Hidden Markov Model Harmonisation on Vocals and Instrumentals

Dimuth Kulasinghe

Project Final

1 Introduction

1.1 Problem Designation

Harmonisation is the process of constructing an accompanying line of music for a given melody. For this project, I consider this line to be a sequence of musical chords, or chord progression. While the definition of a song is open-ended, we can generally distinguish its vocal melodies from the instrumental melodies. It is a known concept in music theory that vocals tend to incorporate a smaller range of pitches than instrumentals, as well as less dramatic changes in pitch. I thus hypothesize that, due to their more consistent melodic sequences, a harmoniser will be more effective in predicting the underlying chords for vocal melodies than instrumental melodies.

1.2 Background Theory

The Hidden Markov Model (HMM) is a common model used in harmonisation [1]. HMM is a 5-tuple $\langle \mathcal{S}, \mathcal{O}, \pi, A, B \rangle$, where

- $\mathcal{S}$ is a set of states
- $\mathcal{O}$ is a set of observations
- π is a vector depicting the initial probabilities of each state
- A is a transition matrix depicting the probabilities of moving from state to state
- B is an emission matrix depicting the probability of an observation from a given state

The intuition of HMM is that the world can only see the observations of the model, and not the states themselves (though the world is aware of the underlying definitions and transition probabilities). Thus, given a sequence of observations $(o_1, ..., o_n), o_i \in \mathcal{O}$, we can use HMM to predict the underlying state sequence $(s_1, ..., s_n), s_i \in \mathcal{S}$. By considering $\mathcal{S}$ to be the set of possible chords, and $\mathcal{O}$ the set of possible melodic sequences, we can generate a chord progression for a given melodic sequence.

The Viterbi algorithm can be used to generate the predictions based on HMM [2]. For an input sequence $Y = (y_1, y_2, ...y_T)$, it outputs a state prediction sequence $X = (x_1, x_2, ...x_T)$ by using transition tables T_1, T_2 to depict the most likely state sequence so far.

2 Approach

2.1 State and Observation Definitions

For simplicity, I restrict the state set in my model to the distinct major and minor chords in each key, giving a total of 24 possible chords. This also helps in determining the accuracy of the model, as there is not a significant tonal difference, for example, in Cmaj as opposed to Cmaj7. I define an observation as a multiset of musical frequencies (and rests) that sum up, in duration, to a half measure in 4/4 time, with the base unit of measurement being a sixteenth note. For example, the following excerpt

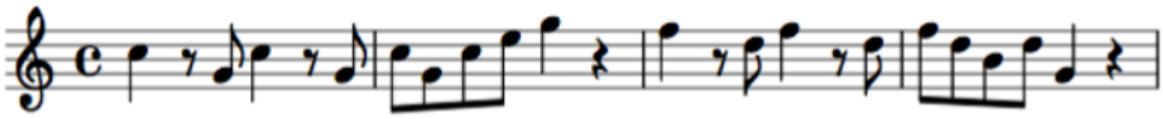


Figure 1: Melody from Eine Kleine Nachtmusik by Mozart, tranposed to Cmaj

would be represented by the following sequence of observations

`{C5:4, G4:2, R:2}, {C5:4, G4:2, R:2}, {C5:4, G4:2, E5:2}, {G5:4, R:4},`

`{F5:4, D5:2, R:2}, {F5:4, D5:2, R:2}, {F5:2, D5:4, B2:2}, {G4:4, R:4}`

and the underlying chord progression `C,C,C,C,G,G,G,G`. Hence, I disregard the order of the notes, as well as the individual notes composing each notes' total duration; I only care about the *proportionate contribution* of each pitch to the half measure's duration. I do, however, take into account the scale of the notes, differentiating between the octaves separating the `C` notes, for example. These steps are taken to reduce the dimensionality of the observation space while retaining the information regarding pitch.

2.2 Data Collection

The specific notation used by my experiment requires me to manually document any songs I use. I decided to document a set of Sinhalese songs from an online database [3]: these songs, while rich in sound, are methodical in their underlying chord progressions and clearly distinguish between vocal and instrumental melodies, making them an ideal case set to work on. All songs are transposed to the key of C major.

