MUSIC TRANSCRIPTION WITH MULTILABEL CLASSIFICATION

Nora Huang
Department of Computer Science
University of Victoria
Victoria, B.C, Canada
norah@uvic.ca

Aazim Lakhani
Department of Computer Science
University of Victoria
Victoria, B.C, Canada
aazimlakhani@uvic.ca

Parul Parul
Department of Computer Science
University of Victoria
Victoria, B.C, Canada
pparul@uvic.ca

ABSTRACT

Automatic music transcription is an active research topic in the field of Music Information Retrieval (MIR). The goal of music transcription is to determine the score-like representation of the input audio music. The technologies of it have been used in many mobile and web applications. However, the accuracy of the current transcription approach is still below that of a human expert. The majority of current research is focused on how to separate the pitches from the audio, then estimate them. In this project, time and frequency domain features for piano music files were extracted. The data was filtered through normalization, and feature subsets with maximum entropy and correlation were selected for use in well known classifiers such as Decision Trees, Random Tree, Nearest Neighbor, and SVM. This process enables the project model to detect musical notes which could be used in automatic music transcription.

1. INTRODUCTION

Automatic music transcription is the process of translating an audio music signal into a musical score sheet[1]. There are many applications of this technology, such as helping a student learn how to play music for well known soundtracks using a specific instrument. It could do so by building playable score sheets they could use for practice. It could also allow a musician or composer to record their improvised performance [2] and share their work for others to reproduce. There are various approaches to achieve these goals [3] [7].

An ideal piece of music for transcription has clear, simple, and easily differentiable sounds. However in reality, song compositions are complex pieces of work with many different instruments and human voices. Addressing this complexity is a significant challenge. This project focuses on the problem of monophonic and polyphonic piano music for up to 4 notes. Monophonic includes a single melodic line with no accompaniment, while polyphonic may have multiple melodic lines mostly independent or in imitation of one another. In other words, this project focuses exclusively on music composed with up to 4 notes played simultaneously.

This project builds a unique model to recognize keys used to played music notes to support music transcription. First, windows are applied to the input music files to segment the audio into frames around 11ms in length. In total, it retrieves 512 digital samples per frame. In total, 63 features are extracted from each frame. These acoustic features are post-processed by removing silent frames or values and labelling each non-silent frame, then analysed with either Weka [8] and Meka [9] to reduce the set into several reasonably sized feature subsets. The selected features are those having high entropy or co-relation, which are then passed through well known classifiers to prepare a model.

Weka is used to analyse monophonic music which can simplified as the single-label classification problem. Conversely, Meka is a multi-label classification extension of Weka to analyse polyphonic music, equivalent to a multilabel classification problem. Polyphonic music exposes the challenge of detecting the right keys used to play a chord when several others notes from potentially many different instruments are played simultaneously. This project reduces the scope of the problem by focusing on single-instrument audio tracks, leaving only the problem of detecting each note in a chord. In total, Multilabel classification [10] is applied using 4 classifiers: IBK [11], J48 [12], SMO [13], and RandomTree [14].

The overview of the project model is described in Figure 1. The classifiers are trained and tested based on this model.

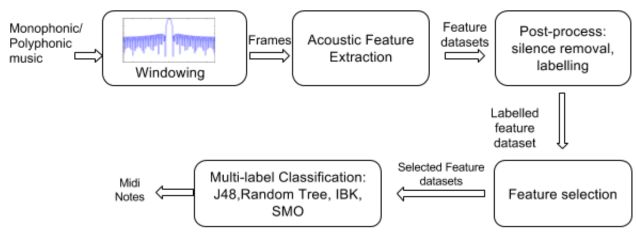

Figure 1. System Flow Diagram

The structure of the report is as follows: first, the dataset for the project is described. The tool and methods used for feature extraction are then described, along with the meaning of each set of features. Next, feature selection describes how the feature subsets were derived, analysed, and grouped. The classification section details the classifiers used and the parameters of each for every feature subset, including details on how they were trained and tested. The results of all the experiments are then evaluated and visualized for each dataset. Finally, the results are concluded and future work is outlined.

