Automatic Music Transcription using Variable Q-Transform and Deep Learning

Sbonelo Mdluli : 1101772, Moshekwa Matthews Malatji : 1387556
Group: 20G04
School of Electrical & Information Engineering, University of the Witwatersrand, Private Bag 3, 2050, Johannesburg, South Africa

***Abstract*—Automatic music transcription is the process of converting music signal into musical notes. This paper presents the project plan and milestones used to accomplish the final product. The milestones for the project are; system design, data collection, digital signals processing, model development and testing methodologies. The methodology which will be followed in development of the Automatic Music Transcription is the Cross-Industry Process for Data Mining methodology. This is a structural approach to a data mining project, is adopted in this context to draw knowledge and insight from the dataset. Variable Q-Transform and Deep Learning algorithms are the main concepts to be explored when implementing the transcription system. Music instruments to be transcribed in real time are the piano and drums using the MAESTRO & Groove MIDI dataset, respectively. The Automatic Music Transcription assumes that audio reconstruction and vocal separation is not required and successful transcription the accuracy of the model is required to be at least 50 %. A user interface implemented with PyQt5, Tkinter or MATLAB GUI will provide the user a graphical means of interacting with the AMT system. The project is be completed using an iterative approach in order to refine and improve the model. The project management and planning for the 8 week long project are discussed and the project schedule is registered on the Gantt Chart.**

I. INTRODUCTION

In this paper we present the project plan to be used when implementing an Automatic Music Transcription (AMT) system. The project plan details the project specifications, implementation, testing methodologies and the project management aspect of the investigation. AMT is defined as the design of computational algorithms to convert acoustic music signals into of music notation [1]. This is a process that is of concern to signal processing and artificial intelligence. AMT is mainly implemented using neural networks and non- negative matrix factorization.

The document is structured as follows; Section III outlines the AMT project specifications including the assumptions made, the relevant success criteria and constraints to be adhered to. This is followed by Section II which is a description of the various literature on AMT which is explored in order to derive the proposed methodology. The fundamental approach to the AMT project is discussed on Section IV, it is a structural and procedural approach which has subsequent descriptive sections. An overview of the musical instruments used for music transcription is on Section V, which is followed by the description of the proposed signals processing technique employed for AMT on Section VI. This is followed by a detailed discussion of the artificial intelligent modelling technique for AMT as illustrated in Section VII. In addition, Section IX & X discusses the Post Processing Unit and Testing methodology for the model, respectively. The penultimate section for this document are Section XI which outlines the techniques employed for building a suitable User Interface for the AMT. The project management overview which also outlines the risk analysis and methods employed to ensure project planning techniques by forms of a Work Breakdown Schedule and Gantt Chart are included on Section XII.

II. BACKGROUND

Various literature on AMT has been explored in deriving in deriving suitable to optimize performance and prioritize efficiency. There has been many approaches to AMT with the a common goal to produce musical notation or score from audio signals using different forms of signals processing and modeling techniques. However, [1] denotes that AMT approaches are classified according to the following categories: Frame level, Note level, Stream level and Notation level. Frame-level transcription refers to the estimation of the pitch and number of notes which are present in a frame.This is usually a common level where AMT transcription occurs such as [2] which focuses on multi-pitch estimation of piano sounds using Probabilistic Spectral Smoothness Principle. Furthermore [3] also focuses on

spectral and temporal representations for multi-pitch estimation of polyphonic music, other literature[4] employs the Bayesian methods. Note level or note tracking is similar to the frame level transcription with the addition of connecting pitch estimates into notes. This level of transcription is often incorporates note tracking of the three music notes elements: pitch,onset time, and offset time[1] and pitch estimates of each frame. Median filtering[3] is an example of literature where transcription is classified to be on this level. As the level of transcription increases from the frame level to the stream level, the complexity and degree of transcription also increases as this introduces more musical variables. Stream level transcription focuses on classification of estimated pitches and notes into streams that correspond to certain musical instruments or voices[1]. The work done in [5][6] involves estimation of musical frames which includes pitch and notes and clustering them into different streams/sources.