2.3 Data Representation

Songs are represented as their individual vocal and instrumental sequences, with the underlying chord progressions for each. The two types of sequences of processed independently, resulting in two HMM models that are used for predictions.

The A matrix for each is represented as a simple `numpy` matrix. The B matrix is represented as a list of the probabilities at the non-zero coordinates, as it is far too large to record formally. There is the inevitable issue of data sparsity with regards to the observation space; to account for this, I use the method from [4] and declare $\mathcal{O}$ as the set of *recorded* observations, and for an input observation g use

$$\text{Observation} = \arg\min_{o \in \mathcal{O}} \left(\text{jaccard}(g, o)\right),$$

where jaccard() is the Jaccard distance between the observations. The model will thus always run on an observation for which data has been collected. Finally, the effectiveness of each model is measured as the percent of chords it accurately predicts on a set of sample songs, which are documented the same way.

3 Results

I measure the results for the vocal and instrumental sections with `Normal`, `NoScale`, and `NoTranspose` models. For example, in the `NoScale` model, the data points {`C5:4, G4:2, R:2`} and {`C2:4, G1:2, R:2`} are equivalent. This is used to further investigate the effect of large interval changes on the harmonisation process. During my main analysis I transposed all songs to C; this is done mainly due to data sparsity issues, as I simply didn't have the time to gather enough data for a model that properly accomodates songs in various keys. Nevertheless, the `NoTranspose` model gives a glimpse at the ability of the current model to harmonise songs in various keys.

3.1 General Statistics

	Chords Recognized	Melodies Recognized	Unique Melodies
Instr Normal	12	1128	241
Vocals Normal	13	2406	277
Instr NoScale	12	1128	172
Vocals NoScale	13	2406	207
Instr NoTranspose	14	1128	275
Vocals NoTranspose	13	2406	308

Table 1: Data Frequency Counts

There is a lot of repetition within both melodies, and there is more vocal data than instrumental data in general. The following are statistics on the step intervals between notes in the melodies, where a single step is measured as a musical half step. The notes `C1, D2`, for example, have an interval of 13.

	Average	Median	Variance	Max	2nd
Vocals	1.426	0	10.344	84	12
Instr	2.040	0	13.499	48	48
Vocals NoScale	0.682	0	1.581	6	6
Instr NoScale	0.897	0	1.800	6	6

Table 2: Statistics on Step Intervals. As a note, the `NoTranspose` intervals will be the same as the originals.

Step intervals for instrumental sections are noticeably higher than the vocal intervals, as expected. `NoScale` intervals are lower in general due to having a maximum value of 6. In all cases, the median value of 0 means that the note generally doesn't change in any given interval.