2. DATASET

The model described in Section 1, Figure1 is evaluated using the MIDI (Musical Instrument Digital Interface) Aligned Piano Sounds (MAPS) dataset [15]. Due to copyright licensing, there are a limited number of datasets available for research. MAPS is a widely used piano database for multipitch estimation and automatic transcription of music. It provides recordings with CD quality (16-bit , 44 kHz sampled stereo audio) sounds. It is freely available under the Creative Commons License. It consists of audio files with corresponding annotations for isolated sounds, chords, and complete pieces of piano music under various different recording conditions, including various closed rooms and close ambient takes. MAPS can be divided into 4 sets:

ISOL (isolated notes and musical excerpts)
RAND (Chords with random pitch notes.)
UCHO (Usual chords from Western music)
MUS (Pieces of piano music)

For monophonic music, we use 2 second notes from the ISOL dataset. This is defined as P1 for the remainder of this report. For polyphonic music, we use polyphonic sounds with 2 to 4 notes from the RAND set. They are similarly defined as P2, P3, and P4 for the remainder of this report.

3. FEATURE EXTRACTION

In the signal processing domain, feature extraction is one of the most widely used methods to analyse a signal. There are several parts to one audio signal offering varying amounts of useful information. This section explores some of these features and selects some for use based on how much they contribute to the goal of automatic music transcription.

3.1 Feature Extraction Tool

Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) is an open source software framework for audio processing with specific emphasis on MIR applications [21]. Bextract is one of the most powerful executables provided by Marsyas. It can be used for complete feature extraction and classification experiments with multiple files. However its classification feature is limited and did not fulfill the requirement of this project. As such, only the feature extraction functionality is used.

3.2 Feature Extraction Strategy

Using Bextract provided, the features for each frame are extracted. By default, the frame size of Bextract is 512 samples which matches the dataset sampled from MAPS. In total, it extracts 63 features from each frame. The results are formatted and stored in .arff format file for later processing. Bextract further handles the labeling for each item by using a single label - it does not support multi-labelling. As such, for the polyphonic music requiring multi-labeling, they are manually labeled with multiple values corresponding to the notes it consists of. This is done by post-processed the output .arff files with a script uniquely created for this project. This enables the data to be run directly on Meka. The script derives the labels for the output from the corresponding .txt annotations provided with the sound from MAPS.

3.3 Acoustic Features

From each frame, 63 acoustic features were extracted. They are combinations of the following 5 parameters:
Time-domain Zero-Crossings is the point where the sign of a mathematical function changes (e.g., from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function [16]. It is a commonly used term in electronics, mathematics, sound, and image processing.
Spectral Centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound [17].
Rolloff is the steepness of a transmission function with frequency [18].
Flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame [19]. More precisely, it is usually calculated as the 2-norm (also known as the Euclidean distance) between the two normalised spectra.
Mel-Frequency Cepstral Coefficients are coefficients that collectively make up an mel-frequency cepstrum (MFC) [20]. In sound processing, the MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

4. FEATURE SELECTION

The feature set produced by Bextract, while effective, is a large feature set making human understanding of the data difficult. Feature selection was needed in order to reduce the usable attribute set from 63 into a more reasonable size. This section outlines the approach and result of feature set selection.

4.1 Filter-based Selection

There are multiple techniques that may be used to reduce a feature set, including using domain knowledge of the

data to manually reduce it, tailoring the set for a specific classifier, or using a systematic filter-based approach [22]. Due to the complexity of the notes and their acoustic features, and the desire to have a general feature set for use in any classifier, the systematic approach was used to decide which features were most useful.
Three approaches were used to create 5 feature sets for each PN dataset: 1) learner, 2) entropy, and 3) pearson correlation coefficient. The accuracy results of each feature set was compared against the original for 4 different classifiers in 5-fold cross validation. In order to create a standardized set selection across the datasets (i.e., they all have the same thresholds and features), the P1 dataset was selected as the baseline. That is, the feature subsets generated for P1 were assumed to be sufficient for P2 - P4.

