., 2017] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks. In *18th International Society for Music Information Retrieval Conference*, pages 23–27.

[Jiang et al., 2018] Jiang, Y. G., Wu, Z., Wang, J., Xue, X., and Chang, S. F. (2018). Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):352–364.

[Jinqiao Wang et al., 2006] Jinqiao Wang, Lingyu Duan, Hanqing Lu, Jin, J. S., and Changsheng Xu (2006). A mid-level scene change representation via audiovisual alignment. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II.

[Jordan and Mitchell, 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

[Katsaggelos et al., 2015] Katsaggelos, A. K., Bahaadini, S., and Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653.

[Kavalerov et al., 2019] Kavalerov, I., Wisdom, S., Erdogan, H., Patton, B., Wilson, K., Roux, J. L., and Hershey, J. R. (2019). Universal sound separation. *arXiv preprint arXiv:1905.03330*.

[Kelley, 1960] Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954.

[Kidron et al., 2005] Kidron, E., Schechner, Y. Y., and Elad, M. (2005). Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95. IEEE.

[Kilgour et al., 2019] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. *Proc. Interspeech 2019*, pages 2350–2354.

[Kim et al., 2013] Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691. IEEE.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Koepke et al., 2020] Koepke, A. S., Wiles, O., Moses, Y., and Zisserman, A. (2020). Sight to sound: an end-to-end approach for visual piano

transcription. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[Korbar et al., 2018] Korbar, B., Tran, D., and Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774.

[Le et al., 2011] Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 265–272.

[Le Roux et al., 2019] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR – half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[Lee et al., 2019] Lee, H.-Y., Yang, X., Liu, M.-Y., Wang, T.-C., Lu, Y.-D., Yang, M.-H., and Kautz, J. (2019). Dancing to music. In *Advances in Neural Information Processing Systems*, pages 3586–3596.

[Lee, 2004] Lee, K. (2004). *An analysis and comparison of the clarinet and viola versions of the two sonatas for clarinet (or viola) and piano Op. 120 by Johannes Brahms*. PhD thesis, University of Cincinnati.

[Leman, 2017] Leman, M. (2017). The Interactive Dialectics of Musical Meaning Formation. In Lesaffre, M., Maes, P.-J., and Leman, M., editors, *The Routledge Companion to Embodied Music Interaction*, pages 13–21. Routledge.

[Li et al., 2017] Li, B., Dinesh, K., Duan, Z., and Sharma, G. (2017). See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910. IEEE.

[Li et al., 2018a] Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2018a). Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535.

[Li et al., 2019] Li, B., Parekh, S., Duan, Z., and Essid, S. (2019). Ismir2019 tutorial 3: Audiovisual music processing. `https://github.com/bochen1106/ISMIR2019-Tutorial3-Audiovisual-Music-Processing`. Accessed on July 6th, 2020.

[Li et al., 2018b] Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., and Snoek, C. G. (2018b). VideoLSTM convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50.

[Liem et al., 2011] Liem, C. C., Müller, M., Eck, D., Tzanetakis, G., and Hanjalic, A. (2011). The need for music information retrieval with user-centered and multimodal strategies. In *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops - MIRUM 2011 Workshop, MIRUM'11*, pages 1–6.

[Lim et al., 2010] Lim, A., Mizumoto, T., Cahier, L.-K., Otsuka, T., Takahashi, T., Komatani, K., Ogata, T., and Okuno, H. G. (2010). Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1964–1969. IEEE.

[Liu et al., 2019] Liu, J., Yang, Y., and Jeng, S. (2019). Weakly-supervised visual instrument-playing action detection in videos. *IEEE Transactions on Multimedia*, 21(4):887–901.

[Liu et al., 2019] Liu, J. Y., Yang, Y. H., and Jeng, S. K. (2019). Weakly-Supervised Visual Instrument-Playing Action Detection in Videos. *IEEE Transactions on Multimedia*, 21(4):887–901.

