MELODY LINE DETECTION AND SOURCE SEPARATION IN CLASSICAL SAXOPHONE RECORDINGS

Estefanía Cano C.[(*)(+)], *Corey Cheng* [(+)(#)]

[(*)]TU Ilmenau/ Fraunhofer Institute IDMT
Ilmenau, Germany

e.cano3@umiami.edu

[(+)]Music Engineering Technology Program
Frost School of Music
University of Miami
Coral Gables, FL USA

coreyc@miami.edu

[(#)]Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL USA

ABSTRACT

We propose a system which separates saxophone melodies from composite recordings of saxophone, piano, and/or orchestra. The system is intended to produce an accompaniment sans saxophone suitable for rehearsal and practice purposes. A Melody Line Detection (MLD) algorithm is proposed as the starting point for a source separation implementation which incorporates known information about typical saxophone melody lines, acoustic characteristics and range of the saxophone in order to prevent and correct detection errors. By extracting reliable information about the soloist melody line, the system separates piano or orchestra accompaniments from the solo part. The system was tested with commercial recordings and a performance of 79.7% of accurate detections was achieved. The accompaniment tracks obtained after source separation successfully remove most of the saxophone sound while preserving the original nature of the accompaniment track.

1. INTRODUCTION

1.1. Melody Line Detection

Melody Line Detection is a problem that has received considerable amount of attention due to the large number of applications that could benefit from a solid and reliable algorithm for this purpose. For example, systems for audio classification often use melody lines to classify or identify tunes from a database, and query by humming systems also use melody lines amongst other attributes to identify a song. Music transcription systems can also benefit, as detecting melody lines allows for music transcription of single lines in polyphonic signals. Audio coding and segmentation can also use melody lines and common musical structures to avoid redundancy in coding schemes.

One of the earliest works in Melody Line Detection is the system for predominant F0 estimation presented by Goto [1]. He proposes a probabilistic method for melody and bass line detection where no assumption is made regarding the number of consecutive sources. The missing fundamental phenomenon is accounted for and the use of tone models in MLD is introduced. The system uses a MAP estimation to obtain the model parameters and the final F0 trajectory is obtained by a salience detector and a multi agent architecture.

Paiva, Mendes and Cardoso [2] propose a melody line detection system which uses a model of the human auditory system as a frequency analysis front end and MIDI-level note tracking. Periodicities within frequency channels are obtained by means of the auto-correlation function and salience curves are used to segment tone trajectories. Candidate fundamentals are eliminated based on their salience, duration and octave relation.

Eggink and Brown [3] propose a system which detects melody lines played by a solo instrument in an accompanied sonata or concerto. The key features of this implementation are the use of knowledge sources and an instrument classification system. Our proposed system is similar to this work in these respects.

1.2. Source Separation

The process of isolating the signals associated with different sources when only the mixture of all the signals is available is called source separation. The complexity of source separation is well known as in most cases no information about the mixing conditions of the signals is available. Furthermore, partial collisions are frequent and inharmonicities present in most musical signals – even in pitched musical instruments- make predictions even harder. The underlying assumption in this implementation is that given the complexity of a blind source separation task, a well-detected melody line along with saxophone instrument specific information will permit a more successful source separation implementation.

Virtanen and Klapuri [4] propose a system for blind source separation in monophonic recordings that represents signals as sinusoids with time varying frequencies, amplitudes and phases which are assumed to be constant in single frame analysis. Perfect harmonicity is not assumed and the system builds upon the fact that frequency ratios remain constant even when the fundamental frequency varies. A linear model is used to force the spectral

envelope of each sound to be smooth. The system incorporates an iterative search for the parameters that best fit the observed spectra.

Woodruff, Pardo and Dannenberg [5] propose a method for informed source separation that uses knowledge of the written score and spatial information from an anechoic, stereo mixture to isolate individual sound sources. The key feature of this model is that it introduces a score alignment algorithm to further enhance separation accuracy. The idea with the alignment algorithm is to recover information about expressive timing and tempo from the audio track that is not available in the score's MIDI file.

Bay and Beauchamp [6] propose a source separation method that represents each instrument as a time varying harmonic series and uses information from the instruments' spectra to enhance detection and improve separation results. An instrument spectra library is created using instrument samples from the University of Iowa Database [7] and a nearest neighbor approach is implemented to find the spectrum in the library that best matches the F0 combination obtained with a Gaussian mixture.

2. PROPOSED MODEL

We propose a system which consists of two main stages: a melody line detection stage where the soloist line is detected and a source separation stage where the soloist line is removed from the track. For the purposes of this paper, melody line detection refers to the process of determining the sequence of notes played by a soloist alto saxophone in a classical recording when the number of simultaneous instruments or sources in the accompaniment is not known and no information about the mixing conditions of the track is available. Fig. 1 shows a block diagram of the proposed model.

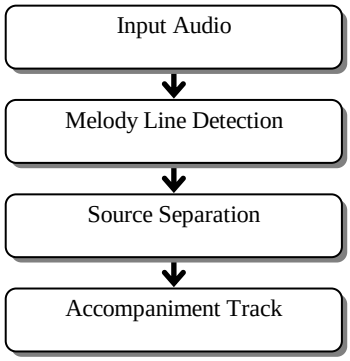

Figure 1: *Block diagram of the proposed model.*

2.1. Melody Line Detection

A detailed block diagram of the melody line detection algorithm is shown in Fig 2. Audio clips from commercial classical saxophone recordings with either orchestra or piano accompaniments are used. A monophonic track is obtained from the stereo recordings and the sampling frequency of 44.1 kHz is kept. In the frequency analysis stage, audio is framed using a 3072 samples long Hanning window with a 50% overlap between consecutive frames. The Discrete Fourier Transform (DFT) is obtained for every audio frame and a spectral compression approach as proposed in [8] is used. The spectral compression stage raises the magnitude spectrum to the power of 0.67 as shown in Eq. 1. Previous results show that the peak picking procedure is facilitated and octave errors decrease with the inclusion of this particular value at this stage.