3.2 HMM Matrices

INSTRUMENTALS

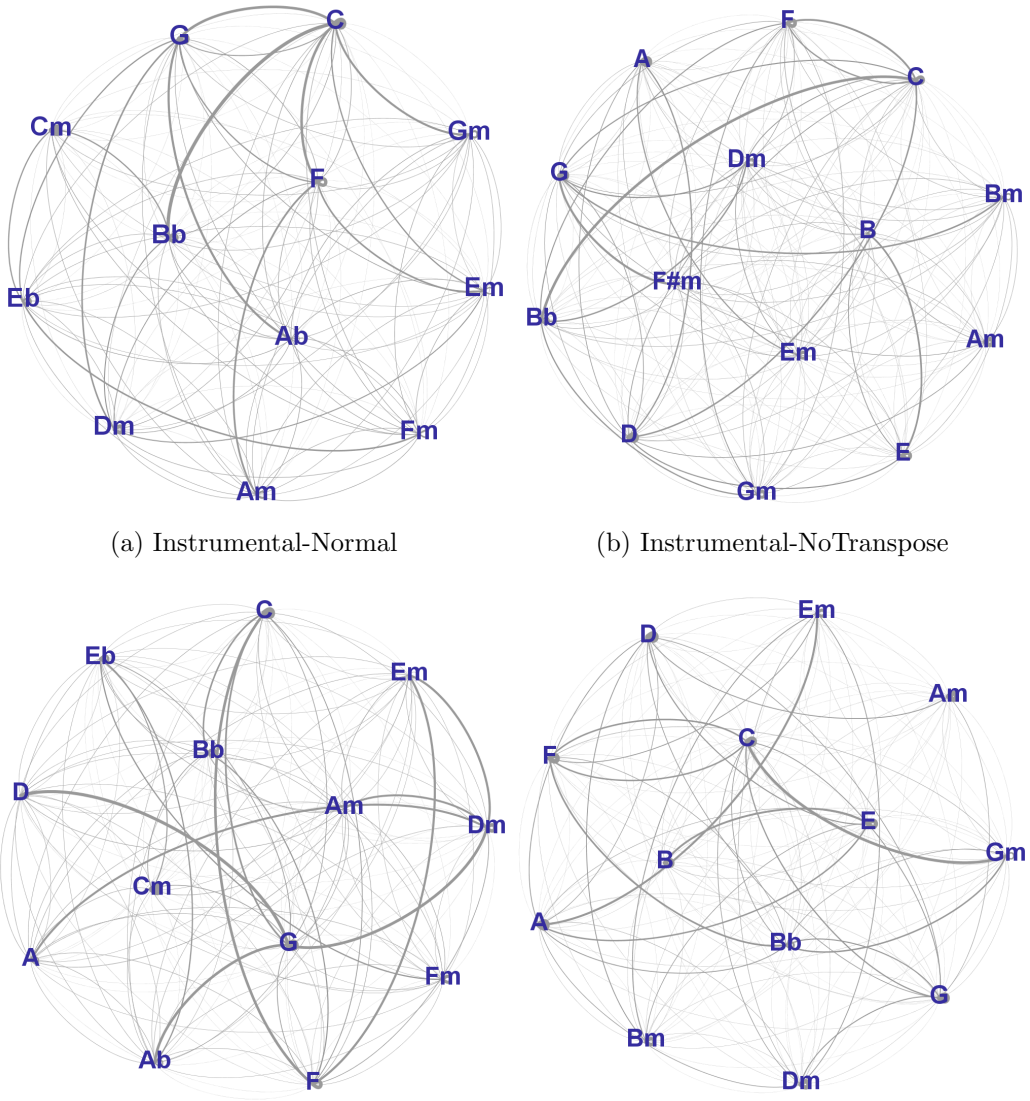
	Bb	C	G	F	Dm	Em	Cm	Ab	Fm	Eb	Gm	Am
Bb	0.44	0.41	0.03	0.01	0.01	0.01	0.01	0.04	0.01	0.03	0.01	0.01
C	0.07	0.69	0.11	0.07	0.03	0.02	0.00	0.00	0.00	0.00	0.01	0.01
G	0.00	0.32	0.55	0.07	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
F	0.01	0.31	0.15	0.41	0.06	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Dm	0.13	0.03	0.19	0.10	0.44	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Em	0.04	0.16	0.08	0.20	0.12	0.16	0.04	0.04	0.04	0.04	0.04	0.04
Cm	0.15	0.02	0.02	0.02	0.02	0.02	0.66	0.02	0.02	0.02	0.02	0.02
Ab	0.07	0.07	0.27	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Fm	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.19	0.19	0.06	0.06
Eb	0.05	0.05	0.15	0.05	0.05	0.05	0.15	0.05	0.05	0.25	0.05	0.05
Gm	0.05	0.25	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.25	0.05
Am	0.06	0.06	0.06	0.22	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.22

VOCALS

	C	G	F	Bb	Dm	Em	A	D	Am	Eb	Cm	Fm	Ab
C	0.73	0.08	0.10	0.02	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
G	0.22	0.69	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
F	0.34	0.10	0.46	0.01	0.03	0.02	0.00	0.03	0.00	0.01	0.00	0.00	0.00
Bb	0.20	0.01	0.15	0.39	0.01	0.01	0.01	0.01	0.01	0.07	0.01	0.05	0.04
Dm	0.01	0.36	0.08	0.01	0.45	0.05	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Em	0.04	0.04	0.28	0.04	0.28	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
A	0.05	0.05	0.05	0.05	0.26	0.05	0.16	0.05	0.05	0.05	0.05	0.05	0.05
D	0.04	0.39	0.04	0.04	0.04	0.04	0.04	0.13	0.04	0.04	0.04	0.04	0.04
Am	0.06	0.06	0.06	0.06	0.25	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Eb	0.03	0.10	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.32	0.13	0.03	0.16
Cm	0.01	0.09	0.01	0.09	0.01	0.01	0.01	0.01	0.01	0.01	0.73	0.01	0.01
Fm	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.16	0.04	0.40	0.04
Ab	0.05	0.33	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.14

Figure 2: A matrices for instrumentals and vocals (normal).

