

Waveform-based music processing with deep learning Sander Dieleman, Jordi Pons, Jongpil Lee



Historical perspective

Why now?

Papers on "neural networks & music": milestones

80

papers

70 **End-to-end** learning for automatic **music composition** (van den Oord et al., 2016) End-to-end learning for music audio classification (Dieleman & Schrauwen, 2014) 60 **CNN** learns from **spectrograms** for music audio **classification** (Lee et al., 2009) 50 **LSTM** from symbolic data for automatic music composition (Eck & Schmidhuber, 2002) 40 **MLP** learns from **spectrograms** to **classify** pop/classical (Matityaho & Furst, 1995) 30 **RNN** from symbolic data for automatic music composition (Todd, 1988) 20 **MLP** from symbolic data for automatic music composition (Lewis, 1988) 10 0 1990

Papers on "neural networks & music": input data



References

Van den Oord et al., 2016. "Wavenet: a generative model for audio" in arXiv.

Dieleman and Schrauwen, 2014. "End-to-end learning for music audio" in International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Lee et al., 2009. "Unsupervised feature learning for audio classification using convolutional deep belief networks" in Advances in Neural Information Processing Systems (NIPS).

Matityaho and Furst, 1995. "Neural network based model for classification of music type" in 18th Convention of Electrical and Electronics Engineers in Israel. IEEE, 4–3.

Eck and Schmidhuber, 2002. "Finding temporal structure in music: Blues improvisation with LSTM recurrent networks" *in Proceedings of the Workshop on Neural Networks for Signal Processing.*

Todd, 1988. "A sequential network design for musical applications" in Proceedings of the Connectionist Models Summer School.

Lewis, 1988. Creation by Refinement: A creativity paradigm for gradient descent learning networks" *in International Conference on Neural Networks.*

Historical review from: Pons, 2019. "Deep neural networks for music and audio tagging". PhD Thesis.

Waveform-based music processing with deep learning

Classification

by Jongpil Lee, PhD candidate at KAIST in Daejeon, South Korea.

Source Separation

by Jordi Pons, researcher at Dolby Laboratories in Barcelona.

Generation

by Sander Dieleman, research scientist at DeepMind in London, UK.

Music Classification

Learning specific properties of music

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

- Problem statement

- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Music classification



Celma, Oscar, 2010 "Music recommendation", *Music recommendation and discovery, Springer*.

Music

Music (artwork)



Scene



Speech (object)



Audio

Image







Multi source polyphonicMulti timbre (texture) polyphonicMostly single speakerUnstructured sound sourcesStructure of sound sourcesA structured sound sourceDynamic control (mixing)



Semantic labels are highly diverse and have different levels of abstraction

Grosche et al., 2012 "Audio content-based music retrieval", Dagstuhl Follow-Ups.

Casey et al., 2008 "Content-based music information retrieval: Current directions and future challenges", Proceedings of the IEEE.

Problem definition



- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

From feature engineering to end-to-end learning



Nam et al., 2018. "Deep learning for audio-based music classification and tagging", IEEE Signal Processing Magazine.

Humphrey et al., 2013. "Feature learning and deep architectures: new directions for music informatics", Journal of Intelligent Information Systems.

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Engineered feature based model



- MFCC, Chroma, Bag of low-level-features

- Mel-spectrogram

- Scattering transform

- 1D CNNs

- 2D CNNs

- RNNs

- Attention

Pons et al., 2018. "End-to-end learning for music audio tagging at scale", *ISMIR*. Choi et al., 2016. "Automatic tagging using deep convolutional neural networks", *ISMIR*.

Purwins et al., 2019. "Deep Learning for Audio Signal Processing", IEEE Journal of Selected Topics in Signal Processing.

Fixed parameters in preprocessing stage



Scattering transform



Andén et al., 2011. "Multiscale Scattering for Audio Classification", ISMIR.

Song et al., 2018. "Music auto-tagging using deep Recurrent Neural Networks", Neurocomputing.

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Frame-level waveform based model



Dieleman and Schrauwen, 2014. "End-to-end learning for music audio", ICASSP.

Multi-scale approaches



Zhu et al., 2016. "Learning multiscale features directly from waveforms", Interspeech.

SincNet



Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", IEEE Spoken Language Technology Workshop (SLT).

SincNet



Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", IEEE Spoken Language Technology Workshop (SLT).

SincNet



Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", IEEE Spoken Language Technology Workshop (SLT).

HarmonicCNN

Harmonic embedding

features are stacked at



Won et al., 2019. "Automatic music tagging with harmonic CNN", ISMIR LBD.

HarmonicCNN

Feature engineering (MFCC + Classifier)



Won et al., 2019. "Automatic music tagging with harmonic CNN", ISMIR LBD.

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Connect 1D CNN model to sample-level model

by reducing the sizes of window/filter and hop/stride, and, accordingly, increasing the number of convolutional blocks.



Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", SMC.

Connect 1D CNN model to sample-level model

by reducing the sizes of window/filter and hop/stride, and, accordingly, increasing the number of convolutional blocks.



Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", SMC.