[Llagostera Casanovas et al., 2010] Llagostera Casanovas, A., Monaci, G., Vandergheynst, P., and Gribonval, R. (2010). Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371.

[Lluis et al., 2018] Lluis, F., Pons, J., and Serra, X. (2018). End-to-end music source separation: is it possible in the waveform domain? *arXiv preprint arXiv:1810.12187*.

[Lostanlen and Cella, 2016] Lostanlen, V. and Cella, C.-E. (2016). Deep convolutional networks on the pitch spiral for musical instrument recognition. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York City, NY, USA.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE.

[Lu et al., 2019] Lu, R., Duan, Z., and Zhang, C. (2019). Audio–visual deep clustering for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1697–1712.

[Luo and Mesgarani, 2018] Luo, Y. and Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

[Maezawa and Yamamoto, 2016] Maezawa, A. and Yamamoto, K. (2016). Automatic music accompaniment based on audio-visual score following. In *17th International Society for Music Information Retrieval Conference (ISMIR), Late-breaking/demo (LDB)*.

[Maragos et al., 2008] Maragos, P., Gros, P., Katsamanis, A., and Papandreou, G. (2008). Cross-modal integration for performance improving in multimedia: A review. In *Multimodal processing and interaction*, pages 1–46. Springer.

[Marchini et al., 2014] Marchini, M., Ramirez, R., Papiotis, P., and Maestre, E. (2014). The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43(3):303–317.

[Marenco et al., 2015] Marenco, B., Fuentes, M., Lanzaro, F., Rocamora, M., and Gómez, A. (2015). A Multimodal Approach for Percussion Music Transcription from Audio and Video. In Pardo, A. and Kittler, J., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings*, pages 92–99. Springer International Publishing, Cham.

[McFee et al., 2015] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

[Meseguer-Brocal and Peeters, 2019] Meseguer-Brocal, G. and Peeters, G. (2019). Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations. In *20th International Society for Music Information Retrieval Conference (ISMIR)*.

[Michelsanti et al., 2020] Michelsanti, D., Slizovskaia, O., Haro, G., Gómez, E., Tan, Z.-H., and Jensen, J. (2020). Vocoder-based speech synthesis from silent videos. *arXiv preprint arXiv:2004.02541*.

[Miron et al., 2016] Miron, M., Carabias-Orti, J. J., Bosch, J. J., Gómez, E., and Janer, J. (2016). Score-informed source separation for multichannel orchestral recordings. *Journal of Electrical and Computer Engineering*, 2016.

[Mishra, 2019] Mishra, D. (2019). Convolution vs correlation. `https://towardsdatascience.com/convolution-vs-correlation-af868b6b4fb5`. Accessed: 2020-04-23.

[Mishra et al., 2017] Mishra, S., Sturm, B. L., and Dixon, S. (2017). Local Interpretable Model-Agnostic Explanations for Music Content Analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*.

[Montavon et al., 2017] Montavon, G., Bach, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222.

[Montes et al., 2016] Montes, A., Salvador, A., Pascual, S., and Giro-i Nieto, X. (2016). Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks. In *1st NIPS Workshop on Large Scale Computer Vision Systems*.

[Montesinos et al., 2020] Montesinos, J. F., Slizovskaia, O., and Haro, G. (2020). Solos: A dataset for audio-visual music source separation and localization. *Under review for IEEE 22nd International Workshop on Multimedia Signal Processing*.

[Nam et al., 1998] Nam, J., Alghoniemy, M., and Tewfik, A. H. (1998). Audio-visual content-based violent scene characterization. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, volume 1, pages 353–357 vol.1.

[Nanni et al., 2017] Nanni, L., Costa, Y. M. G., Lucio, D. R., Jr., C. N. S., and Brahnam, S. (2017). Combining visual and acoustic features for audio classification tasks. *Pattern Recognition Letters*, 88:49–56.

[Naphade et al., 2002] Naphade, M. R., Lin, C. Y., and Smith, J. R. (2002). Learning semantic multimedia representations from a small set of examples. In *Proceedings - 2002 IEEE International Conference on Multimedia and Expo, ICME 2002*, volume 2, pages 513–516. Institute of Electrical and Electronics Engineers Inc.