In both models, it can be seen that states are generally likely to stay the same.



(a) Instrumental-Normal

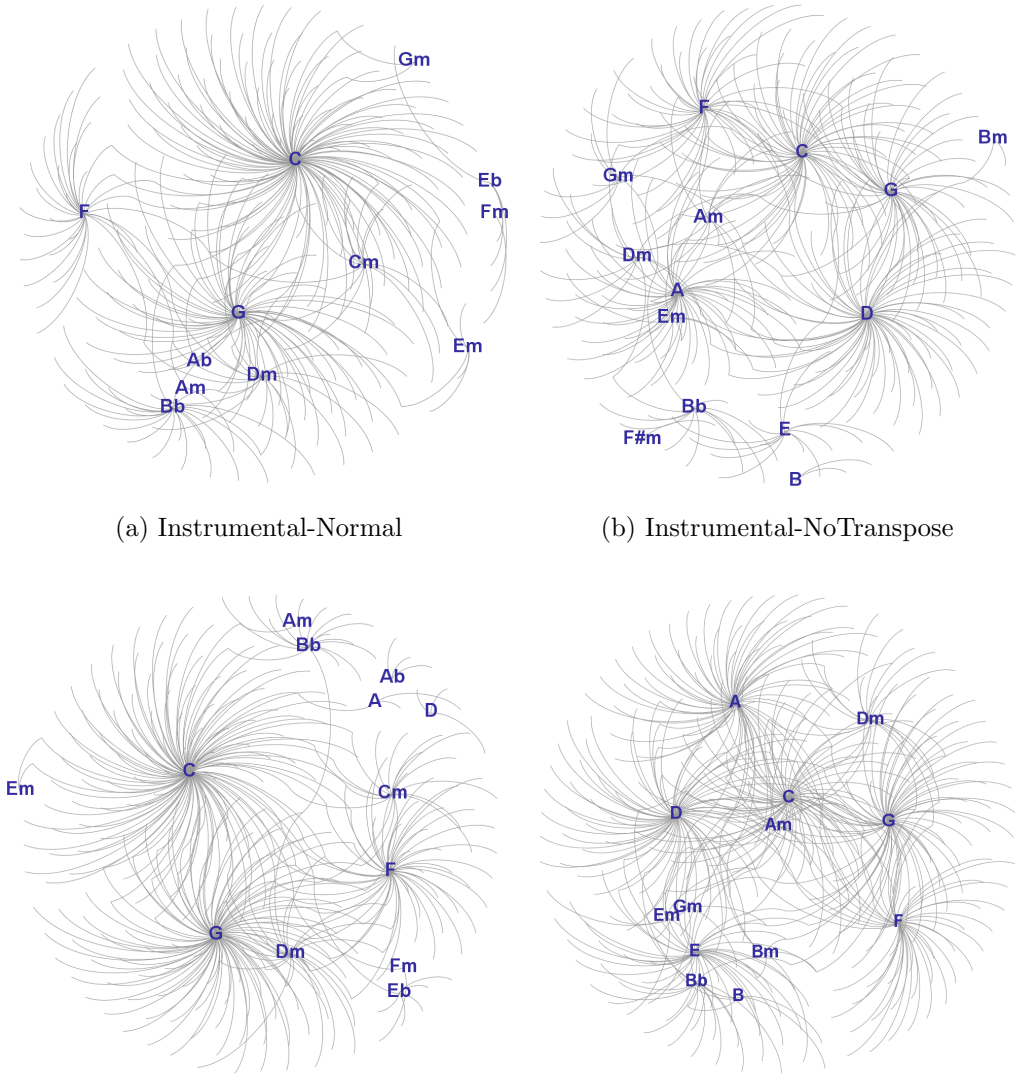
(b) Instrumental-NoTranspose

(c) Vocal-Normal

(d) Vocal-NoTranspose

Figure 3: Visualization of A matrices. The self loops are difficult to see. As a note, the `NoScale` option has no effect on the state observations.