Connect 1D CNN model to sample-level model



Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", IEEE Journal of Selected Topics in Signal Processing.

Sample-level waveform based model



Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", SMC.

More building blocks



- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Learned filter visualization

First convolutional layer filter of frame-level waveform model



Dieleman and Schrauwen, 2014. "End-to-end learning for music audio", ICASSP.

Ardila et al., 2016. "Audio deepdream: Optimizing raw audio with convolutional networks", ISMIR LBD.
Activation Maximization



- 1. Generate random waveform
- 2. Feed-forward to the target filter
- 3. Obtain the gradient of the input layer (waveform)
- 4. Update waveform
- 5. Repeat

Image borrowed from "https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html"

Erhan et al., 2009. "Visualizing higher-layer features of a deep network", University of Montreal.



Layer 1	n an an taini sa na sa na Taini sa na sa n		n panananan Turi pananan	l nevjeta kon sklavne Police se kon svetevne		
Layer 2			4MAN ANA		***	lah-Al-Lahapatela
Layer 3	which me all have		-Malika-Mil	in the second	1411111-144-	MARA AN AN AND
Layer 4	Halparthere	with worth // Hithy	heinderheinden	mathematility	handillanariishha	sill provide
Layer 5	white and the	winter the states	-	www.	Withmonth	Anthemakin
Layer 6	alladelination	www.uniteday.	WHAT AND A WATCH	manashert. April	alisianaliseatile	ninihummah

Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", SMC.





Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", IEEE Journal of Selected Topics in Signal Processing.



Channel index

Extended Architecture

Conv1D

BatchNorm

MaxPool

 $C \times T$

Scale + C×T

relu

GloAvgPool

FC

FC

sigmoid C×1

 $C \times 1$





Excitation Analysis



Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", IEEE Journal of Selected Topics in Signal Processing.

Excitation Analysis



x-axis : loudness (root mean square)

Excitation Analysis



Outline

- Problem statement
- Audio classification models
 - Engineered feature based model
 - Frame-level waveform based model
 - Sample-level waveform based model
- Connect signal processing and deep learning
- Summary

Summary

- Music classification
- Audio classification models
- Waveform based model analysis



References

END-TO-END LEARNING FOR MUSIC AUDIO TAGGING AT SCALE Jordi Pons*1 Oriol Nieto1 Matthew Prockup1 Erik Schmidt1 Andreas Ehmann1 Xavier Serra* * Music Technology Group, Universitat Pompeu Fabra, Barcelona Pandora Media Inc., Oakland, CA

ABSTRACT

age dow wavelorm-based models outper-based ones in large-scale data scenarios.

1. INTRODUCTION

ongs. Provided that a sizable amount of data is avail-

able for that study, we investigate the learning capabili-

tics of these two architectures. Specifically, we investigate whether the architectures based on domain knowledge overly constrain the solution space for cases where large The lack of data tends to limit the outcomes of deep learning research, particularly when dealing with end-to-end learning tacks processing raw data such as waveforms. In this study, 12M tracks annotated with musical latraining data are available - in essence, we study if certain architectural choices (e.g., using log-mel spectrograms as bels are available to train our end-to-end models. This large amount of data allows us to unrestrictedly explore two different design paradigms for music anto-tagging assumption-free models – using waveforms as input with no model assumptions are required for music auto-tarring when operating with large amounts of data. Section 2 discusses the main deep architectures we identified in the audio literature, section 3 describes the datasets used for this work, section 4 presents the architecvery small convolutional filters: and models that rely on tures we study, and section 5 provides discussion about the wo types of deep architectures perform when dataset of variable size are available for training: the MarnaTa-2. CURRENT DEEP ARCHITECTURES gATune (25k songs), the Million Song Dataset (240k songs), and a private dataset of 1.2M songs. Our experi-In order to facilitate the discussion around the current au ments suggest that music domain assumptions are relevant dio architectures, we divide deep learning models into two parts: front-end and back-end - see Figure 1. The front when not encough training data are available, thus showparts: front-end and back-end - see Figure 1. The front-end is the part of the model that interacts with the input signal in order to map it into a latent-space, and the backbased models outperform spectrogramend predicts the output given the representation obtained by the front-end. In the following, we present the main

front- and back-ends we identified in the literatur One fundamental goal in music informatics research is to One fundamental goar in music informance research is to automatically structure large music collections. The music and/o tagging task consists of automatically estimating the Input - Front-end - Back-end - Output musical attributes of a song - including: moods, languag Figure 1 Deep learning nineline Front-ends. These are generally comprised of con-volutional neural networks (CNNs) [5, 9, 20, 21, 27]. automatically organizing musical libraries. since these can learn efficient representations by sharing Many approaches have been considered for this task (mostly based on feature extraction + model [1, 22, 26]), with recent publications showing promising results using weights' along the signal. Front-ends can be drouded into two groups depending on the used input signal: wave-forms [9, 14, 27] or spectrograms [5, 20, 21]. Further, the design of the filters can be either based on domain knowldeep architectures [5,9, 14, 21]. In this work we confirm this trend by studying how two deep architectures conedge or not. For example, one leverages domain knowledge when a front-end for waveforms is designed so that edge when a front-end for wavetorms is designed so that the length of the filter is set to be as the window length of a STFT [9]. Or for a spectrogram front-end, it is used vertione of the datasets being of an unprecedented size: 1.2M

