XTRACK: AUTOMATIC SEGMENTATION OF LIVE MUSIC WITH YAMNET

Florian COLOMBO (florian.colombo@gmail.com)[1,2], **Alain DUFAUX**[1], and **Davide PICCA**[2]

[1]**Lausanne Federal Institute of Technology (EPFL)**, Lausanne, Switzerland
[2]**University of Lausanne (UNIL)**, Lausanne, Switzerland

ABSTRACT

Recorded in UNESCO's Memory of the World Register, the collection of recordings of Montreux Jazz Festival concerts represents an invaluable audiovisual resource. Expert annotators have segmented the audio recordings to separate the music pieces from live concert events such as applause, speeches, and silences. This task is time-consuming and several rounds are necessary to homogenize the subjective decision of music borders. In this paper, we present XTRACK, an algorithm for automated segmentation of music pieces. It integrates the predictions of the sound classifier network YAMNET to create accurate segments. The algorithm is designed to isolate each track from live audio recordings. Finally, we evaluate the algorithm's performances on the human-labeled Montreux Jazz Festival archives.

1. INTRODUCTION

Recordings of live music are numerous and their number is increasing every day. To publish or archive such recordings, one typically needs to segment them into single music tracks: audio segments containing only music. Indeed, there are many moments without music from the moment the recording starts until it ends. Therefore, segmentation of live recordings aims to have a pleasant listening experience of individual tracks. To do so, student annotators were involved in indexing each live concert recording from the Montreux Jazz Festival (MJF). This constitutes a valuable database of audio segmentation.

In this paper, we present XTRACK, a deep-learning algorithm to assist people in segmenting live music recordings into individual music tracks. The segmentation algorithm employs the pre-trained audio classifier YAMNET [1]. Our contributions are the post-processing of YAMNET predictions to segment audio files in music and non-music segments and the performance evaluation on the MJF indexation data.

1.1 Montreux Jazz Festival audiovisual archives

Recorded in professional quality video and audio since the first edition in 1967, the audiovisual collection of the MJF gathers more than 5,000 concerts, representative of the greatest artists and musical trends of the last 50 years. It was inscribed in the UNESCO's Memory of the World Register in 2013, simultaneously with the creation of the Claude Nobs Foundation which is in charge of its preservation. From 2007, and in the frame of the so-called Montreux Jazz Digital Project [1] , EPFL has been responsible for inventory, digitization, cataloging, storage, and enrichment of the collection. The Cultural Heritage and Innovation center of EPFL (CHC) was created for that goal, gathering all data and metadata in a dedicated relational database and defining numerous research, education, and innovation projects to enrich the collection.

1.2 Dataset: Cataloging and Indexing

The dataset represents more than 11'000 hours of video and 6000 hours of audio, digitized in uncompressed formats or directly captured to professional broadcast files in recent years. The sample rate of the digital audio stream is 96 kHz (24 bits precision) when digitized from analog support, but it can be 96, 48, or 44.1 kHz (16 or 24 bits) when captured in a native way from digital tapes. The associated Metadata consists of both technical information obtained from digitization (including hash and media information exports) and concerts' information (concert name, title of the pieces, hall, date, artists' names, associated instruments, authors' rights, original tape format)

Since 2013, a large project has been developed to catalog and index every concert, identifying series of events (introduction, song, speech, interlude, applause, encore, or silence), indexing precisely their respective start/end time locations in the media files, and verifying concert metadata (titles, artists, hall, year, etc). The motivation for this work is to be able to navigate comfortably across the 50'000 songs of the live concerts' collection, being able to properly and precisely playback them like in CDs or DVDs. Consequently, high precision is required in the determination of the start and end time indexes of every event. Indexing is performed by specially trained people, who need extensive experience of the multiple scenarios that can occur within song transitions. Such scenarios include

