Insert here your thesis' task.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Master's thesis

Multi-instrument music transcription

Bc. Yevhen Kuzmovych

Department of Applied Mathematics
Supervisor: Ing. Marek Šmíd, Ph.D.

December 26, 2019

Acknowledgements

THANKS (remove entirely in case you do not with to thank anyone)

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on December 26, 2019 .....

Czech Technical University in Prague
Faculty of Information Technology
© 2019 Yevhen Kuzmovych. All rights reserved.
This thesis is school work as defined by Copyright Act of the Czech Republic. It has been submitted at Czech Technical University in Prague, Faculty of Information Technology. The thesis is protected by the Copyright Act and its usage without author's permission is prohibited (with exceptions defined by the Copyright Act).

Citation of this thesis

Kuzmovych, Yevhen. *Multi-instrument music transcription.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2019.

Abstrakt

V několika větách shrňte obsah a přínos této práce v českém jazyce.

Klíčová slova Replace with comma-separated list of keywords in Czech.

Abstract

Summarize the contents and contribution of your work in a few sentences in English language.

Keywords Replace with comma-separated list of keywords in English.

Contents

List of Figures

Introduction

Problem definition

Tuning

Chapter 1

State-of-the-art

This chapter discusses existing solutions for music transcription including source instruments separation, pitch detection, event detection, etc.

1.1 Source separation

There were many successful attempts for music score source separation[1, 2, 3]. Performance of such projects are commonly measured according to *Source Separation campaign (SiSeC)*[4] on the standard *musdb18* and *DSD100* datasets.

Latest and most successful project in this field is *Spleeter*[1]. It is a project of Deezer[1]. It takes similar approaches to previous solutions by University of London and Spotify[2]. Spleeter's pre-trained models will be used in the module responsible for music source separation described in detail in the chapters 2 and 3.

Following approaches are described in [1, 2, 3].

1.1.1 Approach

The pre-trained models are U-nets[2] and follows similar specifications as in *Singing voice separation: a study on training data*[3]. The U-net is a encoder/decoder Convolutional Neural Network (CNN) architecture with skip connections[1]. Architecture used in this approach showed a state-of-the-art results on DSD100 dataset[2] and in the last SiSeC[5].

1.1.2 U-net architecture

The U-Net shares the same architecture as a convolutional autoencoder with extra skip-connections that bring back detailed information lost during the

[1]Deezer is a French online music streaming service.

encoding stage to the decoding stage. It has five strided[2] 2D convolution layers in the encoder and five strided 2D deconvolution layers in the decoder.

The goal of the neural network architecture is to predict the vocal and instrumental components of its input indirectly: the output of the final decoder layer is a soft mask for each source that is multiplied element-wise with the mixed spectrogram to obtain the final estimate.