filters to learn longer temporal cues [25]. Generally, a sin-gle filter shape is used in the first CNN layer [5, 9, 12, 25]. but some recent works report performance gains when us Schmidt, Andreas Human, Xevier Serra, Licemed under a Creative ing several filter shapes in the first layer [4, 18-2], 27] Example (g) deal runs, Oraz sens, annue recong, tan Schnill, Advess Binnan, Nerez Fors, Larono (unie a Deatrie Common Attribution 4-B International Licence (ICE DY 40). Marthue-fiem: José Poux, Onel Nato, Marthue Prockog, Fais Schnild, Advess Einman, Xavier Sens, "EDD TO EDD LEARNING FOR MUSIC AU-DIO TMORING AT SCALE", 19th International Society for Music Infor-mation-Retrievel Conference, Phys. Pance, 2018. Using many filters promotes a richer feature extraction in the first layer, and facilitates levenging domain knowl-edge for designing the filters' shape. For example: a

cal filters to learn timbral representations [12] or horizonta

REE RURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. VOL. 13. NO. 2. MAY 2019

Comparison and Analysis of SampleCNN Architectures for Audio Classification

Taeiun Kim[®] Ionenil Lee[®] and Juhan Nam[®] Member IEEE

After a Cardia and sensing with sensing the sensitivity of the sensit Abstract-End-to-end learning with convolutional neural netinterventing class provide with increasing due to taken in the second material second material

IndexTerms—Audio dassification, end-to-endfaraming, convolu-tional neural networks, residual networks, squeeze-and-excitation tilterhank when the network was trained with a large amount networks, interpretability.

I. INTRODUCTION

ONVOLUTIONAL Neural Networks (CNNs) have They found out that CNNs using waveform input can be compet moven highly effective in classification tasks where the input is high-dimensional sensory data such as image or audio. In layer is sufficiently deep. Other end-to-end approaches for variimage classification, the end-to-end learning that takes raw eix- ous audio recognition tasks are found in many literatures [25]els of images directly as inputs of the CNNs has become a stan-[28]. All of these previous works used a large size of filters in dard approach [1]-[4]. In audio classification, on the other hand, the first convolutional laver, which typically takes several humthe majority of CNN-based models still use spectrogram-based inputs such as mel-spectrograms that involve significant hand-the filters should learn all possible phase variations within the crafted design in the time-frequency transform. Thus, depending on the domains in audio classification tasks, researchers have to grams with the corresponding waveforms does not improve

Mannoist method Adaré 3, 2014, aviad Fohrang 15, 2019, souring dimensional protein aviant of a source of the sourc Task Welevernin as Hiples and next system better when the first convo-ing the system of Science and Technology, Dagion 31141, Sonh Korea Insigne of Science and Technology, Dagion 31141, Sonh Korea vanced Insidule of science and reconsergy, tracpets seven, soom or mail: topin/0 kaist.ac.kr, rechter@kaist.ac.kr, juhannam@kaist.ac.kr). Dioial.Ohioet.Ldogtdier.10.1109/JSTSP.2019.2909429

performance in plain CNNs [22]. To train CNNs with wave raw wave forms as inputs and has very small sizes of filters [29]. rather than a typical frame-level size (e.g., 256 or 512 samples)

helped reducing phase variation in the time-domain convolution. Dai et al. used Deep Convolutional Neural Networks (DCCNN)

with skip connections for environmental sound recognition [24].

1932-4553 () 2019 IEEE. Personal use is permitted, but republication/redistribu

Deep Learning for Audio Signal Processing Hendrik Purwins⁰, Bo Li⁰, Tuomas Virtanen⁰, Jan Schlüter⁰, Shuo-Yiin Chang, and Tara Sainath⁰

semently, prominent deep learning application areas are covered. ar, audio recognition (automatic speech recognition, music infor-mation retrieval, environmental sound detection, localization and cking) and synthesis and transformation (source separation, autracking) and synthesis and transformation (source separation, au-dio enhancement, generative models for speech, sound, and music synthesis). Finally, key issues and future questions regarding deep

ed to audio signal processing are identified. Inlex Terms—Deep learning, connectionist temporal memory, automatic speech recognition, music information retrieval, source swaration, audio enhancement, ovironmental sounds.

I. INTRODUCTION

RTIFICIAL neural networks have gained widespread often analyzed as a whole or in patches with little order con-A attention in three wayes so far, triggered by 1) the percen-[2] in 1986, and finally 3) the success of deep learning in speech solutions. recognition [3] and image classification [4] in 2012, leading to

a renaissance of deep learning, involving e.g., deep feedforward neural networks [3]. [5], convolutional neural networks (CNNs, [6]) and long short-term memory (LSTM, [7]). In this "deep" paradigm, architectures with a large number of parameters are advances in machine parallelism (e.g., cloud computing, GPUs or TPUs [148]). The recent surge in interest in deep learning

nift@google.com. Z. Vistanen is with Tampere University, Tampere FE-33100, Finland different combinations considered.