The AMT approach presented in this paper is within the Notation Transcription Level which is focused on transcribing music into music scores which are readable by humans. Furthermore, this level of transcription focused on digital signal processing and artificial intelligence approaches to successfully transcribe audio signals into music scores. A common approach is obtain a time-frequency transform of audio signals and applying modeling techniques, but the work done in [7][8] only applies neural networks to audio signals in generating a music score. However, literature[9][10] employ the Constant Q-Transform(CQT)[11] audio signals technique to obtain a spectral representation by form of spectrograms. This technique is commonly used as it shows a much better spectral resolution at low frequencies which outperforms the Discrete Time Transforms employed in [12]. This paper proposed the AMT using the VQT and Deep Learning methods. The proposed digital signals processing technique used to obtain spectral frequency analysis and it is is a modification of the CQT, aimed at providing a well defined spectral analysis at low frequencies. The scope of the document is a description of the AMT process including the development methodologies inspired by the literature explored in this section.

III. Project Specifications

1) Assumptions: Assumptions have been made to reduce the scope and complexity while still satisfying the success criteria of the AMT project.It is assumed that the AMT will only be employed on monophonic audio signals instead of polyphonic. The audio signals from the dataset are assumed to have no vocals, thus eliminating the need to implement vocal separation techniques. The last assumptions denotes that the scope of AMT will not be concerned with audio signals reconstruction.

2) Success Criteria: The AMT is required to transcribe 2 instruments in real time. A user will interact with the system through a Graphical User Interface. For a successful transcription the accuracy of the model is required to be at least 50 %.

3) Constraints: The computational power of the machine will dictate the speed at which the models train. The quality of the data source also plays a role in the accuracy of the model. Well labelled and noise reduced music is required in order to improve the model accuracy.

IV. System Design & Methodology Overview

The proposed AMT methodology or approach in development of the AMT has to be precise and procedural in-order to achieve successful implementation. Noting that the design and implementation of the Automatic Music Transcription is a software programming, data science problem within an Electrical Engineering context. Therefore the methodology in development has to be meticulous and precise such that it accounts for Software and Data Science concepts while adhering to Electrical Engineering practices(emphasis on Digital Signals Processing) .The methodology which will be followed in development of the Automatic Music Transcription is the Cross-Industry Process for Data Mining methodology(CRISP-DM) [13]. This is a structural approach to a data mining project, is adopted in this context to draw knowledge and insight from the obtained music dataset. Figure 1 provides a sequential phases of task to provide a robust, well-defined and accurate output for Automatic Music Transcription.

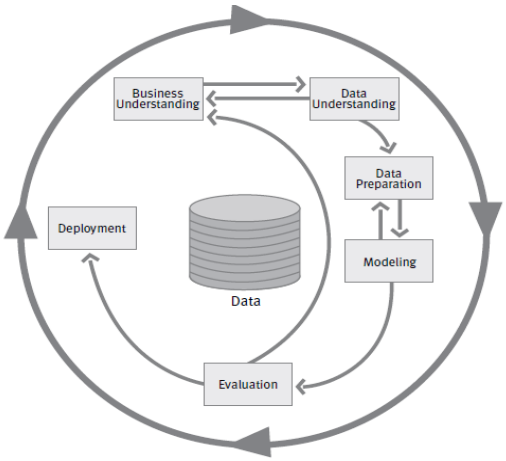

Fig. 1: The CRISP-DM Methodology

A. Business Understanding

Business Understanding is the most important phase in carrying out AMT as this is the skeleton which defines the fundamental processes that lay a foundation for the successive phases. This phase focuses on outlining the project specifications, objectives, assumptions, constraints to be adhered to and the relevant success criteria for successful Automatic Music Transcription. Furthermore, production of a project plan which discusses tools and techniques to be employed and assessing various contingencies i.e. conducting a Risk Assessment are crucial tasks during this phase.

B. Data Understanding

The second phase which is Data Understanding is more involved with the interactions with the data that will be used during AMT. This phase begins with the collection of music data and subsequent tasks involve exploration of the data by identifying patterns, describing the data and drawing insights which verify the quality of the data. For AMT, the dataset is assumed to be monophonic, therefore this phase focuses on accounting for such variables.