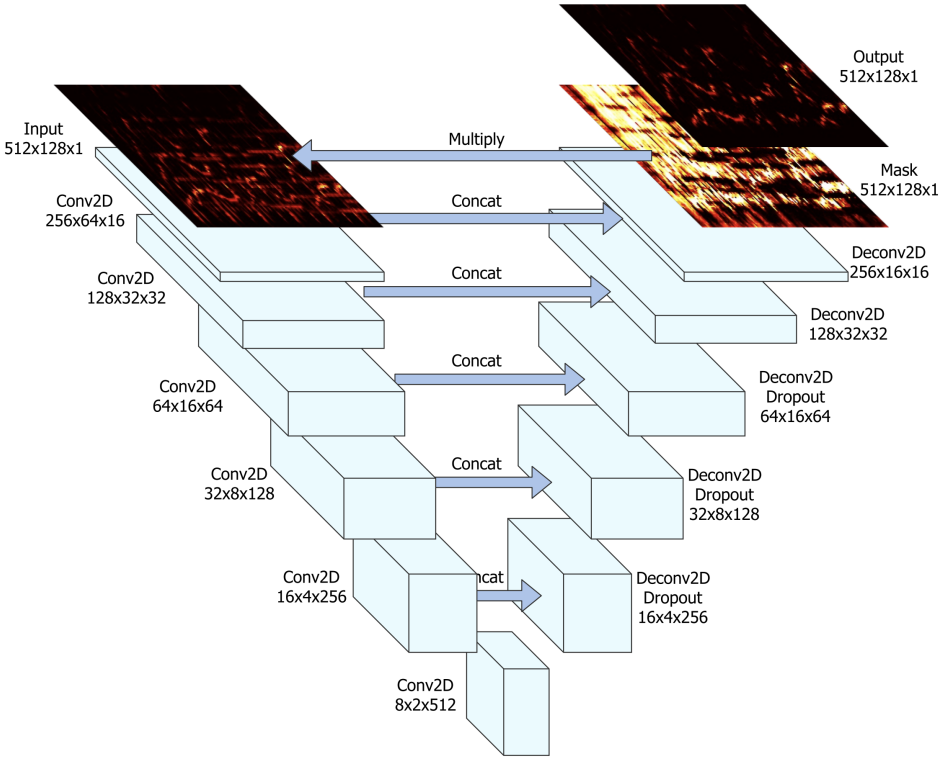


Figure 1.1: Network Architecture[2]

1.1.3 Data and training

Spleeter's training dataset is an internal Deezer's dataset and is not shared (for copyright reasons).

Another project with similar approach, as explained in the dedicated article[3], uses two datasets during training of the models: *MUSDB* and *Bean*.

MUSDB[7] is the largest and most up-to-date public dataset for source separation. It contains 150 songs of western music genres primarily pop/rock,

[2]Transposed convolutions – also called *fractionally strided convolutions* – work by swapping the forward and backward passes of a convolution. One way to put it is to note that the kernel defines a convolution, but whether it's a direct convolution or a transposed convolution is determined by how the forward and backward passes are computed.[6]

some hip-hop, rap and metal songs. And each song consists of four audio tracks: drums, bass, vocal and other. Original mix (and input of the model) is produced by summing tracks of four sources (expected outputs) together.