Schäter is with tampere University, tampere PO3100, Pinano (e-mail: Isomas vietanen@tuni.8).
Schäter is with Université de Toulon. Aix Marseille Univ. CNRS, LIS, DYNI team, Marseille, France and Austrian Research Institute for Artificial

Intelligence, Vienna 1010, Austria (e-mail: jan.schlaeter@ofai.at). Digital Object Identifier 10.1109/JSTSP2019.2908700

Januer - Genes in erun sing is he solutionna de de jane. In a cabida parcial or président in sura san est de solution de la president de la pr representations (in particular, lag-and spectra and raw workers) as work as innervous kannom rakes focus as genomes, square and deep kerning models are reviewed, including constantions of the long short-term memory architector neural networks, variants of the long short-term memory architector term, as well as models are reviewed, school and the long short-term memory architector term, as well as models are reviewed as the long short-term memory architector term, as well as models are reviewed as the long short-term memory architector term as well as models are reviewed as the long short-term memory architector term as well as models are reviewed as the long short-term memory architector term are short and term models. In this short are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term memory architector term are short models are reviewed as the long short-term are reviewed as the long sh hidden Markov models and non-negative matrix factorization applications where sufficient data is available

> image processing, there are important differences between the domains that warrant a specific look at audio. Raw audio samples form a one-dimensional time series signal, which is fundament tally different from two-dimensional images. Audio signals are representations for processing, but the two axes, time and fre quency, are not homogeneous as horizontal and vertical axes in an image. Images are instantaneous snapshots of a target and

porithm [1] in 1957, 2) the backpropagation algorithm chronological order. These properties gave rise to audio-specific

To set the stage, we give a conceptual overview of audio analysis and synthesis problems (II-A), the input representation

They Analise to Equination Activities, leeging an attain team in the analysis of the angle is a start of the input length. Secondly, in Lis. 5: y. Chang, and T. Sisanth are with Gorgeli (n), Mustim were, CA start ULSA is small holding operations, and asymptopic sec. and hilder and the angle of the lass as set of classes, or a summer's analysis of the start of the lasses of the lasses. value. In the following, we will name and give examples for the ¹While the audio signal will often be processed into a sequence of features, we consider this part of the solution, not of the task.

1932-4553 © 2019 IEEE. Presonal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications.atandards/publications/rights/index.html for more information.



Juhan Nam, Keurawaa Chai, Janapil Lee

deploying the Music Genome Project [1] (https://www.pandora om/about/men) based on human-annotated content analysis nother company, Spotify, has a catalog of over 40 million ones and over 180 million users as of mid-2018 (https://rees ify.com/us/about/), making it a leading music service pro der unddwide. Giant technology companies such as Angle service platforms. Furthermore, artificial intelligence speakers rs with a new and easily accessible way to listen to music While music-streaming services have made a huge volme of music accessible to users, the enormous size of the ervice catalogs has created the challenge of finding among so many choices the songs that fit users' tastes. A general approach to this issue has been collaborative filtering, which data, such as play history and song rating. Although collaboutive fikering effectively retrieves songs and accommodates personalized recommendations, its performance is hampered by such issues as popularity bias and the cold-start problem the challenge of recommending new music to users [2]. The tent-based approach is often regarded as a supplementar solution to those problems. Pandora radio is a representative sample as it retrieves songs by exploiting the similarities of song descriptors, such as genre, mood, instruments, and vocal lity. However, high-quality manual annotation is costly and not scalable, suggesting a need for better ways to automate lassification of music content. As a result, much attention in the field of music information retrieval (MIR) over the last few years has centered on finding ways to automate the process of classifying music genre and mood and tagging music. Hereaf er, this article will use the term music classification and tag nine as a general expression for tasks that involve taking musiaudio data as input and automatically annotating them with a certain form of semantic label.

HEE SOME PROCESSING INCODING | Income 2019

Deep Learning for Audio-Based Music

Teaching computers to distinguish rock from Bach

Classification and Tagaing

er the last decade, music-streaming services have grown

ramatically. Pandora, one company in the field, has pio-

ered and popularized streaming music by successfull

11

Music Source Separation

Time to explore end-to-end learning

Why end-to-end music source separation?



Why end-to-end music source separation?



I) Are we missing crucial information when discarding the phase?

II) When using the **phase of the mixture at synthesis time**, are we introducing artifacts that are limiting our model's performance?

Why filtering spectrograms with masks?



III) Filtering spectrograms does not allow recovering masked (perceptually hidden sound) signals

Figure from: Jansson et al., 2017. "Singing voice separation with deep U-Net convolutional networks" in ISMIR.



I) Are we missing crucial information when **discarding the phase**?

- II) When using the **phase of the mixture at synthesis time**, are we introducing artifacts that are limiting our model's performance?
- III) Filtering spectrograms does not allow recovering masked signals

End-to-end music source separation



Other (active) research directions: Use the complex STFT as i/o interface?