C. Data Preparation

The Data Preparation focuses on the tasks involved in preparation of the data for modeling techniques and tools to be used. This requires the processing and transformation of the raw data such that it is adequate for modeling. The tasks responsible for transformation and processing the data do not occur in any order as they may be repeated multiple times. For the purpose of AMT, this phase will consist of transformation of raw monophonic music dataset to produce spectrograms using VQT for modeling techniques.

D. Modeling

This phase focuses on selecting and applying modeling techniques on the transformed data from the preceding phase. Noting that there are several modeling techniques which are employed on the transformed data, the experimentation often requires iterating between data preparation and modeling. The modeling of the dataset is not enough for a robust and resolute output therefore other tasks involved include generating tests and possible model refinement techniques. The spectrograms resulting from VQT will be used as input fro the proposed model in Section VII.

E. Evaluation & Deployment

Evaluation is the preliminary phase of the AMT project at the point where the modeling would have been completed. It will focus on reviewing and assessing the results of the overall AMT process and verify if they are a reflection of the specifications, objectives, constraints and success criteria. An important task in this phase is conducting a critical analysis on the results from the preceding phases of experimentation. Deployment is the final phase of the life-cycle for the AMT process. Noting obtaining data preparation, modeling and evaluation are not the last stages of the AMT process. It is imperative that the results are organized in professional and scientific form, this will require documentation and a presentation. The deployment phase will also outline a review of the entire process to focus on reviewing the process outline future recommendations to the scientific flaws and shortcomings encountered.

V. Datasets

Music transcription can be employed to many different instruments. The proposed instruments to be transcribed are the piano and drums. The choice of these instruments was inspired by the abundance of their respective dataset and their significance in history of the music industry.
Drums are profound for their versatility in music as they are not confined to one genre. They are important as they can adapt to what music requires while producing unlimited tonal, melodic, rhythmic and harmonic shading. Drum sets have often been employed in genres such as Rock 'N' Roll, Jazz and the Blues. The Groove MIDI Dataset [14] is to be used for the automatic drum transcription. The dataset is composed of 1,150 MIDI files and over 22,000 measures of drumming including 13.6 hours of MIDI and audio human-performed, tempo-aligned drumming. Furthermore, the dataset was performed by 10 professional drummers with Roland TD-11 electronic drum kit who were inspired to be versatile and experiment with a wide range of playing styles to ensure the dataset is diverse.

The piano belongs to the keyboard family of musical instruments with stuck strings, and it offers a range of all 88 notes of the music scale which stands out from most instruments. Another interesting piano feature is that when a note is played, the musician has an option of releasing the key or playing it again while the note is still active [15]. MAESTRO (MIDI and Audio Edited for Synchronous Tracks and Organization) [16] dataset is employed for automatic piano transcription. This is a raw dataset contains over 200 hours of paired audio and recorded MIDI data from performances by virtuoso

pianists perform on Yamaha Disklaviers in the International Piano-e-Competition. Furthermore, The MIDI Data and paired audio are aligned by approximately 3ms. The Groove MIDI Dataset drum and the Maestro piano datasets both consists of recorded MIDI data from performances by professional drummers and pianists, respectively. The involvement of professional musicians in the creation of the dataset brings into question the veracity, accuracy and precision of the recordings. As stated by [17], It can be argued that a high level of accuracy can be guaranteed in the alignment of audio and MIDI data if the dataset were created by automated self-playing instruments regulated by a MIDI signal. However, discrepancies are bound to occur with the automated recordings as the Yamaha Disklavier incorrectly plays notes when the MIDI velocity decreases, as [18] reports up to 100ms in audio and MIDI data alignment errors due to automated recording.

VI. Digital Signals Processing

A. Audio Signal Representation

The Automatic Music Transcription of the Drums and Piano begins with the digital processing of audio signals. In general, signals are abstract therefore we model them mathematically to study their behaviour in the time and frequency domain. For AMT, the aim of the digital signal processing of audio signals is to provide a spectral representation of the signals in frequency domain in the form of spectrograms. A spectrogram is 3D matrix which represents frequencies of a signal in variation with time[15], and different frequency magnitudes are represented in a variety of colors. The proposed digital signals technique used to obtain spectrograms is the VQT which is a modification of the Constant CQT by introducing a parameter γ [19]. Therefore, it is imperative to have a primitive understanding of the CQT to draw the distinction its from the VQT for Automatic Music Transcription.