- The end of the song is sudden, a 1-2 seconds silence time occurs before the public starts applauding
- The end of the song can be easily identified and predicted by the public, resulting in a mix of musical song fade-out and applause fade-in period, followed 5-10 seconds later by a few words from a musician, speaking for a while about his last song

[1] https://go.epfl.ch/mjdp

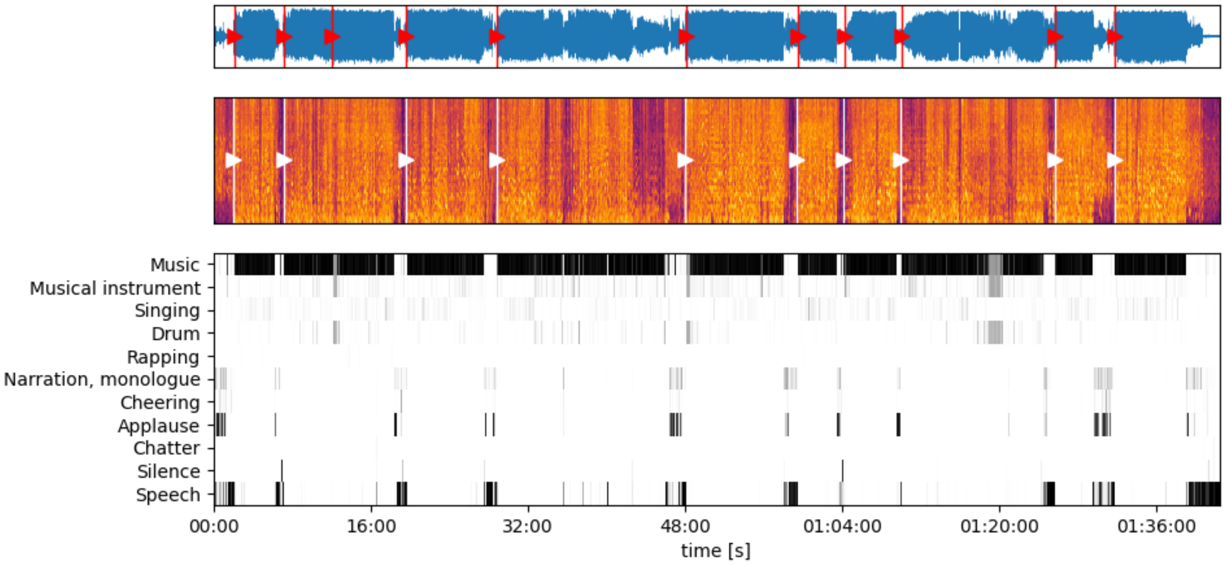

Figure 1. On top, a plot of the waveform from a live concert recording with track start labels from the MJF archives shown as vertical bars. In the middle, the frequency decomposition of the waveform with the XTRACK prediction for the starts of music segments shown as vertical bars. On the bottom, the normalized YAMNET predictions for the 5 music classes and 7 non-music classes.

- Same situation but with the musician speaking immediately at the very beginning of the applause, saying *Thank You!* and possibly presenting the musicians while some music goes on. Then the next song is chained
- The next song starts immediately during the fade-in of the public applause, which then rapidly stops. No speech from the band.
- No transition between two songs, a new one starts suddenly in the previous one, and the audience applauds for a short time while the new song starts.
- The song gets quiet and applause starts but a solo moves on. Applause fades out and stops rapidly, to start again over the music at the end of the solo. The music goes on for a while before its real completion.

Where to place the end and start time codes of every song? Criteria were defined for usual situations but the question is difficult and decisions can often be subjective. The indexing work requires extensive training before it can be performed reliably. Consequently, a detailed workflow was introduced with two phases, involving different documentalists, followed by mixed manual and automated quality checks, and weekly team discussions for solving difficult or uncertain situations.