Kameoka et al., 2009. "ComplexNMF: A new sparse representation for acoustic signals" in ICASSP.

Dubey et al., 2017. "Does phase matter for monaural source separation?" in arXiv.

Le Roux et al., 2019. "Phasebook and friends: Leveraging discrete representations for source separation" in IEEE Journal of Selected Topics in Signal Processing.

Tan et al., 2019. "Complex Spectral Mapping with a CRNN for Monaural Speech Enhancement" in ICASSP.

Liu et al., 2019. "Supervised Speech Enhancement with Real Spectrum Approximation" in ICASSP.

Other (active) research directions: Alternative models at synthesis time?

Virtanen and Klapuri, 2000. "Separation of harmonic sound sources using sinusoidal modeling," in ICASSP.

Chandna et al., 2019. "A vocoder based method for singing voice extraction" in ICASSP.

End-to-end music source separation



Historical perspective: unsupervised & linear models



Linear model example



Unsupervised factorization of the mixture into **bases** (w) and **activations** (h)

Historical perspective: unsupervised & linear models







waveform-based ICA



Problem 1: phase sensitive basis Problem 2: simplicity of the linear model

Figure from: Blumensath and Davies, 2004. "Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music," in ICASSP.





NMF cannot be used with waveforms due to its non-negative constraint! (waveforms range from -1 to 1)



A widely-used set of tools:

filtering spectrograms

linear models

unsupervised learning

audio domain knowledge

...maybe we could try another toolset?

filtering \rightarrow synthesis?

linear models \rightarrow non-linear models?

unsupervised learning \rightarrow supervised learning?

audio domain knowledge \rightarrow data driven?

End-to-end music source separation: 9 publications

Stoller et al., 2018. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation" in arXiv.

Grais et al., 2018. "Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders" in EUSIPCO.

Lluis, et al., 2018. "End-to-end music source separation: is it possible in the waveform domain?" in arXiv.

Slizovskaia et al., 2018. "End-to-end Sound Source Separation Conditioned on Instrument Labels" in arXiv.

Cohen-Hadria et al., 2019. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation" in arXiv.

Kaspersen, 2019. "HydraNet: A Network For Singing Voice Separation". Master Thesis.

Akhmetov et al., 2019. "Time Domain Source Separation with Spectral Penalties". Technical Report.

Défossez et al., 2019. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed" in arXiv.

Narayanaswamy et al., 2019. "Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets" in arXiv.

ALL THE PUBLICATIONS IN CHRONOLOGICAL ORDER AS OF OCTOBER 2019

End-to-end music source separation: architectures



Introduction: the "generative" Wavenet



Van den Oord et al., 2016. "Wavenet: a generative model for audio" in arXiv.

A "regression" Wavenet for music source separation



Lluis, et al., 2019. "End-to-end music source separation: is it possible in the waveform domain?" in Interspeech.

Fully convolutional & deterministic



Lluis, et al., 2019. "End-to-end music source separation: is it possible in the waveform domain?" in Interspeech.
Fully convolutional & deterministic



Real time inference!

1601 samples input \rightarrow denoising time: \approx 0.56 sec per second of music on GPU!

Lluis, et al., 2019. "End-to-end music source separation: is it possible in the waveform domain?" in Interspeech.

End-to-end music source separation: architectures



Autoencoders



Multi-resolution & Convolutional autoencoder





Multi-resolution CNN: efficient way to represent 3 periods!

Multi-resolution CNN = Inception CNN (different filter shapes in the same CNN layer)

Grais et al., 2018. "Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders" in EUSIPCO.

Multi-resolution & Convolutional autoencoder



Grais et al., 2018. "Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders" in EUSIPCO.

End-to-end music source separation: architectures



Wave-U-net



Stoller et al., 2018. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation" in arXiv.

Wave-u-net extensions

• Multiplicative conditioning using instrument labels at the bottleneck.

Slizovskaia et al., 2019. "End-to-end Sound Source Separation Conditioned on Instrument Labels" in ICASSP.

• Data augmentation: \approx 1 dB SDR improvement.

Cohen-Hadria et al., 2019. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation" in arXiv.

Data augmentation strategies

It is used to artificially expand the size of a training dataset by creating modified versions of it.

- Random swapping left/right channel for each source
- Random scaling sources
- Random mixing of sources from different songs
- Pitch-shifting
- Time-stretching

Uhlich et al, 2017. "Improving music source separation based on deep neural networks through data augmentation and network blending" in ICASSP.

Cohen-Hadria et al., 2019. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation" in arXiv.

Wave-u-net extensions

• Multiplicative conditioning using instrument labels at the bottleneck.

Slizovskaia et al., 2019. "End-to-end Sound Source Separation Conditioned on Instrument Labels" in ICASSP.

• Data augmentation: \approx 1 dB SDR improvement.

Cohen-Hadria et al., 2019. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation" in arXiv.

• Add BiLSTMs at the bottleneck: \approx 1 dB SDR improvement.

Kaspersen, 2019. "HydraNet: A Network For Singing Voice Separation". Master Thesis.

• Loss function in the spectral domain.