The Constant Q-Transform is audio signals processing technique which is suited for music transcription[15] is as it offers well defined low frequency spectral representation rather than at high frequencies with the added benefit of efficiency[20]. The CQT is interpreted as a filter-bank, where the filters banks are geometrically spaced by centre frequency (f_{k}) as indicated in Equation 2, and f_{min} is defined as the first center frequency. The bandwidth of the $\mathrm{k_{th}}$ filter is defined by Equation 1

$$B_k = 2^{\frac{k}{n}} B_{min} \tag{1}$$

where n is the number of octaves per filter and B_{min} is the first bandwidth of the first filter bank. Noting that the centre frequencies(f_k) of the filters are geometrically spaced, the centre frequency of the $\mathrm{k_{th}}$ filter is given by Equation 2

$$f_k = f_{min} 2^{\frac{k}{n}}; k = 0, 1, ... \tag{2}$$

where f_{min} is the first centre frequency and n is the number of octaves per filter. The CQT weights have constant equal Quality-factors(Q) which can be represented as a ratio of the centre frequency(f_k) to the Bandwidth(B_k) in Equation 3

$$Q = \frac{f_k}{B_k} = \frac{1}{2^{\frac{1}{n}} - 1} \tag{3}$$

Therefore, the window size(N) of each k filter can be represented in terms of the Q-factor as shown by Equation 4, where f_s is the sampling frequency which is defined in Section VI-A.

$$N[k] = \frac{f_s}{B_k} = Q\frac{f_s}{f_k}; k = 0, 1, ... \tag{4}$$

As the stated before, The VQT is a modification of the CQT with the aim of producing a much better spectral representation for AMT. The modification of CQT into VQT occurs with the introduction of the parameter γ to maintain the a constant Q-factor and decreases it at high and low frequencies, respectively. Furthermore, for CQT representation, $\gamma = 0$, and VQT ,$\gamma > 0$ and its value varies depending on the application, noting that large values of γ significantly increase the time resolution at lower frequencies[21]. Figure 2 and Figure 3 are CQT and VQT spectrograms resulting from the experimentation in [22]. A comparison between the spectograms denote a much better spectral resolution at low frequencies for the VQT representation.