[Nefian et al., 2002] Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002). Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):783042.

[Ng et al., 2015] Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702.

[Ngiam et al., 2011] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Oramas et al., 2018] Oramas, S., Barbieri, F., Nieto, O., and Serra, X. (2018). Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1(1):4–21.

[Owens and Efros, 2018] Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648.

[Ozerov and Févotte, 2009] Ozerov, A. and Févotte, C. (2009). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563.

[Pang and Ngo, 2015] Pang, L. and Ngo, C.-W. (2015). Mutlimodal Learning with Deep Boltzmann Machine for Emotion Prediction in User Generated Videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 619–622, New York, NY, USA. ACM.

[Parekh, 2019] Parekh, S. (2019). *Learning representations for robust audio-visual scene analysis*. PhD thesis, Universite Paris-Saclay, Telecom ParisTech.

[Parekh et al., 2017] Parekh, S., Essid, S., Ozerov, A., Duong, N. Q., Perez, P., and Richard, G. (2017). Guiding audio source separation by video object information. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, volume 2017-Octob, pages 61–65. IEEE.

[Parekh et al., 2019a] Parekh, S., Essid, S., Ozerov, A., Duong, N. Q., Pérez, P., and Richard, G. (2019a). Weakly supervised representation learning for audio-visual scene analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:416–428.

[Parekh et al., 2019b] Parekh, S., Ozerov, A., Essid, S., Duong, N. Q., Pérez, P., and Richard, G. (2019b). Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 268–272. IEEE.

[Patrick et al., 2020] Patrick, M., Kuznetsova, Y. M. A. P., Fong, R., Henriques, J. F., Zweig, G., and Vedaldi, A. (2020). Multi-modal self-supervision from generalized data transformations.

[Perez et al., 2018] Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Pishdadian et al., 2019] Pishdadian, F., Wichern, G., and Roux, J. L. (2019). Finding strength in weakness: Learning to separate sounds with weak supervision. *arXiv preprint arXiv:1911.02182*.

[Platz and Kopiez, 2012] Platz, F. and Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, 30(1):71–83.

[Pons and Serra, 2017] Pons, J. and Serra, X. (2017). Designing Efficient Architectures for Modeling Temporal Features with Convolutional Neural Networks. In *42th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, New Orleans, USA. IEEE, IEEE.

[Pons and Serra, 2019] Pons, J. and Serra, X. (2019). musicnn: Pre-trained convolutional neural networks for music audio tagging. In *20th International Society for Music Information Retrieval Conference (ISMIR), Late-breaking/demo (LDB)*.

[Pons et al., 2017] Pons, J., Slizovskaia, O., Gong, R., Gómez, E., and Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2744–2748. IEEE.

[Qiu et al., 2017] Qiu, Z., Yao, T., and Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541.

[Rafii et al., 2018] Rafii, Z., Liutkus, A., Stoter, F.-R., Mimilakis, S. I., FitzGerald, D., and Pardo, B. (2018). An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(8):1307–1335.

[Reilly and McGrath, 1995] Reilly, A. and McGrath, D. (1995). Convolution processing for realistic reverberation. In *Audio Engineering Society Convention 98*. Audio Engineering Society.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Salamon and Bello, 2017] Salamon, J. and Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3):279–283.

[Sasaki et al., 2015] Sasaki, S., Hirai, T., Ohya, H., and Morishima, S. (2015). Affective music recommendation system based on the mood of input video. In He, X., Luo, S., Tao, D., Xu, C., Yang, J., and Hasan, M. A., editors, *MultiMedia Modeling*, pages 299–302. Springer International Publishing.

[Schedl et al., 2014] Schedl, M., Gómez, E., Urbano, J., et al. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261.

[Schindler, 2019] Schindler, A. (2019). *Multi-Modal Music Information Retrieval: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis*. PhD thesis, TU Wien.