Chapter 2

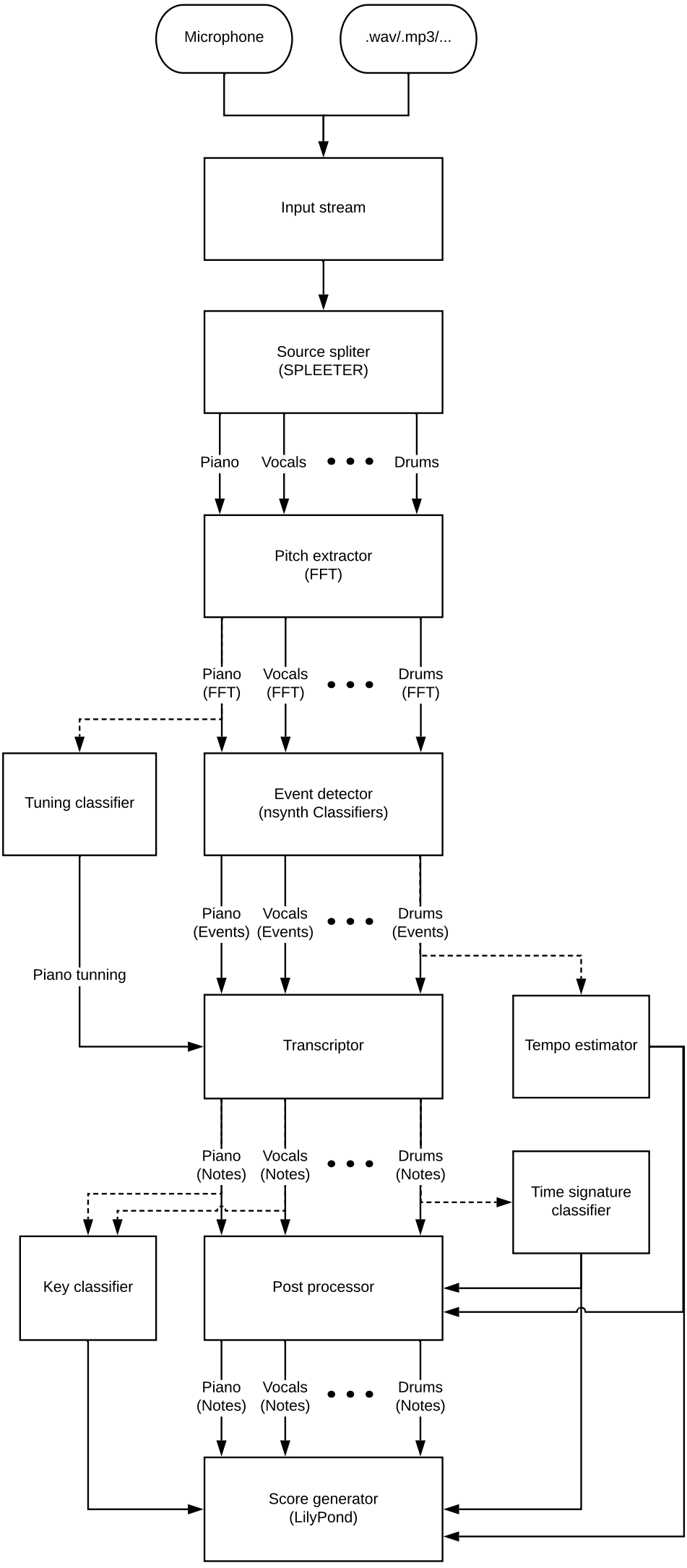
Analysis and design

This chapter defines architecture of the chosen solution. It provides details of used approaches for music sources separation and models used in it, pitch extraction and signal analysis, event detection, etc.

2.1 Architecture

The implementation of the system is separated into logical parts responsible for sound data streaming, music source separation, pitch and events detection, transcription and score generation. Following diagram shows architecture of the solution. Arrows represent data flow. Dotted arrows represent flow that is optional. If given parameters (like tuning, tempo, time signature and key) are specified by user, they are not being estimated. Each rectangular block represents logical module in implementation.

Detailed description of each component is in the dedicated section following the diagram.


Microphone
.wav/.mp3/...
Input stream
Source spliter
(SPLEETER)
Piano
Vocals
Drums
Pitch extractor
(FFT)
Piano
(FFT)
Vocals
(FFT)
Drums
(FFT)
Tuning classifier
Event detector
(nsynth Classifiers)
Piano
(Events)
Vocals
(Events)
Drums
(Events)
Piano tunning
Transcriptor
Tempo estimator
Piano
(Notes)
Vocals
(Notes)
Drums
(Notes)
Time signature
classifier
Key classifier
Post processor
Piano
(Notes)
Vocals
(Notes)
Drums
(Notes)
Score generator
(LilyPond)

2.2 Audio streaming

This implementation directly works only with Waveform Audio File Format (WAVE) (.wav/.wave). Any other format is converted to WAVE first, then processed.

2.2.1 WAVE format

WAVE is an audio file format standard, developed by Microsoft and IBM, for storing an audio bitstream on PCs. What's important for this thesis and implementation is that it stores data in chunks in Linear pulse-code modulation (LPCM) format. This format allows to perform Discrete Fourier transform (DFT) used in pitch extraction.

2.2.2 Sampling rate

LPCM mentioned above stores sampled amplitude of recorded audio at specific sampling rate (frequency, measured in Hz).

The most common sampling rate is 44.1 kHz, or 44100 samples per second. This is the standard for most consumer audio, used for formats like CDs[8].

The sampling rate determines the range of frequencies captured in digital audio. According to *Nyquist Theorem*, a signal which has a Fourier transform having only frequencies upto a certain maximum f_m, we can obtain the analog signal $f(t)$ from the sampled signal $f'(t)$ by passing the sampled signal $f'(t)$ through a low pass filter provided that the sampling frequency f_s is more than twice the maximum frequency f_m present in the signal i.e. , $f_s > 2f_m$[9]. Hence, having 44100 Hz sampling rate, we can reproduce and analise frequencies up to 22050 Hz. The lowest frequency a person can hear is 20 Hz. The highest frequency humans can hear are in the range of 20.000 Hz, but only young people can hear such high tones[10].

The implementation is able to process input sound with any sampling rate, though lower sampling rates may cause inaccuracies for high frequencies due to effect called *aliasing*. The phenomenon of aliasing occurs if the sampling rate is less than the Nyquist rate (2× the highest frequency). If sampling rate is less than or greater than the Nyquist rate, it is called under sampling or over sampling. Aliasing phenomenon occurs only for under sampling. If the sampling frequency is too low the frequency spectrum overlaps, and becomes corrupted[9].

2.3 Music source separation

First step of the sound processing is separation of the sound into source instruments (i.e. voice, guitar, piano, etc.)

As was mentioned in previous chapter, for separation of source instruments, this implementation uses *Spleeter*. *Spleeter* is a fast and state-of-the art music source separation tool with pre-trained models[1]. Its implementation contains three pre-trained models:

- vocals/accompaniment separation
- 4 stems separation as in SiSeC[4] (vocals, bass, drums and other)
- 5 stems separation with an extra piano stem (vocals, bass, drums, piano and other). It is, to the authors knowledge, the first released model to perform such a separation.