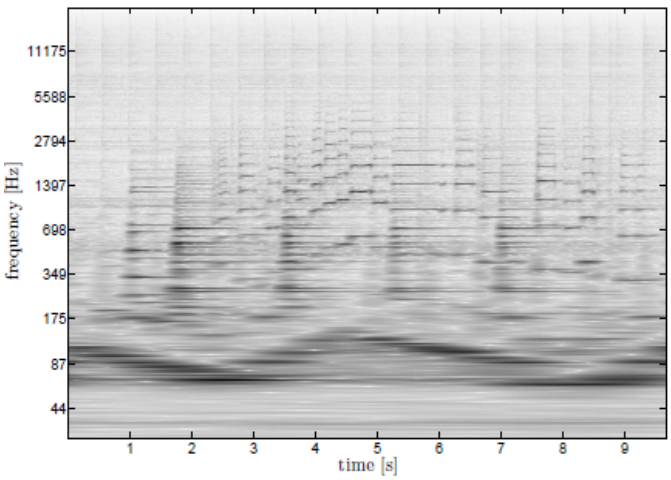

Fig. 2: Constant Q-Transform spectrogram $\gamma = 0$

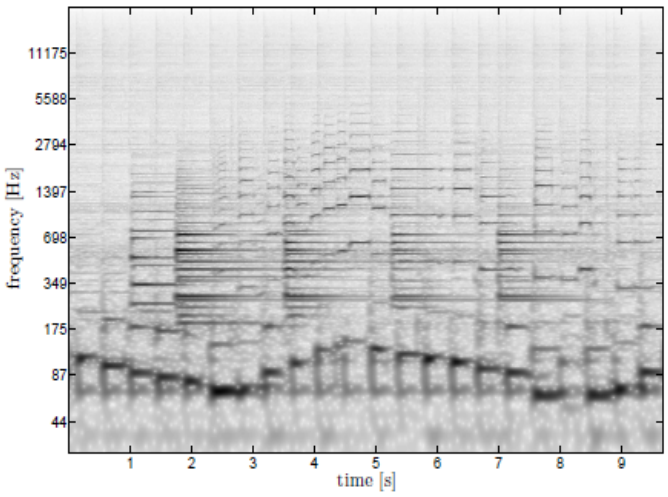

Fig. 3: Variable Q-Transform spectrogram $\gamma = 20$

The bandwidth(B_k) for VQT time-frequency analysis is given by Equation 5, with the additional parameter γ, the centre frequency(f_k) for the k_{th} filter and n the number of octaves per filter.

$$B_k = \alpha f_k + \gamma \tag{5}$$

where

$$\alpha = 2^{\frac{1}{n}} - 2^{-\frac{1}{n}}$$

The Quality factor(Q) can also be represented in terms of γ for the VQT representation using Equation 6 below:

$$Q = \frac{f_k}{(2^{\frac{1}{n}} - 1)f_k + \gamma} \tag{6}$$

The computation of VQT to produce the spectrograms will be achieved using `Jupyter Notebook` & `Librosa`[23] which is a Python package for audio and music digital signal processing. `Librosa` contains predefined computation of CQT which will be modified to introduce γ for VQT representation. However, some parameters require to be modified and redefined for the VQT computations.

Parameters:

1) f_s: Sampling rate/frequency of each filter of window size N such that the Q-factor is satisfied. The sampling frequency is 22050Hz [23].
2) n_{bins} - number of frequency filter banks which determines the maximum frequency of the VQT.
3) n: number of filter banks per octave which influence the frequency resolution of the VQT. Typical values are 12, 24, 36 & 48, however, a suitable on will have to selected during experimentation[15].
4) N(Window Size): the resolution of the VQT spectrogram is dependent on the window size. A variable window size as described by Equation 4 ensures is ideal to obtain high spectral resolution at lower frequencies and high temporal resolution at high frequencies [24].
5) f_{min}(minimum frequency): the minimum frequency will be the first center frequency of the filter bank. According to [15] it should be as low as the frequency of the first notes such that the spectrogram should show the fundamental frequency and the neural networks recognizes the notes.

VII. Proposed Model

The proposed model is based on the work by [25],[19],[26] and the Kelz architecture [15][27]. The model consist of sub tasks which perform different tasks in the training stage namely; onset, offset, velocity and frame prediction as depicted in figure 4. Different machine learning libraries will be explored for the implementation of the model `PyTorch, Keras, TensorFlow` and `matlab machine learning toolbox`. Ultimately the library with the most support, easy of use and less development time will be selected.

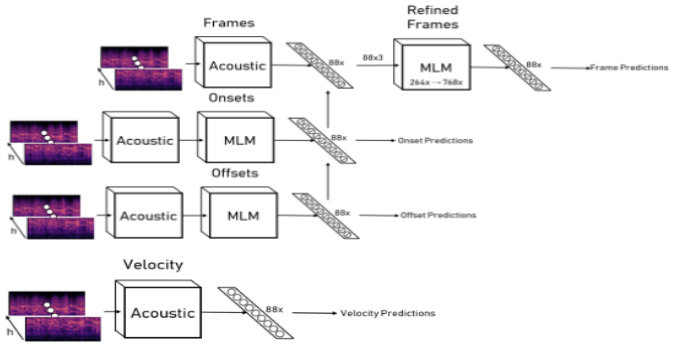

Fig. 4: Transcription model, where h represents the total number of frames

A. Acoustic model

The model consists of a Convolutional Neural Network (CNN) and a post processing unit which is to be determined through experiments. The spectograms produced by VQT are used as input images to the CNN. The spectograms have dimensions $t \times f_s$ where t is the duration of the song in seconds and f_s is frequency (Hz). Unlike normal object images, music is an ordered sequence which means spectograms cannot be transformed to different configurations. The first operation to be done is to convolve a region of the spectrogram with a filter. The filter is a matrix of weights. In this context convolution is an element wise dot product between the considered region and the filter. Each filter has a