We can further see from fig 3 that the I-IV-V chords for a key (C-G-F in C major) are likely to transition amongst each other. The distribution amongst less common chords within the key are more uniform. The instrumental `Normal` model has a strong C-Bb transition, while the Vocal `Normal` model doesn't.



(a) Instrumental-Normal

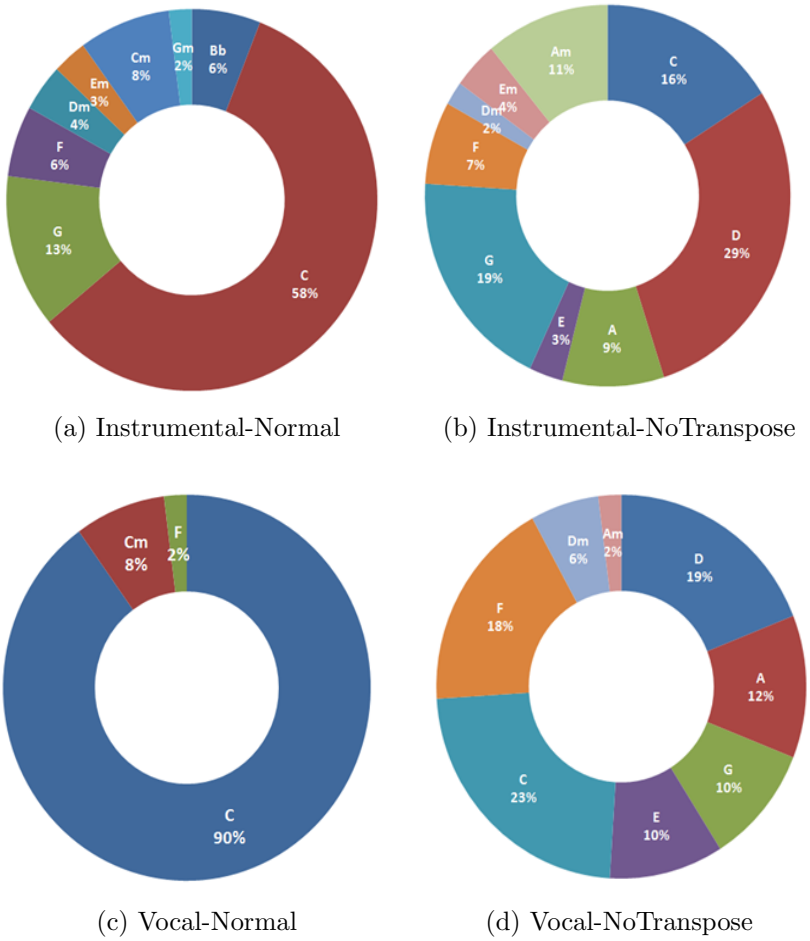
(b) Instrumental-NoTranspose

(c) Vocal-Normal

(d) Vocal-NoTranspose

Figure 4: Visualization of B matrices. Intuitively, the `NoScale` will have little effect on the visualization, as it will simply merge certain output melodies from the states.

There is a heavy concentration of melodic outputs for the chords `C,F,G` in the `Normal` models, while the `NoTranspose` models have a more balanced dispersal. It can also be noted that most melodies are produced from a single chord.



(a) Instrumental-Normal (b) Instrumental-NoTranspose

(c) Vocal-Normal (d) Vocal-NoTranspose

Figure 5: Visualization of π matrices. The `NoScale` option has no effect on the state observations.

There is a noticeably low variety in initial chords in the vocal `Normal` model, while the instrumental `Normal` model has a more varied distribution. This could result in a high rate of correct initial predictions in the vocal model, which the instrumental model would struggle with. This will likely not be as apparent in the performance of the `NoTranspose` models.

Out of the HMM matrices, the π vector displays the biggest intuitive difference in the vocal and instrumental models.

