# AUTOMATIC TRANSCIPTION AND ARRANGEMENT ANALYSIS FOR A CAPPELLA MUSIC

**Joseph James DiPassio III**
Ph.D. Student in Electrical Engineering
Computer Audition (ECE-477)
University of Rochester
JDiPass2@rochester.edu

## ABSTRACT

This work explores the implementation of an automated transcription suite specifically designed for A Cappella singing groups. In this work, a system for recording vocalists during an improvised musical work will be proposed, as well as a system by which their improvisation is rated against the conventions of classical music theory. Via this rating system, improvements for the improvisation will be presented to the users. The main research elements explored in this paper are the use of cascaded statistical filtering over the passage's beats combined with a priori knowledge about the tendencies of the singing group to improve the accuracy of the pitch detection system, as well as the use of bi-gram methods borrowed from language processing to analyze the passage's chord progression. This paper will show that the use of this a priori knowledge will allow greater pitch detection accuracy versus the baseline algorithm for a majority of the voice parts present in the test passages.

## 1. INTRODUCTION

Collegiate A Cappella groups are widely known for their ability to make music anywhere and anytime, due to their ability to perform without instruments (and therefore without extensive set-up time) and their passion for making unique and interesting music. Therefore, it is no surprise that a common pastime for these groups is to improvise a short and often groovy musical passage from the ground up. This is often done at the end of a group's practice or before performances as a warm-up. Often, these improvised passages sound aesthetically pleasing to the group, however given the circumstances under which they are performed they are not always easy to remember, and it is not always convenient to immediately transcribe them to sheet music for future reproduction.

### 1.1 Motivation

The author of this paper was previously in an A Cappella group at the Rochester Institute of Technology, called RIT's Brick City Singers. Therefore, it can be proposed first-hand that an automated transcription system for improvised A Cappella passages would be of immense use to these groups. These improvised grooves would often be useful for composing future arrangements for the group, either via inspiration or by directly plugging a version of the passage into a new piece. The system proposed in this paper will attempt to be a standard and all-encompassing solution to this problem, and empower the directors of these groups an easy way to save, reproduce, and draw from these improvised passages.

## 2. THEORY

It is important to establish a framework for how single-source vocal transcription can be achieved for this application.

### 2.1 Pitch Detection

A comprehensive look at spectral and temporal approaches to this problem is given in Cheveigne's chapter in the Computation Auditory Scene Analysis book [2]. It is this spectral approach involving pattern matching that will be most closely implemented in this work, however other approaches must be considered to account for the challenges introduced when utilizing singing voice. Work coming from Stanford's CCRMA lab delves into a fundamental frequency estimator for singing voice [3]. The theory presented in this paper also utilizes a maximum likelihood approach to determining the fundamental frequency of vocals.

The general theory behind pattern matching is that an optimal choice of fundamental frequency (F0) can be selected for a given frame by utilizing the known frequency of every peak in the frame, and diving the frequency of each peak by successive integer multiples to determine which harmonics could be associated with that peak. Every division results in the increment of a bin within a histogram, and upon the completion of this iterative process, F0 is estimated as the largest bin [2]. The following equation was generated for this project to show the specific implementation used:

$$Bin_f = \sum_j \sum_i I[(Peak_j/i) = f] \tag{1}$$

### 2.2 Use of A Priori Information

A unique proposed method for improving the accuracy of the single-channel transcription algorithm will be the use of personalized distribution functions for each singers range. An important element of any A Cappella group is their repertoire. As digitized composition software has become more and more accessible, a good number of collegiate A Cappella groups have made the switch to completely digitizing their repertoire. This has the added advantage of allowing MIDI tracks to be stored for each piece.

Utilized in the MATLAB workspace was an academic library for enabling MIDI functionality within MATLAB, called the MATLAB MIDI Toolbox [1]. With code written off the framework of this toolbox, the entire repertoire of a group can be imported into the workspace and analyzed, with the result being a probability distribution of all notes within each voice parts range. For each MIDI file in the group's repertoire, a histogram based on MIDI value can be created, storing the frequency of each note observed for each channel (or voice part) in the piece. These histograms are then added together, and normalized to create a probability distribution. By the principle of the central limit theorem, and assuming each piece in the repertoire is relatively independent and independently distributed, the resulting probability function should resemble a Gaussian. This exploitation of a groups repertoire will allow the proposed system to be even further informed about the tendencies of the group's singers that comprise the system's users.