[Schindler and Rauber, 2015] Schindler, A. and Rauber, A. (2015). *An Audio-Visual Approach to Music Genre Classification through Affective Color Features*, pages 61–67. Springer International Publishing, Cham.

[Schluter and Bock, 2014] Schluter, J. and Bock, S. (2014). Improved musical onset detection with convolutional neural networks. In *39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pages 6979–6983. IEEE.

[Schulze-Forster et al., 2019] Schulze-Forster, K., Doire, C., Richard, G., and Badeau, R. (2019). Weakly informed audio source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 268–272.

[Schwartz et al., 2002] Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2002). Audio-visual scene analysis: evidence for a" very-early" integration process in audio-visual speech perception. In *Seventh International Conference on Spoken Language Processing*.

[Seetharaman et al., 2019] Seetharaman, P., Wichern, G., Venkataramani, S., and Le Roux, J. (2019). Class-conditional embeddings for music source separation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305. IEEE.

[Senocak et al., 2019] Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. (2019). Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Shaffer and Pletzer, 2009] Shaffer, J. R. and Pletzer, K. (United States Patent, US8481839B2, 2009). System and methods for synchroniz-

ing audio and/or visual playback with a fingering display for musical instrument.

[Shams et al., 2002] Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. In *Cognitive Brain Research*, volume 14, pages 147–152. Elsevier.

[Shrestha et al., 2010] Shrestha, P., Barbieri, M., Weda, H., and Sekulovski, D. (2010). Synchronization of multiple camera videos using audio-visual features. *IEEE Transactions on Multimedia*, 12(1):79–92.

[Sigtia et al., 2016] Sigtia, S., Benetos, E., and Dixon, S. (2016). An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939.

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576.

[Slizovskaia et al., 2016] Slizovskaia, O., Gómez, E., and Haro, G. (2016). Automatic musical instrument recognition in audiovisual recordings by combining image and audio classification strategies. In *13th Sound and Music Computing Conference (SMC 2016)*, Hamburg, Germany.

[Slizovskaia et al., 2017] Slizovskaia, O., Gómez, E., and Haro, G. (2017). Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture. In *ICMR '17: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, Bucharest, Romania.

[Slizovskaia et al., 2020] Slizovskaia, O., Haro, G., and Gómez, E. (2020). Conditioned source separation for music instrument performances. *Under review for The IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[Slizovskaia et al., 2019] Slizovskaia, O., Kim, L., Haro, G., and Gomez, E. (2019). End-to-end sound source separation conditioned on instrument labels. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 306–310. IEEE.

[Smaragdis and Casey, 2003] Smaragdis, P. and Casey, M. (2003). Audio/visual independent components. In *International Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, pages 709–714.

[Sohn et al., 2014] Sohn, K., Shang, W., and Lee, H. (2014). Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149.

[Srivastava and Salakhutdinov, 2012] Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

[Stoller et al., 2018a] Stoller, D., Ewert, S., and Dixon, S. (2018a). Adversarial semi-supervised audio source separation applied to singing voice extraction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2391–2395. IEEE.

[Stoller et al., 2018b] Stoller, D., Ewert, S., Dixon, S., et al. (2018b). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*. 19th International Society for Music Information Retrieval Conference (ISMIR).

[Stöter et al., 2018] Stöter, F.-R., Liutkus, A., and Ito, N. (2018). The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305. Springer.

[Stöter et al., 2019] Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-Unmix – A Reference Implementation for Music Source Separation. *Journal of Open Source Software*.

[Su et al., 2020] Su, K., Liu, X., and Shlizerman, E. (2020). Audeo: Audio generation for a silent performance video. *CoRR*, abs/2006.14348.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

[Thompson and Luck, 2012] Thompson, M. R. and Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, 16(1):19–40.

[Thompson et al., 2005] Thompson, W. F., Graham, P., and Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, 2005(156):203–227.

[Tran et al., 2015] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA. IEEE Computer Society.