corresponding bias matrix which is summed with the convolution operation result to give the output from the convolution layer. The filter and biases are then optimised with each forward pass. The filter shape is structured such that is captures maximal note information (fundamental frequency and corresponding harmonics), as such the filter shape for the proposed model spans the frequency axis and has stride equal to the tone duration [28].

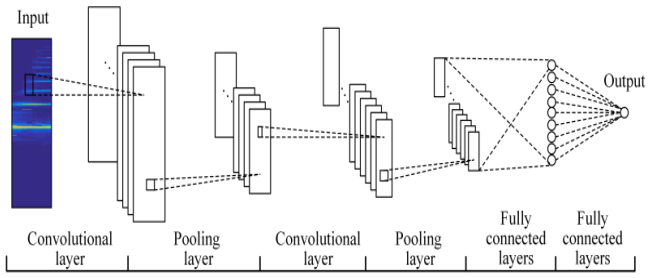

Fig. 5: Acoustic model

An activation layer is introduced between the convolutional layer and the pooling layer. This layer is added to decrease over-fitting between features by introducing non linearity. The activation function used is the Rectified Linear Unit (ReLu) to produce and activation map. The advantage of using this function compared to activation functions such as sigmoid and tanh is that the ReLu converges faster and does not require input normalization because the output is guaranteed to be between 0 and the maximum value x. The ReLu function is defined as:

$$relu(x) = max(0, x) \tag{7}$$

A pooling layer is applied to perform a non-linear down sampling of the activation map. The pooling performed should be such that it reduces the activation map whilst maximising information gain. Usually pooling is done through a statistical parameter namely the average or maximum value. Max pooling deemed more appropriate for this application compared to average pooling. This is because max pooling ensures that the dominant feature is recorded, the problem with average pooling is that the average might be feature that is not dominant in that region. This is important for this application because we want to capture the most dominant note per frame.

Another method to avoid over-fitting is to introduce a drop out which randomly disconnects neurons between layers. This layer is added in between the pooling layer and activation layer. This step is only introduced in the training stage.

A binary cross-entropy loss function is used to measure the difference between the predicted values $\hat{y_n}$ and ground truth y_n. The loss function is defined below, where N represents the number of labels.

$$L = \frac{1}{N}\sum_{n=0}^{N} y_n log(\hat{y_n}) + (1 - y_n) log(1 - \hat{y_n}) \tag{8}$$

The loss function is continuously optimised through gradient descent and is used to update the weights and biases through back propagation until it reaches diminishing results.

VIII. Acoustic Topology

The topology presented is a first iteration approximation design of the training model. Experiments are going to be carried out in order to determine the most optimal parameters.

A. Filter

Various filter sizes will be evaluated those presented in literature and others that appropriate for the selected instruments. Filter sizes to be explored from literature are $\{f_s \times window\_size/2k, 3 \times 3, 5 \times 5\}$ where k is the number of bins per octave and $window\_size$ being the time slice.

B. Learning Parameters

The learning rate is an important parameter used during back propagation in gradient descent. It specifies the amount by which the network weights from be updated. Setting the parameter too low may result in the taking too long to converge to a local minima and setting it too high may result in the model never converging because it would have missed the local minima. It is therefore important to increment this parameter at a reasonable rate.

Dropout is a less computational demanding form of regularisation. Different keep rates used in literature for the dropout layer are to explored specifically $\{0.2, 0.25, 0.5\}$. A threshold value of 0.5 is used for the activation function, where $\hat{y_n} = y_n > 0.5$.