4.2 Results

WEKAs select attributes toolset [23] was used to generate data for each approach. Two feature subsets were created from the entropy results; any attribute that provided an information gain greater or equal to 0.05 was used in the first entropy feature data set. A second set was created using a threshold of 0.05. Similarly, the pearson result was split into two feature subset using threshold values of 0.03 and 0.02. The threshold values were chosen by observation as they provided either approximately half the original feature set or slightly less which was agreed to be a reasonable attribute set size. Table4.2 summarizes the results.

Table 1. Feature subset selection attribute size results

Filter Algorithm	Threshold	Attributes Size	Reduced %
None	N/A	63	–
Learner	N/A	28	66
Entropy	0.05	14	77
	0.03	31	51
Pearson	0.03	27	57
	0.02	36	43

5. PREPROCESSING WITH NORMALIZATION

Normalization scales all numeric variables in the range [0,1] to ensure the outputs are scale invariant. In other words, it makes sure big values dont dominate smaller ones. One possible formula is given below:

$$\chi_{new} = \frac{\chi - \chi_{min}}{\chi_{max} - \chi_{min}}$$

Normalization was used on the P1 original dataset (i.e., with all 63 attributes) to determine if this technique could improve the accuracy of the results. As seen in Table5, the results of normalization did not significantly improve the accuracy of 4 different classifiers. This result meant that normalization was not applied, and was not attempted for the remaining P2-P4 datasets.

Table 2. Accuracy Results Compare (%)

	IBK	J48	SMO	Random tree
Without Normalization	97.98	88.98	54.75	87.65
With Normalization	97.98	88.98	54.76	87.67

6. MULTILABEL CLASSIFICATION

In MIR, multi-label classification are usually used to identify genres or emotion recognition, but are not used in transcription. Current research into automatic transcription tends to focus only on either the main melody, or on the approach by separating the multiple melodic lines and processing them individually. This project explores multilabel classification for polyphonic transcription by treating each note from the multiple melodic lines as a label, where each label is one of 88 classes corresponding to 88 piano keys (i.e., 11 octaves of 8 notes).
Using the feature attribute subsets derived in Section 4, four classifiers were used to experiment with results on all datasets from P1 to P4. The four classifiers were: 1) IBK, 2) J48, 3) SMO, and 4) RandomTree. Training and testing was executed in 5 fold cross-validation on each dataset. This section presents the raw data results and some visualization of the data.

7. EVALUATION

This section presents the results of running the training and testing classifiers on the P1 dataset in Weka, and the same approach for P2 - P4 in Meka. All datasets use the feature subsets derived in Section 4, and each dataset is separated into its own subsection.

7.1 P1 Monophonic Dataset Evaluation

This section presents the results of the first dataset, provides visualizations of the J48 tree, and analyses the J48 confusion matrix. J48 was chosen for further analysis as it was quick to reproduce on any machine while not sacrificing accuracy.
Feature subset results
Table 7.1 summarizes the result of each classifier on each feature subset. Anecdotally, SMO took the longest amount of time to run (approximately 5-10 minutes) and was assumed to have the worst performance, validated by its low correct percentage classified results. IBK took a moderate amount of time to run while providing the highest accuracy. RandomTree was slightly faster than J48 but produced consistent lower accuracies.
J48 Analysis
It was found that the initialization of all J48 trees for each feature subset was identical. That is, each set of attributes derived the same root node and first layer of their J48 classifier. However, to emphasize the differences between the

Table 3. Accuracy results of different feature subsets on different classifiers (%)

Filter (Threshold)	IBK	J48	SMO	Random tree
Original	97.98	88.98	54.75	87.65
Learner	97.37	88.54	46.11	87.63
Entropy (0.05)	91.48	87.08	27.88	86.87
Entropy (0.03)	98.61	88.59	51.86	87.83
Pearson (0.03)	99.41	87.27	48.96	87.26
Pearson (0.02)	99.14	88.10	48.58	87.86