[Tsuchida et al., 2019a] Tsuchida, S., Fukayama, S., and Goto, M. (2019a). Query-by-dancing: a dance music retrieval system based on body-motion similarity. In *International Conference on Multimedia Modeling*, pages 251–263. Springer.

[Tsuchida et al., 2019b] Tsuchida, S., Fukayama, S., Hamasaki, M., and Goto, M. (2019b). Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands.

[Tzinis et al., 2019] Tzinis, E., Wisdom, S., Hershey, J. R., Jansen, A., and Ellis, D. P. W. (2019). Improving universal sound separation using sound classification. *arXiv preprint arXiv:1911.07951*.

[Uhlich et al., 2017] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 261–265. IEEE.

[van den Oord et al., 2016] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609:03499*, abs/1609.0.

[Varol et al., 2017] Varol, G., Laptev, I., and Schmid, C. (2017). Long-term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469.

[Virtanen, 2007] Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074.

[Visentini et al., 2011] Visentini, I., Rodà, A., Canazza, S., and Snidaro, L. (2011). Audio-video analysis of musical expressive intentions. In Maino, G. and Foresti, G. L., editors, *Image Analysis and Processing – ICIAP 2011*, pages 219–228, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Von Hornbostel and Sachs, 1914] Von Hornbostel, E. M. and Sachs, C. (1914). Systematik der musikinstrumente. ein versuch. *Zeitschrift für Ethnologie*, 46(H. 4/5):553–590.

[Voulodimos et al., 2018] Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

[Wang et al., 2015] Wang, S., Ewert, S., and Dixon, S. (2015). Compensating for asynchronies between musical voices in score-performance alignment. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 589–593. IEEE.

[Wang et al., 2020] Wang, W., Tran, D., and Feiszli, M. (2020). What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

[Wang et al., 2007] Wang, Y., Zhang, B., and Schleusing, O. (2007). Educational violin transcription by fusing multimedia streams. In *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*, pages 57–66. Association for Computing Machinery.

[Weng, 2019] Weng, L. (2019). Self-Supervised Representation Learning.

[Wikipedia contributors, 2020] Wikipedia contributors (2020). Loudness — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Loudness. Accessed: 2020-04-23.

[Wisdom et al., 2019] Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., and Saurous, R. A. (2019). Differentiable consistency constraints for improved deep speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 900–904. IEEE.

[Wu et al., 2015] Wu, Z., Wang, X., Jiang, Y.-G., Ye, H., and Xue, X. (2015). Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 461–470.

[Xu et al., 2016] Xu, B., Fu, Y., Jiang, Y.-G., Li, B., and Sigal, L. (2016). Video Emotion Recognition with Transferred Deep Feature Encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 15–22, New York, NY, USA. ACM.

[Xu et al., 2019] Xu, X., Dai, B., and Lin, D. (2019). Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–891.

[Yesiler et al., 2020] Yesiler, F., Serrà, J., and Gómez, E. (2020). Accurate and scalable version identification using musically-motivated embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25.

[Yuhas et al., 1989] Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71.

[Zhang et al., 2019] Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., and Chen, D.-S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005.

[Zhang et al., 2016] Zhang, S., Zhang, S., Huang, T., and Gao, W. (2016). Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 281–284. ACM.

[Zhang and Kuo, 2001] Zhang, T. and Kuo, C.-C. J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4):441–457.

[Zhao et al., 2019] Zhao, H., Gan, C., Ma, W.-C., and Torralba, A. (2019). The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744.

[Zhao et al., 2018] Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586.

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

[Zhuang et al., 2020] Zhuang, W., Wang, C., Xia, S., Chai, J., and Wang, Y. (2020). Music2dance: Music-driven dance generation using wavenet. *arXiv preprint arXiv:2002.03761*.

[Zijl and Luck, 2013] Zijl, A. G. W. V. and Luck, G. (2013). Moved through music: The effect of experienced emotions on performers' movement characteristics. *Psychology of Music*, 41(2):175–197.