2. RELATED WORKS

Trained on Google's AudioSet data comprising over 2 million annotated audio clip [2], YAMNET is a deep convolutional neural network based on the MobileNets architecture [3]. In particular, YAMNET processes audio files (16kHz sampling rate) and outputs at every 48ms frame a probability distribution over 521 classes (e.g.`speech`, `happy music`, `toothbrush`, ...).

In 2013, [4] proposed to automate part of the MJF indexation work by detecting occurrences of applause. Of course, applause does not correspond to song transitions in all cases but the possibility of a semi-automated approach was considered for the indexing work. New AI solutions such as YAMNET are promising, and the question is raised again. This time to fully try and mimic human work, taking thousands of hours of indexed concerts as a benchmark dataset, a new study was launched, and the first outcomes are described in this paper.

3. METHODS AND DATA

Whereas the MJF archives' indexation follows a well-defined protocol, many boundary decisions are subjective and do not directly correspond to audio events. For example, the end of music segments does not happen at the end of music but rather some flexible times later (from 0 to 8 seconds depending on the context and the subjective decisions of the human annotators). From a deep learning perspective, these observations directed us not to train a model directly on these subjective (inaccurate) data. Rather, we inquired how the audio classifier YAMNET [1] can be used for segmenting audio files and take advantage of the MJF data to validate the boundary decisions of XTRACK.

3.1 From YAMNET predictions to audio segments

We start by processing the audio of live concert recordings with YAMNET. To adapt YAMNET to our task, we constrain the probability distributions over the complete set of 521 classes to a smaller set of meaningful classes. XTRACK focuses on the 5 music classes and 6 non-music classes listed in Table 1. The temporal evolution of the

Description	Equation	Value
YAMNet music classes	C_{music}	[Music, Musical instrument, Singing, Drum, Rapping]
YAMNet non-music classes	$C_{non-music}$	[Silence, Speech, Narration, monologue, Chatter, Cheering, Applause]
Probability threshold to classify the nth frame of audio file x as music	if $\sum_{c \in C_{music}} Pr(x[n] = c) > \Theta_{music}$ then $n \rightarrow$ music	$\Theta_{music} = 0.2$
Probability threshold to classify the nth frame as non-music	if $\sum_{c \in C_{music}} Pr(x[n] = c) < \Theta_{non-music}$ then $n \rightarrow$ non-music	$\Theta_{non-music} = 0.1$
Consecutive music frames to place a music start index	if $m \in [n - L_{music} : n] \rightarrow$ music then $n \rightarrow$ start	$L_{music} = 20[s]$
Consecutive non-music frames to place a stop index	if $m \in [n - L_{non-music} : n] \rightarrow$ non-music then $n \rightarrow$ stop	$L_{non-music} = 4[s]$
Music start offset	start$[i] \rightarrow$ start$[i] + d_{start}$	$d_{start} = -0.1[s]$
Music stop offset	stop$[i] \rightarrow$ stop$[i] + d_{stop}$	$d_{stop} = 4[s]$

Table 1. Hyper-parameters of the XTrack audio segmentation algorithm.

probability distributions over these classes (e.g. bottom plot of Figure 1) is the input signal from which XTrack infers music segments.

Table 1 presents the fundamental operations of the XTrack algorithm. To detect when the music starts, we look for frames start$[i]$ where the probabilities associated with the music class are above Θ_{music}=0.2 for L_{music}=20 consecutive seconds. After a music start frame has been identified, we look for the next time frame stop$[i]$ where the probabilities associated with the music class are below $\Theta_{non-music}$=0.1 for L_{music}=4 consecutive seconds. We iterate this process until the end of the audio file.

We then combine these boundary frames to segment the original audio file into music and non-music segments. According to the value of the class-constrained YAMNet predictions within the non-music segments, we predict its content (speech, applause, or silence) to label the predicted segments automatically.

For a more pleasant playback of segmented music tracks, we index the start of a music segment $d_{start} = -0.1[s]$ before the predicted start$[i]$ and $d_{stop} = 5[s]$ after the detected stop$[i]$.