Estimations for all the models is performed in a frequency domain of the sound. Meaning that sound data from time domain is converted to frequency domain using Fast Fourier transform (FFT), passed to the models described in section 1.1.2 about U-net architecture. Output of the model is separated tracks for each instrument and voice. To get sound of each instrument and voice in time domain (as it would be represented in WAVE), we'd need to pass it through Inverse Discrete Fourier transform (IDFT). Though it is unnecessary, as all the subsequent processing will be performed on the sound in frequency domain.

More about FFT in the following section 2.4 about pitch extraction.

2.4 Pitch extraction

2.5 Event detection

2.6 Tuning classification

2.7 Transcription

2.8 Tempo estimation

2.9 Time signature estimation

2.10 Key classification

2.11 Post processing

2.12 Score generation

Chapter 3

Realisation

This chapter provides details of implementation, used tools, training and testing of the models.

3.1 Used tools

3.2 Music source separation

Conclusion

Possible improvements

Bibliography

[1] Hennequin, R.; Khlif, A.; et al. Spleeter: A Fast And State-of-the Art Music Source Separation Tool With Pre-trained Models. Late-Breaking/Demo ISMIR 2019, November 2019, deezer Research.

[2] Jansson, A.; Humphrey, E.; et al. Singing voice separation with deep U-Net convolutional networks. October 2017. Available from: `https://openaccess.city.ac.uk/id/eprint/19289/`

[3] Pretet, L.; Hennequin, R.; et al. Singing Voice Separation: A Study on Training Data. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, doi:10.1109/icassp.2019.8683555. Available from: `http://dx.doi.org/10.1109/ICASSP.2019.8683555`

[4] Stöter, F.-R.; Liutkus, A.; et al. The 2018 Signal Separation Evaluation Campaign. 2018, `1804.06267`.

[5] Liutkus, A.; Stöter, F.-R.; et al. The 2016 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, edited by P. Tichavský; M. Babaie-Zadeh; O. J. Michel; N. Thirion-Moreau, Cham: Springer International Publishing, 2017, pp. 323–332.

[6] Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. 2016, `1603.07285`.

[7] Rafii, Z.; Liutkus, A.; et al. The MUSDB18 corpus for music separation. Dec. 2017, doi:10.5281/zenodo.1117372. Available from: `https://doi.org/10.5281/zenodo.1117372`

[8] Digital Audio Basics: Sample Rate and Bit Depth. 7 2019, [Cited 2019-11-24]. Available from: https://www.izotope.com/en/learn/digital-audio-basics-sample-rate-and-bit-depth.html

[9] Ghosh, S. *Signals and Systems*. Always Learning, Pearson Education Canada, 2005, ISBN 9788177583809. Available from: https://books.google.com.ua/books?id=Sg6y_EBYCEsC

[10] Wendt, S. *Roots of Modern Technology: An Elegant Survey of the Basic Mathematical and Scientific Concepts*. Springer Berlin Heidelberg, 2010, ISBN 9783642120626. Available from: https://books.google.com.ua/books?id=c8TdQmtOD-AC

[11] McGrain, M. *Music Notation: Theory and Technique for Music Notation*. Berklee guide, Berklee Press, 1990, ISBN 9780793508471. Available from: https://books.google.cz/books?id=S_y7JAZqx6QC

[12] Reference. An Explanation of Clefs: Treble, Bass, Alto, Tenor. [Cited 2019-10-05]. Available from: https://makingmusicmag.com/explanation-clefs-treble-bass-alto-tenor/

[13] Key signature and music staff. 4 2012, [Cited 2019-10-23]. Available from: https://www.aboutmusictheory.com/key-signature.html

[14] of Encyclopaedia Britannica, T. E. Time signature. 11 2017, [Cited 2019-10-23]. Available from: https://www.britannica.com/art/time-signature

Appendix A

Acronyms

WAVE Waveform Audio File Format

LPCM Linear pulse-code modulation

FFT Fast Fourier transform

DFT Discrete Fourier transform

IDFT Inverse Discrete Fourier transform

CNN Convolutional Neural Network

SiSeC Source Separation campaign

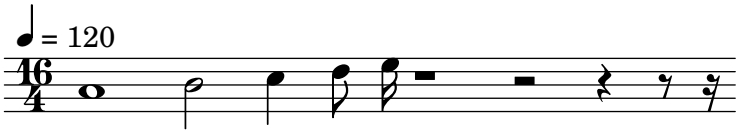
Appendix **B**

Musical notation

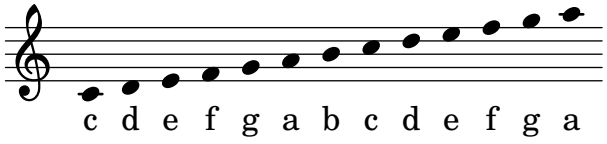
Music notation, when properly applied, can completely describe any musical score in a simple, concise manner. In order to achieve this, music notation must describe all definable parameters of each sound, specifically[11]:

- duration
- pitch
- dynamic
- timbre

Duration is described by time signature ($\frac{4}{4}$, $\frac{3}{4}$, $\frac{7}{8}$, etc.), tempo (primarily, beats per minute: ♩ = 120), and duration values of note-heads (𝅝, 𝅗𝅥, ♩, ♫, ♪, etc.) and rests (𝄻, 𝄼, 𝄽, 𝄾, etc.):



Pitch is defined by position of the note on the staff, key, accidentals (♭, ♯, ♮) and the specified clef (𝄞, 𝄢, 𝄡, etc.):



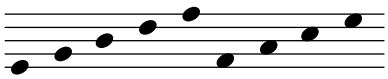
Dynamic of a sound describes its amplitude or loudness (***pp***, ***p***, ***mf***, ***f***, ***ff***, etc.), its emotional intensity and change through time.

Timbre describes specific color of a played note/sound. Timber primarily depends on the instrument played but also can define other instrumental directions (i.e. *on bell of cymbal*, etc.)

B.1 The Staff

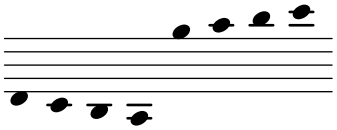
The base for all musical scores is the *staff*. All other music symbols go are placed on the staff or in relation to it.

The staff consists of five horizontal lines and four spaces between the lines. Every note-head is placed on one of the lines or on one of the spaces between the lines. The higher the note-head on the staff - the higher the pitch of the produced note.



B.2 Leger Lines

Obviously, five lines and five spaces can provide only limited range of notes (precisely, eleven places to put the note-head, including just beneath the first(bottom) line and above fifth(top) line). If notes from outside this range are needed, they are placed on or between so-called *Leger lines*. These are the lines placed above or beneath the main staff only in places where they are needed, so for each note individually.

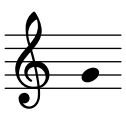


B.3 Clefs

The specified *clef* defines location of each pitch on the staff. The most commonly used clefs are the Treble and the Bass clefs[12].

B.3.1 The Treble Clef

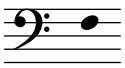
The *Treble Clef* (or *G clef*, because the middle curl of it encircles line on the staff that represents a G-note) is used for most high-sounding instruments (i.e. violin, guitar, ukulele, flute, clarinet, saxophone, trumpet, etc.).



As it defines second line as G, the lines on the staff, from bottom to top, are E, G, B, D, F. The spaces then are F, A, C, E. The middle C[3] goes on the first leger line below the treble staff.

B.3.2 The Bass Clef

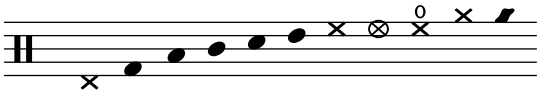
The *Base Clef* (or *F clef*, because line between two dots on the symbol represents an F-note) is used for low sounding instruments (i.e. bass guitar, cello, trombone, tuba, etc.)



As it defines fourth line as F, the lines on the staff are G, B, D, F, A, and the spaces are A, C, E, G. The middle C goes on the first leger line above the bass clef.

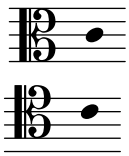
B.3.3 The Percussion Clef

The *Percussion Clef* is commonly used for drum-set notation. Each line and space represent different part of the drum kit. They are often predefined at the start of the part in so-called *key* or *legend*, or when they first appear in the score.



B.3.4 The Alto and Tenor Clefs

Alto Clef (or *C clef*, because line in the middle of the alto staff represent middle C) and The *Tenor Clef* are less often used clefs. The viola and the alto trombone are generally the only instruments that use the Alto clef. Tenor clef is occasionally used to represent the upper ranges of the cello, double bass, bassoon, and trombone.

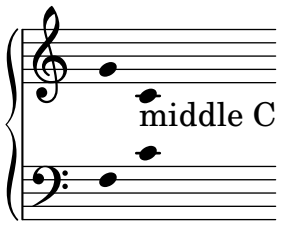


The lines of the alto staff are F, A, C, E, G, and the spaces are G, B, D, F. Similarly, for tenor clef, C is moved up one line from alto clef, making the notes on the lines D, F, A, C, E and notes in the spaces E, G, B, D.