IX. Post processing unit

The MLM is appended at the end of each acoustic model in order to smooth out the output and capture temporal dependencies between frames such as how notes evolve over time. The output from acoustic model is be used as the input for the MLM with the exception velocity prediction. Prediction velocity is a standalone task, velocity takes into account the speed and loudness of a note [29]. The MLM can be implemented using different

techniques namely Recurrent Neural Network (RNN), Bidirectional Long-Short-Term Memory (BiLSTM) or Hidden Markov model (HMM) [30]. The BiLSTM has been shown to be superior compared to the RNN and HMM, nonetheless different variations of the MLM will still be evaluated to get the optimal one. The idea is to follow the Google Brains Onset and Frames Network and Kelz baseline model whereby the results from the sub tasks are appended together in order to get the final frame prediction as depicted in figure 4 In this model a ReLu is after the BiLSTM instead sigmoid in mentioned models.

X. Testing and validation process

The testing stage is used to optimise and fine tune model hyper parameters in order to improve accuracy and performance. The testing stage includes using unseen data in the training stage. This prevent the model to being over fit to a particular dataset.

The model will probably need to be on the trained cloud because of the complex and costly training process. The training process also needs a lot of data which may not be feasible stored on a single machine. Each song is split into defined window sizes and produce a spectrogram for each window. Fine tuning the model may also be time consuming task. Possible cloud providers are google, aws and azure which mainly provide a Python SDK.

A. Metrics

The confusion matrix, word error rate, character error rate, precision(P), recall(R) and f-measures(F1) are some of the common performance metrics used for machine learning models. The metrics to be used to evaluate the model are precision, recall and f-measure. This step also makes use of new unseen data in both the training and testing stage. F1 is a ratio of the total number of errors compared to the number of detectable notes in the music. The equations below are used to calculate the mentioned metrics.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2RP}{R + P} \tag{11}$$

Where TP, FP and FN represent true positive, false positive and false negative respectively. The `mir_eval` library provides and implementation to evaluate the stated metrics [31]. The onset and offset predictions are considered a TP within a ± 50 ms tolerance of the ground truth. Each frame correctness is verified per 10 ms [15].

XI. User interface

The user interface will provide the user a graphical means of interacting with the AMT system. The particular choice will be dictated by the language used to implement the AMT model. PyQt5, Tkinter or matlab gui. Both PyQt5 and Tkinter are Python wrapper which provide a framework for GUI development in Python. The GUI will provide a window in which a user will select a song to be transcribed. The transcription processes will occur in the same window. The output representation will be at a note level.

XII. Project Management

This section outlines the work breakdown assignment and the project schedule. The scope of the AMT system has two main components which digital audio signals processing and modeling using neural networks. This makes the work breakdown schedule of the tasks convenient as each member on the group can focus on one main component at the time while sharing common tasks and constantly helping each other in efforts of improving productivity, team moral and producing an immaculate output. The high level representation of the work breakdown assignment is registered in Table I. The project duration for the AMT project is 8 weeks. The system development will be guided by the project schedule represented by the Gantt chart on Figure 1 in Appendix A. In the Gantt chart the critical tasks are indicated in red, noting that it will define the shortest duration to complete the project. Furthermore, team members are to commit themselves to working from 8am to 5pm weekdays, however, weekends will also be used to review the work done during the week, complete the tasks which were not done and plan/work ahead for the successive week. Prior to the project commencement team members understand that only good teamwork, conflict resolution, constant effective communication by the form of weekly meetings, hard work and determination will guarantee project success. The project will be done in stages in an iterative manner. The project is broken down to consist of milestones to be completed per iteration. The development process consist of design, prototype, test and deployment stages.

TABLE I: Suggested Work Breakdown

Tasks	**Sbonelo**	**Matthews**
AMT research, specifications and planning	X	X
Dataset Collection		X
Digital Signals Processing		X
Modelling, Testing & Validation	X	
User Interface	X	X
Documentation & Presentation	X	X

The analysis of the contingencies that might occur is within the scope of the fundamental stages of the CRISP-DM methodology described in SectionIV. The risk assessment is registered on Table II in Appendix B.