3.2 Evaluation

To evaluate the segmentation decisions of XTrack, we compare the algorithm results with the indexed database of the MJF archives. In particular, we look at how the onsets of predicted and human-labeled segments differ. In the Results section, we report the statistics of these timing differences.

Because of the subjective nature of the indexation process, we also conducted a qualitative evaluation of the segmentation algorithm. To do so, we randomly select examples where the XTrack predictions and the MJF data differ and report the audio contexts before and after these segment onsets.

4. RESULTS AND DISCUSSION

To generate the following results, we applied our segmentation algorithm to 10 years (from 2008 to 2017) of MJF live recordings. It takes less than 1 minute for XTrack to process 1 hour of audio on a MacBook Pro Retina 2013 (CPU 2.6GHz Intel Core i5, 16GB RAM).

4.1 Quantitative evaluation

In Figure 2 we reported the statistics of the timing differences between the automatic segmentation of XTrack and the human-labeled MJF data. These values were found by aligning predicted and human-labeled indexes. XTrack recovered 75% of the indexes. Out of every predicted start/stop, 75%/60% fall before the human-labeled timing and 25%/40% after. When XTrack finds a music segment, 50% of the time it places its start/end within 0.9/2.9 seconds from the human-labeled index. These larger differences for the music stops are expected as the MJF stop indexes are placed with some flexibility (i.e. a few seconds after the end of the music). In summary, these statistical values suggest that the playback of individual music tracks, as determined by XTrack, will be pleasant and respect the actual music boundaries of the live concert.

XTrack decisions are based on several hyper-parameters (see Table 1). These values were found with trial and error to optimize the algorithm's quantitative and qualitative performances. In case the concert setlist is available (the number of individual tracks is known), an