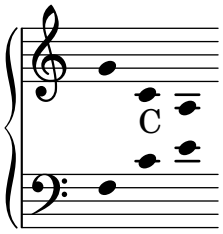
[3] *Middle C* is a commonly used reference note. It is a closest C to the middle of a standard 88 key piano (specifically, fourth C from the left). It is around 261.63 hertz.

B.3.5 The Great Staff

The *Great Staff* or the *Grand Staff* is a combination of the treble staff and the bass staff. Usually used by piano or harp musicians.

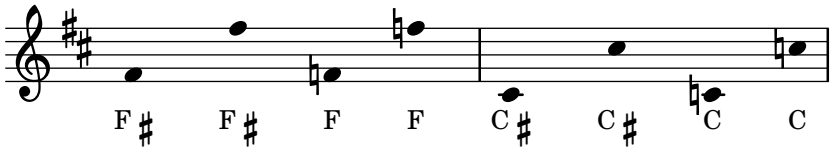


Often they also divide score into to parts played by left and right hand (i.e. on piano, treble clef part with the right hand, bass clef part with left hand). So, even if some notes belong to treble clef they may be put on leger lines above bass clef if played by left hand and vice versa.



B.3.6 Key signature

Key signature is a series of sharp symbols or flat symbols placed on the staff, designating notes that are to be consistently played one semitone higher or lower than the equivalent natural notes (for example, the white notes on a piano keyboard) unless otherwise altered with an accidental. Key signatures are generally written immediately after the clef at the beginning of a line of musical notation, although they can appear in other parts of a score, notably after a double bar.[13]



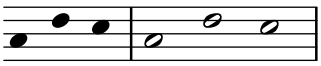
Key *D major* (defined in example above) consists of notes D, E, F♯, G, A, B, C♯. So, after the clef, notes F and C marked with a ♯, so, when they occur in a score without any accidentals, they are played one semitone higher (C♯ instead of C, etc.)

B.4 Rhythmic Description

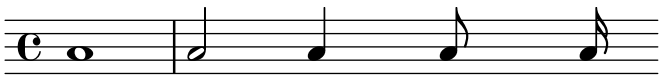
Alongside with pitch, it is required to describe rhythm. *Rhythmic description* determines exactly when note should be played and when it should stop playing. Notationally it is defined by note-heads, stems, flags, beams, rests, and time signature.

B.4.1 Note-heads, stems, flags, beams

There are two types of note heads open and closed.



Stems are vertical lines attached to the side of the notes-head. Together with flags, beams, and augmentation dots they define duration value:



Whole note Half note Quarter note Eighth note Sixteenth note

Two half note have the same duration as one whole note, two quoter notes have the same duration as one half note and so on.

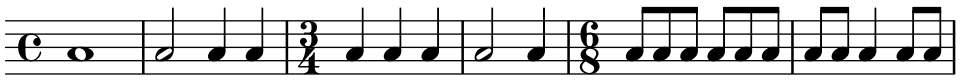
B.4.2 Rests

Same as for notes, we can define pauses in music - *rests*:

Whole rest, half rest, quoter rest, and so on accordingly.

B.4.3 Time signatures

Time signature is a sign that indicates the metre of a composition. Most time signatures consist of two vertically aligned numbers, such as $\frac{2}{2}$, $\frac{3}{4}$, $\frac{6}{8}$, and $\frac{11}{16}$. The top figure reflects the number of beats in each measure, or metrical unit; the bottom figure indicates the note value that receives one beat (here, respectively, half note, quarter note, eighth note, and sixteenth note). When measures contain an uneven number of beats falling regularly into two subgroups, the division may be indicated as, for instance, $\frac{3+4}{4}$ instead of $\frac{7}{4}$[14].



$\frac{4}{4}$ is such a common time signature that sometimes it is specified with $\mathbf{c}$ and $\frac{2}{2}$ as $\mathbf{¢}$.

Appendix C

Contents of enclosed CD

- `readme.txt` ..... the file with CD contents description
- `exe` ..... the directory with executables
- `src` ..... the directory of source codes
  - `wbdcm` ..... implementation sources
  - `thesis` ..... the directory of LaTeX source codes of the thesis
- `text` ..... the thesis text directory
  - `thesis.pdf` ..... the thesis text in PDF format
  - `thesis.ps` ..... the thesis text in PS format