Akhmetov et al., 2019. "Time Domain Source Separation with Spectral Penalties". Technical Report.

• Use dilated convolutions and dense CNNs.

Narayanaswamy et al., 2019. "Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets" in arXiv.

• Achieve comparable results to a spectrogram-based model: Demucs.

w/ BiLSTMs at the bottleneck, data augmentation, and some additional architectural changes: ≈ 1.6 dB SDR improvement. Défossez et al., 2019. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed" in arXiv.

Wave-u-net extensions: Demucs



Wave-u-net extensions: Demucs



Wave-u-net extensions: Demucs



Wave-u-net extensions: Wave-U-net vs. Demucs



Wave-U-net: building blocks

Demucs: building blocks

Wave-u-net extensions: Wave-U-net vs. Demucs



Wave-U-net: building blocks

Demucs: building blocks

Deconvolutions and high-frequency artifacts



Checkerboard artifacts in images



High-frequency buzz in audio

Odena et al., 2016. "Deconvolution and Checkerboard Artifacts" in Distill.

Comparison: a perceptual study

Lluis, et al., 2019. "End-to-end music source separation: is it possible in the waveform domain?" in Interspeech.



Evaluation metrics: SDR, SIR, SAR

$$SDR := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}$$

"overall performance"

$$\text{SIR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}$$

"interference from other sources"

$$SAR := 10 \log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2}$$
 "algorithmic artifacts"

http://craffel.github.io/mir_eval/

https://github.com/sigsep/sigsep-mus-eval/

Vincent et al., 2006. "Performance measurement in blind audio source separation" in IEEE TASLP.

Comparison: a perceptual study

Lluis, et al., 2019. "End-to-end music source separation: is it possible in the waveform domain?" in Interspeech.



MOS	Wavenet-based	Wave-U-Net
Vocals	3.0 ± 1.0	3.3 ± 0.85

(t-test: p-value=0.049, 15 participants)

Additional references







Estefanía Cano, Derry FitzGerald, Antoine Liutkus, Mark D. Plumbley, and Fabian-Robert Stöter



End-to-end music source separation: is it possible in the waveform domain?

Francesc Lluís* Jordi Pons* Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona. name.surname@upf.edu

Abstract

Most of the currently successful source separation techniques use the magnitude spectrogram as input, and are therefore by default omitting part of the signal: the phase. To avoid omitting potentially useful information, we study the viability of using end-to-end models for music source separation - which take into account all the information available in the raw audio signal, including the phase. Although during the last decades end-to-end music source separation has been considered almost unattainable, our results confirm that waveform-based models can perform similarly (if not better) than a spectrogram-based deep learning model. Namely: a Wavenet-based model we propose and Wave-U-Net can outperform DeepConvSep, a recent ctrogram-based deep learning model. Index Terms: source separation, end-to-end learning.

1. Introduction

20

Jun

28

SD]

CS

N

87

121

0

v:1810

Xi

When two or more sounds co-exist, they interfere with each other resulting in a novel mixture signal where sounds are superposed (and, sometimes, masked). The source separation task tackles the inverse problem of recovering each individual sound source contribution from an observed mixture signal. With the recent advances in deep learning, source separa-

tion techniques have improved substantially [1]. Interestingly, though, nearly all successful deep learning algorithms use the magnitude spectrogram as input [1, 2, 3] - and are therefore, by default, omitting part of the signal: the phase. Omitting the potentially useful information of the phase entails the risk of finding a sub-optimal solution. In this work, we aim to take full advantage of the acoustic modeling capabilities of deep learning to investigate whether it is possible to approach the problem of music source separation directly in an end-to-end learning fashion. Consequently, our investigation is centered on studying how to separate music sources (e.g., singing voice, bass or drums) directly from the raw waveform music mixture. During the last two decades, matrix decomposition methods have dominated the field of audio source separation. Sev-

eral algorithms have been proposed throughout the years, with independent component analysis (ICA) [4], sparse coding [5], or non-negative matrix factorization (NMF) [6] being the most used ones. Given that magnitude or power spectrogram representations are always non-negative, imposing a non-negative constraint (like in NMF) is particularly useful when analyzing these spectrograms - but less appropriate for processing waveforms, which range from -1 to 1. For that reason, methods like ICA and snarse coding have historically been used to process waveforms [7, 8, 9]. Waveform representations preserve all the information available in the raw signal. How ever, given the unpredictable behavior of the phase in real-life

*Contributed equally

introduce an additional source of error [10]. Notably, many modern spectrogram-based deep learning models are also relying on this same (potentially problematic) approach [2, 12]. To overcome this issue, some tried to consider the phase when

separating the sources [13, 14, 15]2, or some others relied on a sinusoidal signal model at synthesis time [16]. However, in our work, we do not want to rely on any time-frequency transform or any signal model. Instead, we aim to directly approach the problem in the waveform domain.

sounds, it is rare to find identical waveforms produced by the same sound source. As a result of this variability, a single

basis1 cannot represent a sound source and therefore, one re-

quires i) a large amount of bases, or ii) shift-invariant bases to

obtain accurate decompositions [8, 10]. Although several ma-

trix decomposition methods have been used for decomposing

waveform-based mixtures [7, 8, 9], these have never worked

Due to the above mentioned difficulties, the phase of com-

plex time-frequency representations is commonly discarded as-

suming that magnitude spectrograms already carry meaningful

information about the sound sources to be separated. Phase re-

lated problems disappear when sounds are just represented as

magnitude or power spectrograms, since different realizations

of the same sound are almost identical in this time-frequency

plane. This allows to easily overcome the variability problem

assuming that sources add linearly in the time domain [10]1.