XIII. Conclusion

In this paper we presented the project plan outlines for the AMT system. The system consists of multitude of tasks which form the complete product. The system is designed to transcribe 2 musical instruments viz. the Drum and Piano using the he GrooveMIDI and the MAESTRO Datasets, respectively. From the literature explored, The Variable Q-Transform outperforms the Constant Q-Transform hence it is the preferred audio signals processing technique to produce spectrograms which are fed into neural networks.The proposed model consist of an acoustic model and MLM unit. These are the essential parts used in training the model. A CNN is used for the acoustic model and HMM, BiLSTM or RNN are some of the options to be explored for the MLM.

REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[2] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[3] L. Su and Y. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1600–1612, 2015.

[4] P. H. Peeling, A. T. Cemgil, and S. J. Godsill, "Generative spectrogram factorization models for polyphonic piano transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 519–527, 2010.

[5] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 138–150, 2014.

[6] V. Arora and L. Behera, "Multiple f0 estimation and source clustering of polyphonic music audio using plca and hmrfs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 278–287, 2015.

[7] M. Bereket, "An ai approach to automatic natural music transcription," 2017.

[8] R. G. C. Carvalho and P. Smaragdis, "Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 151–155.

[9] S. Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 327–337, 2008.

[10] R. A. Dobre and C. Negrescu, "Automatic music transcription software based on constant q transform," in *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2016, pp. 1–4.

[11] C. Schörkhuber, "Constant-q transform toolbox for music processing," 2010.

[12] C. Marghescu and A. Drumea, "Modelling and simulation of energy harvesting with solar cell," 02 2015, p. 92582L.

[13] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01 2000.

[14] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, "Learning to groove with inverse sequence transformations," in *International Conference on Machine Learning (ICML)*, 2019.

[15] M. Karioun and S. Tihon, "Deep learning in automatic piano transcription."

[16] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[17] A. C. Jaedicke, "Improving polyphonic piano transcription using deep residual learning," June, 2019.

[18] S. Ewert and M. Sandler, "Piano transcription in the studio using an extensible alternating directions framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1983–1997, 2016.

[19] Cwitkowitz and C. Frank, "End-to-end music transcription using fine-tuned variable-q filterbanks," 2019.

[20] R. A. Dobre and C. Negrescu, "Automatic music transcription software based on constant q transform," in *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2016, pp. 1–4.

[21] E. Benetos and S. Dixon, "Multiple-f0 estimation and note tracking for mirex 2012 using a shift-invariant latent variable model."

[22] C. Schrkhuber, A. Klapuri, N. Holighaus, and M. Drfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Jan 2014. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17112

[23] C. R. D. L. D. P. E. M. M. E. B. McFee, Brian and O. Nieto, "librosa: Audio and music signal analysis in python," in *In Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[24] K. O. U. T. M. Conforto Silvia, Nisar Shibli, "An efficient adaptive window size selection method for improving spectrogram visualization," in *Computational Intelligence and Neuroscience*. Hindawi Publishing Corporation, 2016. [Online]. Available: https://doi.org/10.1155/2016/6172453

[25] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic music transcription," *ArXiv*, vol. abs/1508.01774, 2015.

[26] M. Mnguez Carretero, "Automatic music transcription using neural networks," 2018-07-02.

[27] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," *CoRR*, vol. abs/1612.05153, 2016. [Online]. Available: http://arxiv.org/abs/1612.05153

[28] J. Sleep, "Automatic music transcription with convolutional neural networks using intuitive filter shapes," 2017.

[29] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *CoRR*, vol. abs/1710.11153, 2017. [Online]. Available: http://arxiv.org/abs/1710.11153

[30] B. S. Gowrishankar and N. U. Bhajantri, "An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016, pp. 140–152.

[31] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

APPENDIX

A. Project Schedule

Fig. 1: Gantt Chart

B. Risk Management

The table below indicated the risk and its associated attributes. The risk register is used to enure that the project is delivered will less hindrance.

TABLE II: Risk register

Risk	**Probability**	**Impact**	**Response**	**Action**
Interruption of training process	High	High	Avoid	Ensure machine is always charged
Data loss	High	High	Avoid	Use redundancy (cloud storage and external hard drive)
Computational intensive training	High	High	Mitigate	Cloud training or model refinement
Conflict	Low	High	Solve	Effective communication, Confront the conflict & find common solution