Timbre Comparison in Note Tracking from Onset, Frames and Pitch Estimation

Carlos Hernandez-Olivan, Ignacio Zay Pinilla, Jose R. Beltran

Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: {carloshero, 628123, jrbelbla}@unizar.es

Abstract

Note Tracking (NT) is a subtask of Automatic Music Transcription (AMT) which is a critical problem in the field of Music Information Retrieval (MIR). The aim of this work is to compare the performance of two models, one for onsets and frames prediction and another one with pitch detection and a note tracking algorithm in order to study the behaviour of different timbres and families of instruments in note tracking subtasks.

Introduction

The Automatic Music Transcription problem can be separated into several subtasks, including multipitch estimation or frame-level transcription on pitches (MPE), note-level transcription on pitches, onset, and duration, also known as note tracking (NT) or instruments identification. Although transcribing a monophonic recording is considered to be a solved problem, ATM still remains an open research problem when it comes to multiple instruments (mixed signals) and polyphonic music [1].

Previous studies adressed ATM by two principal methods: Non-Negative Matrix Factorization (NMF) and Neural Networks (NNs). NN methods usually use spectrograms as inputs to later process them with long short-term memory layers or CNNs. Most of ATM works that use NN are based on polyphonic piano transcription such as Magenta Onsets and Frames (OaF) [2]. More recent studies adress ATM with multi-task deep learning techniques by taking a mixed signal and they attempt to transcribe the output stems of the source's separation subtask [3].

Fundamental Frequency Estimation

Fundamental frequency (f_0) estimation has been studied over decades. Recent approaches are based on template matching with the spectrum of a waveform and other use a Hidden Markov Model (HMM) to decode the most probable sequence of pitch values. Most recent and best performing methods such as Crepe NN [4] address the monophonic pitch estimation by estimating the fundamental frequency (f_0) of the input with CNNs.

Methods

In our work, we take a clean stem of an instrument as an input, so we have to address music transcription task by performing one after other MIR tasks such as pitch estimation or onsets detection. We perform pitch estimation with Crepe NN to estimate the pitch and we use the results of the model to perform the note tracking. We test our results over multiple music instruments with different timbres, and we compare the results with the state-of-the-art Onsets and Frames model.

Minimum Pitch Confidence Estimation

The minimum confidence value (c) that we use in our note tracking algorithm can be estimated by different approaches based on the histogram of the estimated confidences that frequencies have for every time step. In our work, we have used a triangulation algorithm, a gaussian distribution over the frequencies histogram and the Otsu's thresholding algoritm which perform the best results. A comparison between Magenta OaF and the tracking algorithm are shown in Table 1 and the note identification results of the tracking algorithm are shown in Fig. 1.

Tracking Algorithm

The tracking algorithm designed for this work takes as its inputs the outputs of the Crepe NN that are arrays of frequencies, time and confidences and it outputs a MIDI file by writing the note on, note off and pitch events. There is an additional input passed to the algorithm that is the *minimum confidence* which is the minimum pitch confidence that our algorithm uses to group the same pitch values over time. Pitch confidences below some values are

discarded. After discarding these notes, we group notes over time. We set the output of Crepe NN to a time step of 10ms, so the model predicts an estimated frequency and confidence every 10ms along the duration of the audio file.

Algorithm 1: Monophonic note tracking

```
Input: Array or lists of frequency f,
       confidence c and time t, and
       minimum confidence value.
initialize pitch, note and off lists and time step
for time step in t to length(t) do
    if confidence > minimum confidence do
        while frequency = next frequency do
            agrupate frequencies in the same note
        end while
        write note on and note off and append
        convert note frequencies into pitches
end for
```

Results

The dataset used in this work is the Slakh2100 dataset [5]. The MIDI files are aligned with the audio files and are used as the grund truth picth and duration of the notes.

Music transcription is evaluated with Precision (P), Recall (R) and F-measure (F). An estimated note is considered correct if its onset is within a tolerance of 50ms of the reference note and if its pitch is within a tolerance of 50cents (which corresponds to a quarter tone). Note offsets and velocities have not being analyzed. The results in terms of F-measure are presented in Table 1.

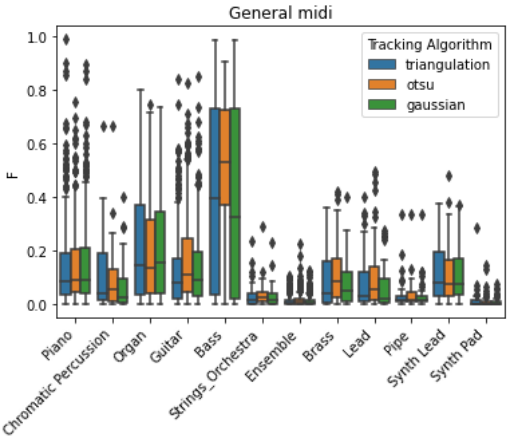

Fig. 1. Note identification reults in terms of F-measure of isolated instruments in Slakh2100 database with crepe and the tracking algorithm for triangulation, gaussian and Otsu's minimum confidence estimation methods

Conclusions

This work shows an overview of how different timbres affect some subtasks of automatic music transcription such as note tracking from an estimated f_0 or onsets and frames prediction. We test isolated instruments with polyphony of Slakh2100 dataset with Magenta OaF model and with a pitch estimation model followed by a note tracking algorithm based on the predicted f_0 confidence, so we do not have to perform the onsets detection subtask. By comparing the results, we can see that timbre and instrument onsets are variables that affect the results of music transcription in different subtasks.

REFERENCES

[1]. Benetos, Emmanouil, et al. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 2018, vol. 36, no 1, p. 20-30.

[2]. Hawthorne, Curtis, et al. Onsets and frames: Dual-objective piano transcription. arXiv preprint arXiv:1710.11153, 2017.

[3]. MANILOW, Ethan; SEETHARAMAN, Prem; PARDO, Bryan. Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments. En *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020. p. 771-775.

[4]. Kim, Jong Wook, et al. CREPE: A convolutional representation for pitch estimation. En 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. p. 161-165.

[5]. Manilow, Ethan, et al. Cutting music source separation some slakh: a dataset to study the impact of training data quality and quantity. En 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019. p. 45-49.

Tabla 1. Results of music transcription from Slakh2100

Instrument	Method	Note F_1
Bass	Note Tracking (Otsu)	0.5575
	Magenta OaF	0.6694
Guitar	Note Tracking (Otsu)	0.3168
	Magenta OaF	0.6432
Synth pad	Note Tracking (Otsu)	0.0679
	Magenta OaF	0.1842
Synth lead	Note Tracking (Otsu)	0.3016
	Magenta OaF	0.3459
Brass	Note Tracking (Otsu)	0.3215
	Magenta OaF	0.5899
Strings	Note Tracking (Otsu)	0.1631
	Magenta OaF	0.4888
Organ	Note Tracking (Otsu)	0.2826
	Magenta OaF	0.2767
Piano	Note Tracking (Otsu)	0.3205
	Magenta OaF	0.9068
Chromatic Percussion	Note Tracking (Otsu)	0.2464
	Magenta OaF	0.5627