### 2.2 Improvement Guidelines

The problems with single-voice transcription of vocalists are well known and documented. Singing elements such as vibrato, imperfections in the singer's technique, approaching a note from a lower frequency (leading to a "swooping" effect), and noise in general can severely inhibit the accuracy of transcription systems.

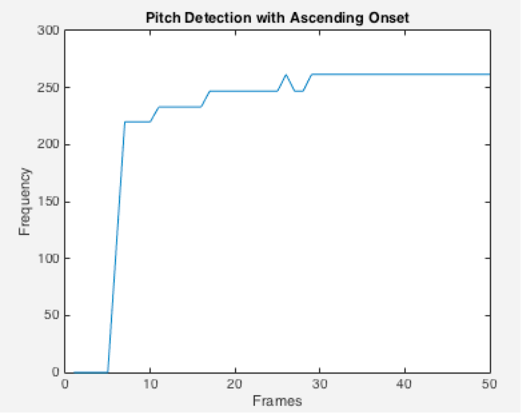

Figure 1: Swooping Onset Effect for a Single Note

In order to alleviate these issues, two systems will be proposed. Firstly, after performing pattern matching weighted with the a priori information, median filtering will be performed across the sequence of estimated fundamental frequencies. This type of filtering should alleviate some noise concerns, especially where pre-filtering is not implemented. Secondly, quantizing each beat of the passage via mode filtering along this median filtered contour will not only allow a standard subdivision of the audio to be defined, but it will be tested for it's ability to filter out swooping effects (such as the result shown in Figure 1), and partially eliminate the artifacts caused by singer vibrato.

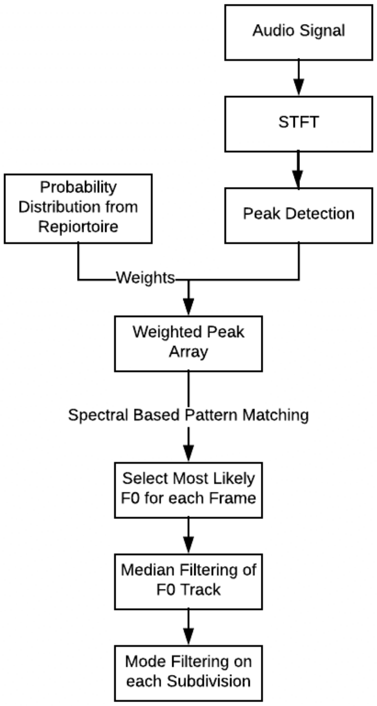
Audio Signal
STFT
Probability Distribution from Repiortoire
Peak Detection
Weights
Weighted Peak Array
Spectral Based Pattern Matching
Select Most Likely F0 for each Frame
Median Filtering of F0 Track
Mode Filtering on each Subdivision

Figure 2: System Chart for Transcription Algorithm

### 2.3 Chord Recognition

An addition feature of the proposed system will be to recognize the chords within the improvised passage. A paper published by Oudre et al. discusses a method for doing so based on chord templates [4]. A naïve and simplified approach will be implemented for this paper. In essence, a chord table was implemented with the makeup of each chord in a key. Using the transcribed audio of each beat in the musical passage, the observed notes will be compared to the templates in the chord table, and the template chord with the largest percentage match will be selected.

Once the chords have been assigned to the subdivisions of the passage, a bi-gram model will be implemented to determine if the selected chord makes sense in the context of the piece. Bi-gram models are longstanding items in the field of speech processing and natural language understanding to analyze when how words relate to each other within a sentence [5]. By thinking of the musical passage as a pseudo-sentence that is governed by the rules of music theory, the bi-gram model can instruct the user when a chord is performed erroneously in context with the chords around it, using classical progression rules noted in the textbook Tonal Harmony [6]. The following figures demonstrates an example of this process:

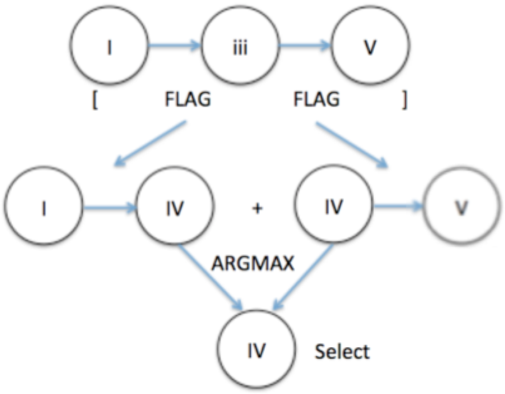

Figure 3: Sample Chord Recommendation Based on Context

## 3. IMPLIMENTATION

The solution explored in this paper is proposed with the idea that the end-user will be able to record, transcribe, and analyze their group's improvised piece all within this one software suite. The workflow of the solution is described below in the following sections.