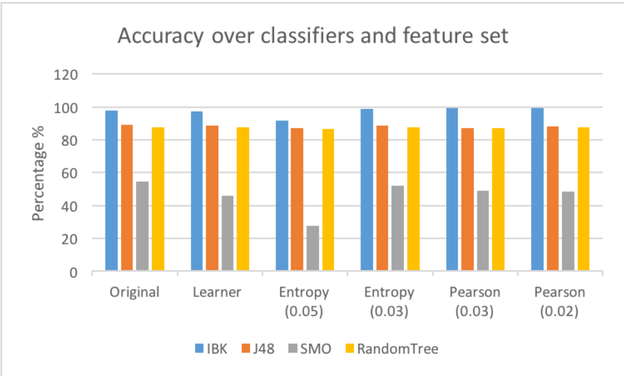

Figure 2. Accuracy for each dataset and classifiers for P1 datasets

feature set trees, Figures 5 through 10 depict the general structures of each.

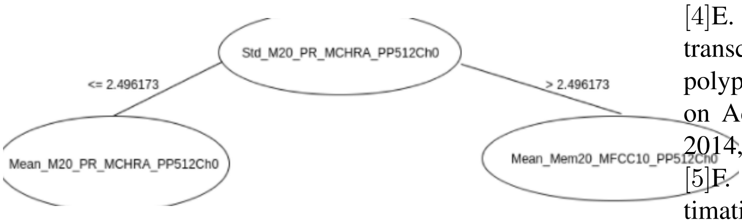

Figure 3. Root Node and First layer for all P1 instances

The confusion matrix of each feature set was reviewed using the matlab script included in Appendix A. In sum, the script found the intersection of notes with the highest misclassified percentage value. Table 5 summarizes the results.

It is not surprising that the feature sets have similar misclassified note intersections. It is reasonable to expect that similar sounding notes would be consistently misclassified no matter which subset and classifier were used. Figure 11 is a visual comparison of the results.

Table 4. P1 Feature subset highest confusion results

Filter (Threshold)	Intersection	Notes
Original	(m28, m27)	55 / 3420
Learner	(m28, m27)	66 / 3420
Entropy (0.05)	(m23, m24)	73 / 3074
Entropy (0.03)	(m28, m27)	75 / 3420
Pearson (0.03)	(m22, m21)	72 / 3370
Pearson (0.02)	(m28, m27)	61 / 3420

7.2 P2 Dataset

7.3 P3 Dataset

7.4 P4 Dataset

7.5 Exact Match Accuracy Comparison for PN Dataset

8. CONCLUSION

9. FUTURE WORK

10. REFERENCES

[1]O. Abraham, Suggested Methods for the Transcription of Exotic Music. .

[2]E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, Automatic music transcription: challenges and future directions, Journal of Intelligent Information Systems, no. 10.1007/s10844-013-0258-3, pp. 128, Jul. 2013.

[3]E. Benetos and S. Dixon, A Shift-Invariant Latent Variable Model for Automatic Music Transcription, Computer Music Journal, vol. 36, no. 4, pp. 8194, 2012.

[4]E. Benetos, S. Ewert, and T. Weyde, Automatic transcription of pitched and unpitched sounds from polyphonic music, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 31073111.

[5]F. Elvander, J. Swrd, and A. Jakobsson, Online Estimation of Multiple Harmonic Signals, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 2, pp. 273284, Feb. 2017.

[6]S. Ewert, M. D. Plumbley, and M. Sandler, A dynamic programming variant of non-negative matrix deconvolution for the transcription of struck string instruments, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 569573.

[7]C. T. Lee, Y. H. Yang, and H. H. Chen, Multipitch Estimation of Piano Music by Exemplar-Based Sparse Representation, IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 608618, Jun. 2012.

[8]H. Ming, D. Huang, L. Xie, and H. Li, Learning optimal features for music transcription, in 2014 IEEE China Summit International Conference on Signal and

Information Processing (ChinaSIP), 2014, pp. 105109.
[9]F. Pishdadian and J. K. Nelson, On the transcription of monophonic melodies in an instance-based pitch classification scenario, in 2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), 2013, pp. 222227.
[10]Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music, ResearchGate.
[11]Classification-Based Music Transcription - PDF. [Online]. Available: http://docplayer.net/39879732-Classification-based-music-transcription.html. [Accessed: 28-Feb-2017].
[12]Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription (PDF Download Available), ResearchGate.