[Zinemanas et al., 2017] Zinemanas, P., Haro, G., and Emilia, G. (2017). Visual music transcription of clarinet video recordings trained with audio-based labelled data. In *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media (CVAVM)*.

[Žitnik and Zupan, 2014] Žitnik, M. and Zupan, B. (2014). Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53.

Appendix A

Appendix

A.1. Hyperparameters of the experiments from Section 5.3.5

We provide the full set of model hyperparameters used in the experiments in Section 5.3.5 and Section 5.3.6 in Table A.1. Note, that there is only a single difference within each pair of the experiments compared in Table 5.3. For the experiments in Section 5.3.6 the model parameters are set as described in Section 5.3.5.

ID	STFT F-scale	STFT V-scale	model	noise	mask	bias	loss	curr.	cond. type
1	log	dB-norm	U-Net	No	Binary	No	BCE	Yes	None
2	linear	dB-norm	U-Net	No	Binary	No	BCE	Yes	None
3	linear	dB-norm	MHU-Net	No	Binary	No	BCE	Yes	FiLM-bottleneck
4	linear	dB-norm	MHU-Net	No	Ratio	No	L_1^{smooth}	Yes	FiLM-bottleneck
5	linear	dB-norm	MHU-Net	No	Binary	Yes	BCE	Yes	FiLM-bottleneck
6	linear	dB-norm	MHU-Net	Yes	Binary	Yes	BCE	Yes	FiLM-bottleneck
7	linear	log	MHU-Net	No	Binary	No	BCE	Yes	FiLM-bottleneck
8	linear	dB-norm	MHU-Net	No	Binary	Yes	BCE	No	None
9	linear	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-bottleneck
10	linear	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-encoder
11	linear	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-final
12	linear	dB-norm	U-Net	No	Binary	No	BCE	Yes	Label-multiply
13	log	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-encoder
14	log	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-bottleneck
15	log	dB-norm	U-Net	No	Binary	No	BCE	Yes	FiLM-final
16	log	dB-norm	U-Net	No	Binary	No	BCE	Yes	Label-multiply

Table A.1: Ablation study parameters and corresponding experiment IDs for Conditioned U-Net.

A.2. Per-experiment bar plots with source separation performance results

Figure A.1 shows source separation metrics (SI-SDR, SD-SDR, PES) pictured as bar plots with mean and standard deviation for each experiment conducted in Section 5.3.5 and Section 5.3.6. The experiment are referenced by id as in Table 5.3 and 5.4.

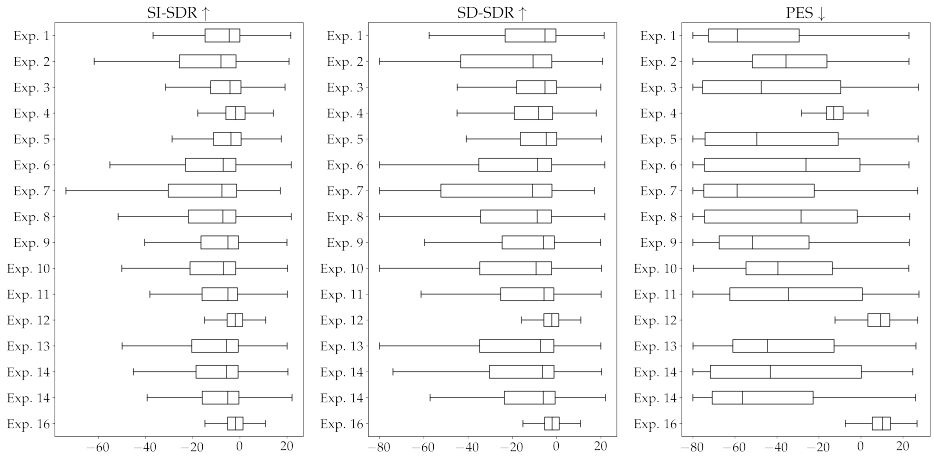


Figure A.1: SI-SDR, SD-SDR and PES boxplots for the experiments from Section 5.3.3. Experiments are referenced by ID.