Most matrix decomposition methods rely on a signal mode

However, the addition of signals in the time and frequency

domains is not equivalent if phases are discarded. Only in ex-

pectation: $E\{|X(k)|^2\} = |Y_1(k)|^2 + |Y_2(k)|^2$, where X(k) =

 $DFT{x(t)}$. This means that we can approximate the time-

domain summation in the power spectral domain. For that rea-

son, many approaches utilize power spectrograms as inputs Although magnitude spectrograms work well in practice [11],

there is no similar theoretical justification for such an inconsis-

trograms still need to deliver a waveform signal. To this end, the

main practice is to filter the original magnitude or power spec-

trogram with (predicted) time-frequency masks. Accordingly,

the original noisy phase of the mixture is used when synthe-

sizing the waveform of the estimated sources - which might

Finally, note that these methods operating on top of spec

tency with the signal model when the phases are discarded.

as well as the spectrogram-based ones.

found when operating with waveforms.

As seen, many issues still exist around the idea of discard ing the phase: are we missing crucial information when discarding it? When using the phase of the mixture at synthesis time, are we introducing artifacts that are limiting our model's performance? Or, since magnitude spectrograms (differently from

1ICA, snarse coding & NMF model the mixture signal as a weighted sum of bases, which represent a source or components of a source. ²Using the full complex STFT number, instead of utilizing phaseless representations (either at the input or when applying the masks).

S. Dubnov, 2002. "Extracting sound objects by independent subspace analysis" in AES Conference.

Blumensath and Davies, 2004. "Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music," in ICASSP.

Jang and Lee, 2003. "A maximum likelihood approach to single-channel source separation" in Journal of Machine Learning Research.

Music Generation

Expressive music modeling

Overview

Why audio? Why raw audio?

Generative models

Likelihood-based models of raw audio

Adversarial models of raw audio

Summary

Why audio? Why raw audio?

Why audio?

Music generation is typically studied in the symbolic domain







Why audio?

Many instruments have complex action spaces Rich palette of sounds and timbral variations

Guitar

- pick vs. finger
- picking position
- frets
- harmonics
- ...





Magnitude spectrogram



Phase spectrogram

Phase is often unimportant in discriminative settings, but is very important perceptually!





original phase

random phase

Phase is hard to model:

- it is an angle, so it wraps around
- it becomes random as the magnitude tends to 0
- absolute phase is less meaningful, but relative phase differences matter



What is "raw audio" anyway?



Discretising audio

- Time
- Amplitude







Generative models

Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**
Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**

Explicit: given $x \in X$, model can infer p(x)

Implicit: model can produce new samples x ~ p(X)

Generative models



Brock et al., 2019. "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR.

Likelihood-based models

Likelihood-based models parameterise **p(X)** directly

Objective function: maximise $\sum_{x \in X} \log p(x)$

Autoregressive models

Autoregressive models factorise **p(X)** into simpler (scalar) distributions

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n)$$

 $\mathbf{p}(\mathbf{x}) = \prod_{i} \mathbf{p}(\mathbf{x}_{i} | \mathbf{x}_{< i})$ chain rule of probability

We can use the same model $\mathbf{p}(\mathbf{x}_{i}|\mathbf{x}_{<i})$ for all i!

Flow-based models

Flow-based models transform **p(X)** to a simple (factorised) distribution with an invertible mapping

$$p(x) = p(z) \cdot |det J|^{-1}$$
 change of variables theorem
 $J = d(g(z))/dz$ $x = g(z)$

Important constraints:

g(z) must be invertible det J must be tractable

Dinh et al., 2014. "NICE: Non-linear Independent Components Estimation", arXiv. Dinh et al., 2016. "Density estimation using Real NVP", arXiv.

Variational autoencoders (VAEs)



VAE maps latents **z** from a simple distribution to **x** with a (non-invertible) generative network

The inference network approximates the inverse operation

p(x) cannot be computed exactly, the Evidence Lower BOund (ELBO) is maximised instead

Kingma & Welling, 2014. "Auto-Encoding Variational Bayes", ICLR 2014.

Rezende et al., 2014. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models", ICML 2014.

Adversarial models



 $\mathcal{L} = \mathbb{E}_{x}[\log(D(x))] + \mathbb{E}_{z}[\log(1 - D(G(z)))]$

Goodfellow et al., 2014. "Generative Adversarial Nets", NeurIPS.