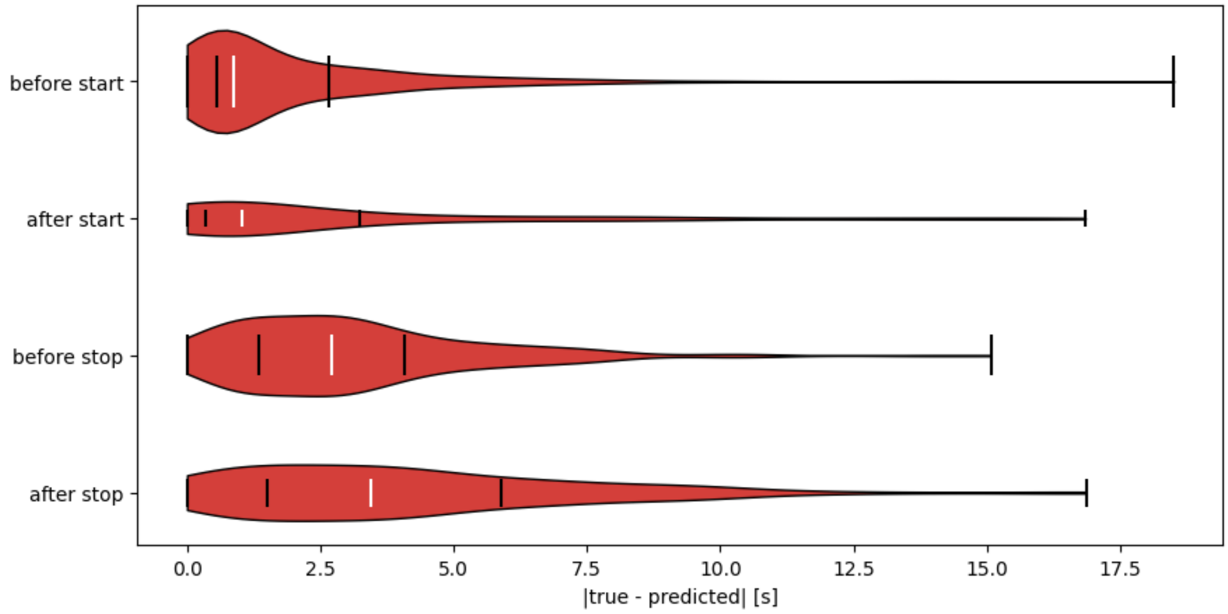

Figure 2. Densities of the timing differences between the predicted and human-labeled starts and stops of music segments. The density is computed using Gaussian kernel estimators of the python `matplotlib.pyplot.violinplot` function. To better understand the values, we chose to visualize separately the timing differences when the predicted transition is before (too early) or after (too late) the human-labeled index. The width of the densities represents the number of predictions falling in each category (before or after). The vertical bars indicate the two extrema and the [0.25 0.5 0.75] quartiles. The values in seconds are: before start [0.56, 0.87, 2.65], after start [0.36, 1.04, 3.25], before stop [1.35, 2.71, 4.07], and after stop [1.51, 3.45, 5.89].

automated grid search or a manual correction for the hyperparameters values can be performed to find the correct number of tracks. For example, the number of consecutive non-music frames required to add a music stop index can be adapted to find fewer or more tracks.

4.2 Qualitative evaluation

In Table 2 we describe six randomly selected segments where there is a difference between the MJF indexes and the automated segmentation of XTRACK.

10% of predicted music starts cannot be aligned with the MJF data. These are boundaries found by XTRACK that are not present in the MJF data. Row 4 of Table 2 describes such an event. In that case, this additional index does not harm the pleasantness of the playback of individual music tracks.

More than 75% of `music` segments are found by XTRACK. The remaining includes track changes without interruption in the music. In particular, 25% of the `music` indexes from MJF data are music-to-music transitions. To cope with that, XTRACK sometimes adds a `stop` index between two consecutive `music` indexes of the MJF data. Indeed, 20% of the predicted `stop` could not be aligned with the MJF data and are therefore supplementary indexes where XTRACK predicted an end of a music track before the start of the next track. The MJF annotators sometimes decided to add a new `start` index without a `stop` before.

`start`$[i]$ at 10:00 in Figure 1 and rows 2+6 of Table 2 illustrate a continuous music-to-music transitions that is not detected. As XTRACK makes its boundary decisions based on when the music starts and stops, these transitions without interruption in the music cannot be detected. In these cases, the playback from XTRACK includes more than one track. However, as the music does not stop, it is still a pleasant cut to listen to. Additional analysis (e.g. by adding more classes to C_{music} and $C_{non-music}$) could help detect changes in musical contents and find these music-to-music transitions.

To sum up, many differences between the XTRACK segmentation and the MJF data can be explained. They can be due to subjective choices in the manual indexation of the Montreux Jazz Archive. As a final check, we randomly listened to individual tracks determined with XTRACK. These listening sessions were fluid and a more pleasant experience than listening to the full concert.

5. CONCLUSIONS AND FUTURE WORK

In this study, we introduced XTRACK, an innovative algorithm designed to automate the segmentation of live music recordings. By leveraging the sound classification capabilities of YAMNET, our approach has significantly streamlined the process of isolating musical segments from live concert recordings, addressing a critical need within the Montreux Jazz Festival archives and potentially other music collections.

For automated indexation of digital music (e.g. on YouTube) or as a tool to index musics in large collections, XTRACK can suggest a first segmentation. This proposition can be fine-tuned by a human checker or used as it is.