### 3.1 Recording Workflow

When an A Cappella group uses this proposed system, recordings will be made with a user-defined tempo, and played in the MATLAB environment via the sound buffer. The recordings will consist of 16 beats, assuming a 4/4 time signature. The user will then enter the voice parts of everyone to be recorded, and then the system will prompt each singer to begin singing. As each singer finishes their part, their audio gets added to playing track, similar to a loop pedal, and also saved to a separate array for transcription. Once everyone has recorded, the post-processing algorithms begin to execute. The system can also use imported audio by calling the designed functions directly from the command window.

### 3.2 Post Processing for Transcription

At this point in the system's operation, it has saved an array containing the audio from every singer that is to be transcribed. It is also assumed that the user has already provided the system with all appropriate MIDI files making up the group's repertoire, and the probability distributions for each voice part over the entire repertoire have been created.

Each singer's audio is broken into frames using STFT. The frames are 512 samples long with 50% overlap, and are windowed with a hamming window. The peaks of each frame are detected, and weighted via the voice parts probability distribution. Via pattern matching, the fundamental pitch in each frame is estimated for each frame of audio, and the successive median and mode filters are applied. This process is described in Figure 2.

### 3.3 Post Processing for Chord Recognition

From here, the template-based approach to chord recognition was implemented, whereby every note being sung is fed into an estimator, and the chord with the highest template match for those notes is selected as the chord. Every subdivision of the passage is analyzed for chord contents, and smoothed by looking for multiple of the same successive chords.

### 3.4 Suggesting Improvements

Finally, treating the chords as words in a sentence, a bi-gram model is employed. In essence, the system is "trained" to know which chords can follow or precede other chords based on the conventions of music theory [6]. If a chord violates this rule, the bi-gram model will be employed to see what chords make sense in-line to replace the erroneous chords.

## 4. RESULTS

The following section will show the results of extracting a group's voice part distributions, applying the distributions and the filtering systems to provide transcription, and a primitive look at the chord recommendation system. In order to actually test the robustness of the output, singers will perform several pre-written musical passages within the workflow of the solution. The results of the transcription and chord recognition can thus be compared to ground truth realizations, and therefore the accuracy of the system can be analyzed.

### 4.1 Test Group's A Priori Information

By using the repertoire of the author's old group, RIT's Brick City Singers, with the created MIDI distribution function, the tendencies of each voice part based on the given arrangements was determined. The following shows the distribution of the voice part's expected range:

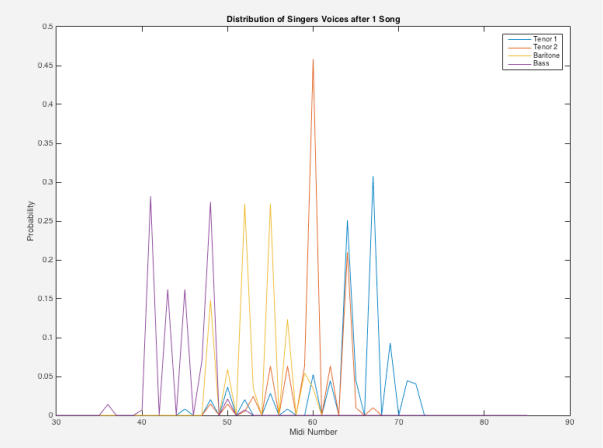

Figure 4: Distribution Over One Song

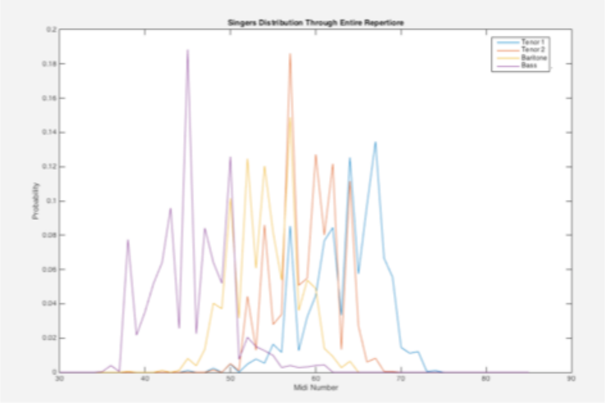

Figure 5: Distribution Over Entire Repertoire

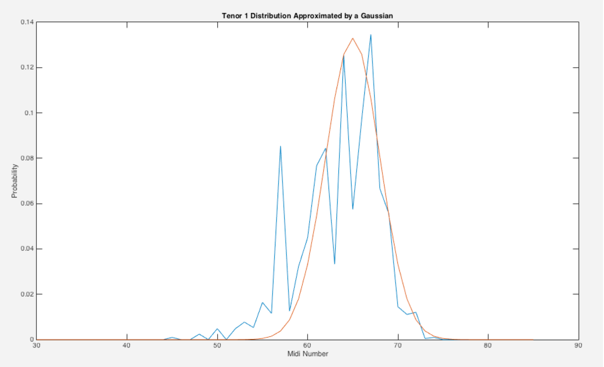

Figure 6: Gaussian Nature of Tenor 1 Distribution

In an encouraging result, the distribution functions of all voice parts seemed to successively become closer and closer to a Gaussian shape. There are some limitations to achieving a better Gaussian result, such as collegiate arrangers sticking to certain "comfortable keys" for the bulk of their arrangements, and the typically smaller size of a group's current repertoire. However, the approach to a Gaussian shape is still very apparent in these results.

### 4.2 Transcription Results

A test passage was used to demonstrate the effect of the cascading filters on the system:

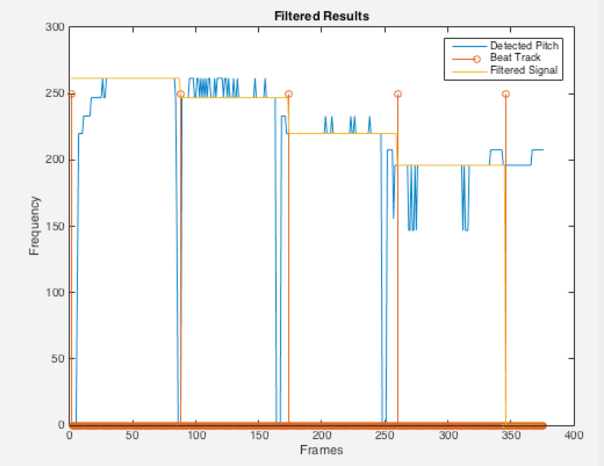

Figure 7: Simple Transcribed Passage

As can be seen from the above figure, the cascading filters show promising results in terms of eliminating the impacts of swoops, vibrato, and noise. The filtered signal actually proved to be a perfect transcription in this case, as the short passage sung was [C4 B3 A3 G3], which is exactly what is seen in the frequency domain. The transcription system can now be tested with larger data sets with appropriate length signals.

Several passages were pre-arranged such that the accuracy of the system could be compared against ground-truth. The following shows one result of this testing with a baritone singer. The transcribed results, which get saved as MIDI files by the system, were imported into Finale software for visualization:

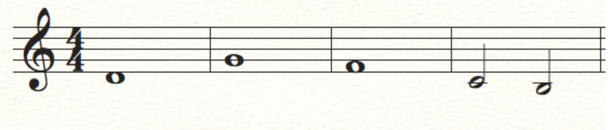

Figure 8: Ground Truth Passage Performed

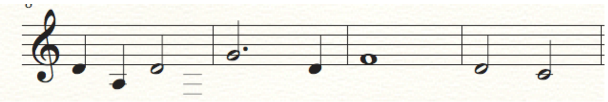

Figure 9: Transcription Result Without A Priori Info

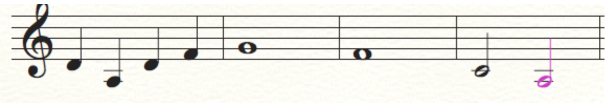

Figure 10: Transcription Result With A Priori Info

As can be seen from the results above, without the use of the group's distribution function, a total of 10 beats were transcribed correctly. With the group's distribution function used for weighting, a total of 12 beats were transcribed correctly. This is an encouraging result, although with an extremely limited sample size. It was discovered that in general, using the distribution weights only caused a small impact on the transcription correctness, as shown in the following table. These results are derived from each singer performing four passages, each containing 16 beats. These results will be discussed in section 4.4.

| | **% of Beats Correctly Transcribed** | |
|---|---|---|
| **Part** | **Without Weights** | **With Weights** |
| Tenor 1 | 72.91% | 75.00% |
| Tenor 2 | 81.25% | 85.41% |
| Baritone | 81.25% | 87.50% |
| Bass | 85.41% | 79.20% |

Table 1: Summary of Transcription Results

### 4.3 Chord Recognition Results

At the point of this paper's submission, the actual chord recognizer system still exists in two parts: the template matching portion and the bi-gram model portion. It will be implemented in future work to tie these two together, although their independent results will be briefly discussed.