More exotic flavours

- Implicit quantile networks
- Energy-based models
- Optimal transport (e.g. Wasserstein autoencoders)
- Score-based generative modelling



Conditional generative models

Conditioning is "side information" which allows for control over the model output

p(x|c) vs. **p(x)**

Conditional generative models

Conditioning is "side information" which allows for control over the model output

p(x|c) vs. **p(x)**



Conditional generative models

Conditioning is "side information" which allows for control over the model output

p(x|c) vs. **p(x)**

	sparsely conditioned			densely conditioned		
music audio models	composer instrument(s) tempo timbre 	note density musical form	score	MIDI	other audio signals	
			, Mileteszzene Mileteszte Mi			



- Likelihood-based models are mode-covering
- Adversarial models are (typically) mode-seeking
- In more densely conditioned settings, we tend to care less about covering all the modes

Likelihood-based models

WaveNet



WaveNet



WaveNet: dilated convolutions



WaveNet: dilated convolutions



















SampleRNN



Mehri et al., 2017. "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model", ICLR.

Parallel WaveNet, ClariNet



van den Oord et al., 2019. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis", ICML.

WaveGlow, FloWaveNet



Prenger et al., 2019. "Waveglow: A flow-based generative network for speech synthesis", ICASSP. Kim et al., 2019. "FloWaveNet: A generative flow for raw audio", ICML.



... but memory usage ~ receptive field length

Required model depth is **logarithmic** in the desired receptive field length

Required memory usage during training is still **linear** in the desired receptive field length!

⇒ We cannot scale indefinitely using dilation

Autoregressive discrete autoencoders decoder y code ... 5 207 131 18 168 ... modulator local model q' = VQ(q)quantised query quantisation q = f(x) query encoder input Х

van den Oord et al., 2017. "Neural discrete representation learning", NeurIPS.



Dieleman et al., 2018. "The challenge of realistic music generation: modelling raw audio at scale", NeurIPS.



Wave2Midi2Wave and the MAESTRO dataset



Hawthorne et al., 2019. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset", ICLR.

Sparse transformers

Attention (with sparse masks) instead of recurrence / convolutions.



(a) Transformer



Child et al., 2019. "Generating long sequences with sparse transformers", arXiv.

())

Add & Norm

Output Probabilities

Softmax

Linear

Add & Norm

Feed Forward

Add & Norm

Multi-Head

N×

Positional

Encoding

Universal music translation network



Mor et al., 2019. "A Universal Music Translation Network", ICLR.

http://dadabots.com/

Neural networks

generating death metal via livestream 24/7 to infinity

We make raw audio neural networks that can imitate bands

Carr & Zukowski, 2018. "Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands", arXiv.

Adversarial models

WaveGAN



Donahue et al., 2019. "Adversarial Audio Synthesis", ICLR.

WaveGAN



Non GAN-activated cat

WaveGAN activated cat

SpecGAN activated cat

Donahue et al., 2019. "Adversarial Audio Synthesis", ICLR.



Engel et al., 2019. "GANSynth: Adversarial Neural Audio Synthesis", ICLR.

GANSynth



Engel et al., 2019. "GANSynth: Adversarial Neural Audio Synthesis", ICLR.



Binkowski et al., 2019. "High Fidelity Speech Synthesis with Adversarial Networks", arXiv.

MelGAN



Kumar et al., 2019. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis", arXiv.
Most popular generative modelling paradigm:

GANs

Most popular generative modelling paradigm:

GANs

Most popular generative modelling paradigm for music:

likelihood-based (autoregressive)

- We are still figuring out the right architectural priors for audio discriminators
 - For images, a stack of convolutions is all you need
 - What do we need for audio? Multiresolution? Dilation?
 Something phase shift invariant?

- We are still figuring out the right architectural priors for audio discriminators
 - For images, a stack of convolutions is all you need
 - What do we need for audio? Multiresolution? Dilation?
 Something phase shift invariant?
- The sparsely-conditioned setting is dominant
 - We care about "creativity" and capturing diversity
 - GANs are worse at this than likelihood-based models

Ulyanov et al., 2018. "Deep Image Prior", CVPR. Pons et al., 2019. "Randomly weighted CNNs for (music) audio classification", ICASSP.

Alternatives to modelling raw audio directly

- Model complex-valued spectrograms and "deal" with phase (GANSynth)
- Model magnitude spectrograms
 Use a vocoder or Griffin-Lim to invert (Tacotron 1 & 2, MelGAN, MelNet, ...)

Wang et al., 2017. "Tacotron: Towards end-to-end speech synthesis", ISCA. Shen et al., 2018. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", ICASSP. Vasquez & Lewis, 2019. "MelNet: A Generative Model for Audio in the Frequency Domain", arXiv.

Alternatives to modelling raw audio directly

 Differentiable Digital Signal Processing (anonymous authors, ICLR, in review) Use raw audio input, but put DSP components in the model https://openreview.net/forum?id=B1x1ma4tDr



Summary

- Generative modelling of raw audio is feasible, even in the sparsely conditioned setting
- Likelihood-based models dominate, but GANs are making in-roads in the densely conditioned setting
- Modelling large-scale structure from raw audio is an unsolved problem

Thank you

Sander Dieleman, Jordi Pons, Jongpil Lee