MJF	XTrack	Description	Comment
08:12 applause 09:04 music	08:17 speech 09:03 music	music > 08:12 applause 08:30 speech 09:04 music	On one side, the MJF index starts at the very beginning of the 08:12 applause. On the other side, XTrack adds 5 seconds of applause after the end of the music. There is a difference in the labeling as well: XTrack predicts a speech segment. The MJF label is applause. There is indeed a 34s-long speech validating this XTrack labeling decision.
24:39 music	None	music > 24:30 cheering 24:40 singing	A single high guitar note is held from 24:30 to 24:55. The music did not stop but there is a transition to a new track. Not stopping here do not disturb a pleasant playback but a new track begins.
01:05:30 music	01:05:41 music	silence > 01:05:30 music 01:05:36 silence 01:05:41 music	The silence between the small introduction and the rest of the track is too long for XTrack to include it in the music segment. A playback starting as indexed by XTrack (01:05:41) is still very pleasant.
None	01:27:28 music	cheering > 01:27:28 music	The music starting at 01:27:28 is a background music played from the concert hall loudspeakers on top of some crowd noises. Whereas, this is not a concert track, our algorithm predicted a music start index. However, that is because YAMNet detected the background music. Adapting the Θ_{music} can be made to be more strict and might be a way to prevent such mistakes. However, listening to a few seconds of this segment is sufficient to throw it away.
40:10 applause 43:10 music	39:56 speech 43:10 music	music > 40:00 cheering 40:45 quiet 42:30 speech 43:10 music	This is a pre-recorded electronic music concert. The music is displayed through loudspeakers in the hall. At 39:50, the electronic music outputs recorded applause that confuses our algorithm, which cut earlier than the MJF data. This cut is valid for a pleasant playback with a smooth fade-out. Moreover, the speech prediction fits the audio context.
06:14 music	None	music > 06:14 music	Transition to a new track without interruption not found by XTrack.

Table 2. Qualitative description of six randomly chosen discrepancies between XTrack decisions and the human-labeled MJF archives. The **Description** column illustrates the analysis of a musician listening to the original audio files. In the **Comment** column, we explain in words the context.

Its code is available on GitHub [5].

Specifically, the algorithm demonstrated a high success rate in accurately identifying and segmenting music tracks, thereby facilitating a more efficient archival process. The outstanding performance of XTrack in music segmentation has inspired us to explore additional applications of this technology.

Given the rich emotional labels attached to YAMNet dataset, we propose extending the use of XTrack beyond traditional music segmentation. By integrating these emotional labels, we aim to construct segments that not only identify different musical pieces but also classify the underlying emotions conveyed in each segment. This development opens up new avenues for emotional analysis within music, allowing for a deeper understanding of the impact of live music on audiences. In fact, this approach promises to enhance the user experience by providing more nuanced and emotionally resonant music listening experiences. Moreover, it can serve as a valuable tool for musicologists, psychologists, and other researchers interested in the interplay between music and emotions.

The XTrack segmentation algorithm can be used for other interesting audio segmentation tasks. For example, we can detect in- and out-of-tune segments by analyzing the magnitude of high-frequencies of the spectrogram (fast Fourier transform fft of the audio files). By imposing a threshold on the amplitude of a high-frequency range, we can find the longest segment of in- or out-of-tune music. Used as a pedagogical tool, XTrack segmentation algorithm is promising.

Acknowledgments

We thank the Cultural Heritage and Innovation Center (CHC) of École Polytechnique Fédérale de Lausanne (EPFL) and the Information and Language Section (SLI) of the Lausanne university (UNIL) for hosting this project. This project was possible thanks to the BNF national qualification program of Bern university (UNIBE). Finally, we are grateful for the recording work of the Montreux Jazz Festival, for its preservation by the Claude Nobs Foundation, and to all the musicians who contributed to this amaz-

ing concert collection.

6. REFERENCES

[1] M. Plakal and D. Ellis, “Yamnet,” https://github.com/tensorflow/models/blob/master/research/audioset/yamnet, 2020.

[2] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[4] D. El Badawy, P. Marmaroli, and H. Lissek, “Audio novelty-based segmentation of music concerts,” in *Acoustics 2013*, 2013.

[5] Florian Colombo, “Xtrack,” https://github.com/FlorianColombo/xtrack, 2024.