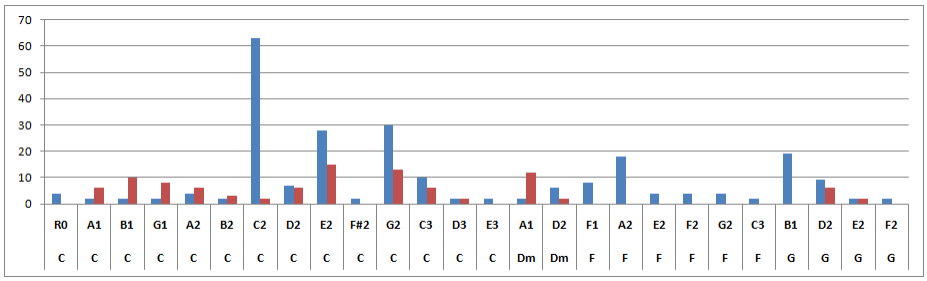
3.3 Performance

Due to the pervasiveness of the C chord in the models (except `NoTranspose`), I also measured accuracy on measures where the actual chord was not C, to see if there was a noticeable difference.

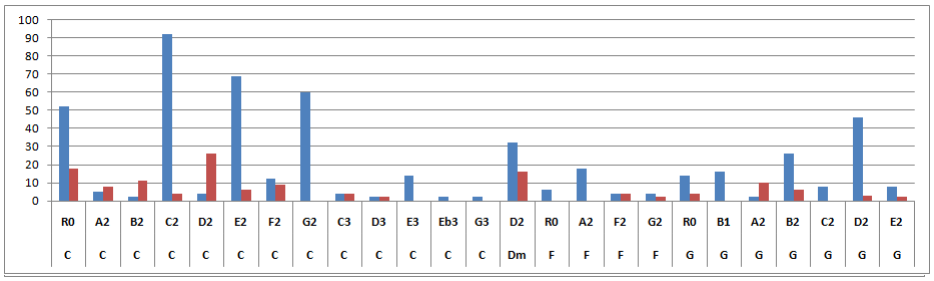
	Accuracy	Accuracy Without C
Instr Normal	0.612	0.345
Vocals Normal	0.778	0.590
Instr NoScale	0.673	0.448
Vocals NoScale	0.790	0.615
Instr NoTranspose	0.347	-
Vocals NoTranspose	0.309	-

Table 3: 10-fold cross validations using various metrics.

The instrumental harmonisations were generally lower than the vocal harmonisations. Though the accuracy dropped significantly when removing C, the vocals still recorded accuracies around 10% higher. The `NoTranspose` model, however, resulted in a significant performance drop, with the instrumental harmonisations giving a higher accuracy.



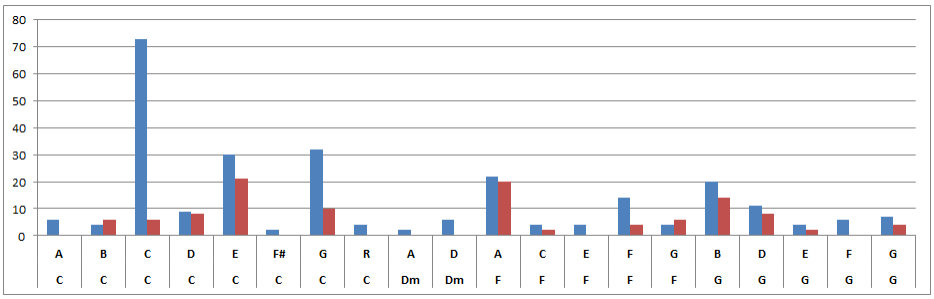
(a) Instrumental-Normal



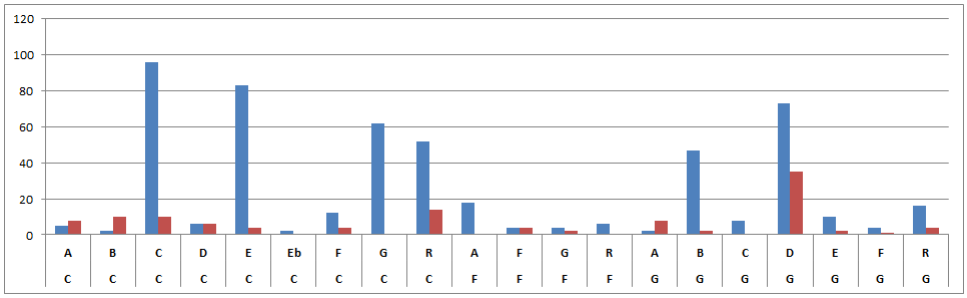
(b) Vocal-Normal

Figure 6: Hit/Miss counts of a predicted chord when the given note was played during the corresponding half-measure.