The template-matching portion of the implementation is successfully able to determine which chord is present based on the notes in the MIDI. The following table shows some results:

| Input Notes | Possible Chords | % Match |
|---|---|---|
| **C, E, G** | C Major | 100% |
| | E Minor | 66% |
| | A Minor | 66% |
| | G Major | 33% |
| **D,F#,A,B** | B Minor 7 | 100% |
| | D Major | 63% |

Table 2: Summary of Template Matching Results

As can be seen from Table 2, the system is able to identify chords correctly in any inversion state, and applies a penalty to extraneous notes not in the chord template. To date, only triads and seventh chords are implemented, but future work will be to expand this functionality to more complex chords.

For the bi-gram model, Figure 3 shows a sample result that was made into a figure. For the progression [I iii V], the maximum accepted chord by the algorithm to replace [iii] is the [IV] chord because of its relation to both the [I] and the [V] chords. This is a selected result, and admittedly a small sample size, however once the template system and the recommendation system are merged, results can be reported in a much more efficient way. The current weighting system reports only a 0%, 50%, or 100% match, whereby a suggested chord fits the motion to neither, one, or both of the chords around it. Priority is then given to the dominant and subdominant chords that aren't already a part of the progression, and then randomly selected after that. This weighting system will have to be expanded to include priority-ranking scores beyond the dominant and subdominant.

### 4.4 Discussion of Results

As reported in section 4.1, the distribution functions of the voice parts started to approach Gaussian distribution functions as more and more songs were added. However, it is interesting to note that the distribution function for the Bass singer appears to be the most irregular, with spikes on each end of the total distribution. This might have lead to the accuracy of the transcription to worsen when using this distribution. In all, it appears that the use of this distribution is helpful to the transcription method, and Figure 7 shows a clear benefit for quantizing and filtering the signal in the proposed method, as it leads to a clean and useable output for the system to analyze.

## 5. FUTURE WORK

In all, the results reported in this paper demonstrate a good first pass at implementing this system. However, several areas of this implementation will need to be improved before the end goal of this system can be achieved.

Firstly, improving the implementation of the vocal transcription method will be important. Even though the use of the distributions is shown to lead to improvements in the single-channel transcription results, a baseline of 70% - 80% is still unacceptable, with modern consumer products and research results showing near perfect transcription results with similar constraints. Further literature search and advanced methods using neural networks will need to be considered. Additionally, testing on a much wider set of data will be important to prove robustness.

As mentioned earlier in the paper, the template matching and the recommendation algorithm must be merged in order to achieve the end goal of the recommendation system. While they both show they are working independently at this point in the project, they must function together on the transcribed MIDI data to fully achieve the desired results.

Additionally, implementation of some sort of vocal percussion tracking features will be vital to a complete solution. A dictionary should be trained for each element of the group's vocal percussionist's repertoire of sounds. Using this dictionary, the vocal percussionists track will be separated, and the sound's onsets will be used to "line up" and transcribe the percussionists track. This future work will make use of non-negative matrix factorization techniques derived from the work by Smaragdis and Brown [7].

A final point of future work would be to receive and implement feedback from the Contemporary A Cappella Society (CASA) and their community, both in terms of accuracy and usefulness, and in the actual ease of use of the program. It is expected that a proper GUI will be necessary for widespread distribution for this system, and CASA's feedback will be vital to the development of this system. The use case will also have to be analyzed, and perhaps further subdivisions of the measures will have to be considered, among other suggested improvements.

## 6. REFERENCES

[1] T. Eerola, and P. Toiviainen: *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä: Kopijyvä, Jyväskylä, Finland.

[2] D. Wang, and G. J. Brown: *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[3] H.L. Lu: "A Hybrid Fundamental Frequency Estimator for Singing Voice."

[4] L. Oudre, Y. Grenier, and C. Févotte: "Template-based Chord Recognition: Influence of the Chord Types," *ISMIR*, pp. 153-158, 2009.

[5] B. Suhm, and A. Waibel: "Towards better language models for spontaneous speech," *Third International Conference on Spoken Language Processing*.

[6] S. Kostka, and D. Payne: *Tonal Harmony*, 5th ed., McGraw-Hill, 2004.

[7] P. Smaragdis, and J. C. Brown, J. C: "Non-negative matrix factorization for polyphonic music transcription," *IEEE workshop on applications of signal processing to audio and acoustics*, Vol. 3, No. 3, pp. 177-180, 2003.