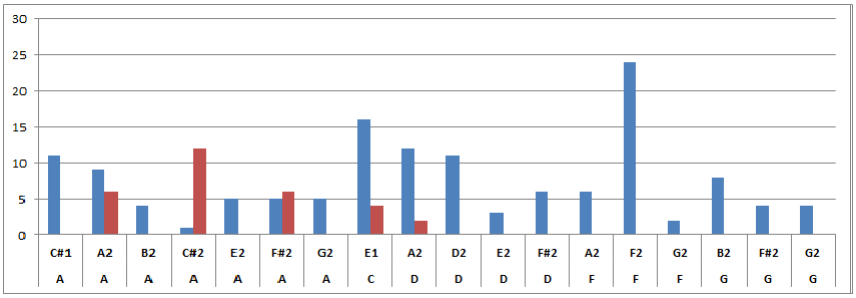
Both models greatly benefitted from high accuracy counts on the arpeggio-based notes of the relevant chord (C-E-G for C major).



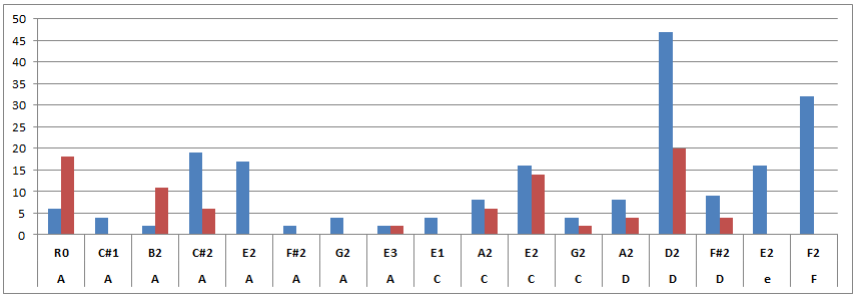
(a) Instrumental-NoScale



(b) Vocal-NoScale



(c) Instrumental-NoTranspose



(d) Vocal-NoTranspose

Figure 7: Hit/Miss counts of a predicted chord when the given note was played during the corresponding half-measure.

The `NoScale` models displayed similar arpeggio-based dependencies to the `Normal` models. The `NoTranspose` instrumental model was surprisingly accurate in the D,F,G chords,

while less accurate on the A chord, which constituted the majority.

4 Conclusion

The results, while promising, are inconclusive. While vocal harmonisations generally had higher performance, this was due more to them being more strongly based in C (fig 5), as opposed to being a direct effect of the sharper interval changes. While the -C experiments also returned higher vocal performances, the resultant data counts were too low, and could possibly vary given a slightly shifted test sample.

It's difficult to make an assessment on the `NoTranspose` models due to the low data counts. Though they performed significantly worse, they were also effective in predicting certain chords, as seen in fig 7. The biggest problem introduced with the `NoTranspose` models is the introduction of ambiguity to the previously more straightforward chord transitions, as seen in fig 3, since the model could potentially "float" between various keys based on the B matrices, if the A matrices aren't structured properly.

This model also does not incorporate the various interpretations of a given standard chord, such as C7, Cdim5, and so on. A true harmoniser should effectively be able to recognize such variations as well; this again would introduce ambiguity and require much more supporting data.

Ultimately, the experiment would need to be performed using more data, possibly from other genres of music as well to possibly take into account the C-centric behavior of the music in the dataset.

References

[1] "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77.2 (1989): 257-86. *http://ieeexplore.ieee.org/*. IEEE Xplore. Web. 31 Oct. 2013.

[2] "The Viterbi Algorithm." *Proceedings of the IEEE* 61.3 (1973): 268-78. *http://ieeexplore.ieee.org/*. IEEE Xplore. Web. 31 Oct. 2013.

[3] "Miyuru Gee." *Miyuru Gee*. N.p., n.d. Web. 31 Oct. 2013.

[4] Borrel-Jensen, Nikolas, and Andreas H. Danielsen. "Harmonisation in Modern Rhythmic Music Using Hidden Markov Models." *http://www.hjortgaard.net/*. University of Copenhagen, n.d. Web. 31 Oct. 2013.