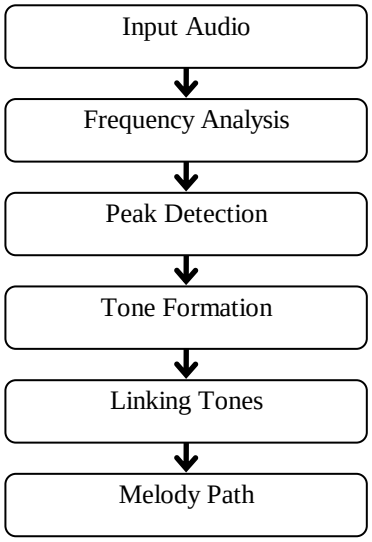

Figure 2: *Block diagram of the melody line detection algorithm.*

$$f(n) = \left(\left|DFT(x(n))\right|\right)^{\alpha} \text{ with } \alpha = 0.67 \quad (1)$$

2.1.1 Peak Detection

The magnitude spectrum is analyzed in the Peak Detection stage and local maxima are found. The number of spectral peaks or tone candidates is decimated in order to reduce complexity and remove spurious peaks. Three approaches are taken: (1) a frequency dependant threshold is used and all peaks whose amplitude is below the threshold are eliminated. (2) Information from the saxophone's register is incorporated so any spectral peak lying outside the range is eliminated. Both the regular register and the altissimo register of the saxophone are incorporated. (3) A Perceptual Pruning stage is also incorporated, which keeps only the spectral peaks that are perceived the most amongst all candidates. Specifically, using the critical band boundaries presented in [9], spectral peaks within a distance of 0.5 Bark are replaced by the strongest peak. The spectrum is normalized and scaled to the equal loudness contours as defined in [10], and only the 5 peaks with the highest loudness levels are kept for each frequency frame.

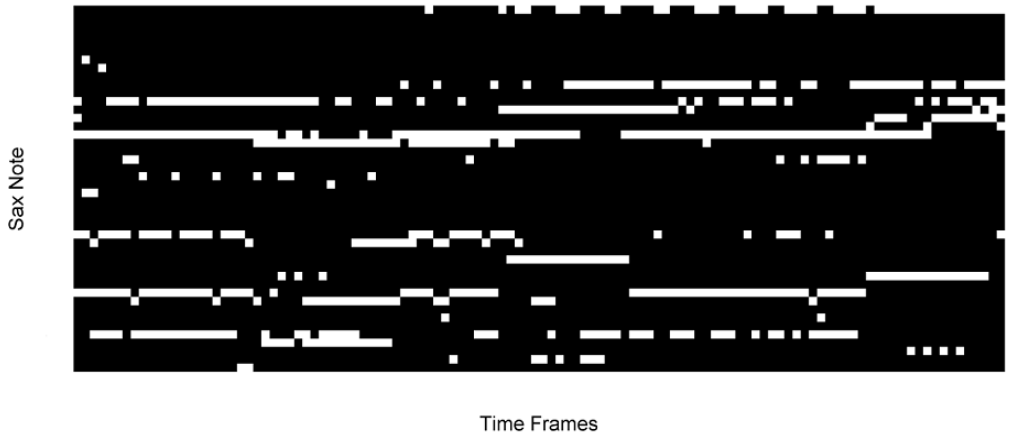

Figure 3: *Binary image representation of the tone candidates in an audio clip*.

2.1.2. Tone Formation

The purpose of the tone formation stage is to build longer tones from spectral peaks in different frames. In this stage the tone candidates of the entire clip are represented as a two dimensional array. The spectral peaks are represented as MIDI notes in a binary image where the columns are time-frequency frames and the rows represent saxophone notes. Fig. 3 shows a representation of an audio clip as a binary image. It can be seen that for every time frame – column – there are a maximum of 5 tone candidates. An Image Processing stage uses morphological operations on the binary image to remove isolated pixels that represent tones that are only one tone long and that are here assumed to be detection errors. The binary image is converted into a grayscale image whose intensity values represent the peak's amplitude normalized to a [0, 1] range. An Error Correction stage is then introduced where tones that seem to be continuous in time but have one pixel gaps are filled with the mean amplitude from adjacent pixels. Fig. 4 shows the grayscale image obtained after the Image Processing and Error Correction stages. It can be seen that all the isolated pixels have been removed and the one pixel gaps have been filled. The intensity values of the image represent spectral amplitudes.

At this point, information about the usage of the different notes within the saxophone's repertoire is used to weigh the tone candidates. We use note likelihoods to bias the tone candidates towards those notes which are more likely to be found in classical saxophone music. The use of note likelihoods in MLD applications was introduced in [3] and here we take this approach one step further by also using saxophone instrument specific information. By analyzing 4 pieces from the classical saxophone's repertoire, Note Likelihoods were obtained. The pieces analyzed were carefully selected as to have the most general sample space possible, both in terms of register and note usage and in terms of the different periods of the instrument's history. As was to be expected, the middle register is more frequently used that the lower and altissimo register.

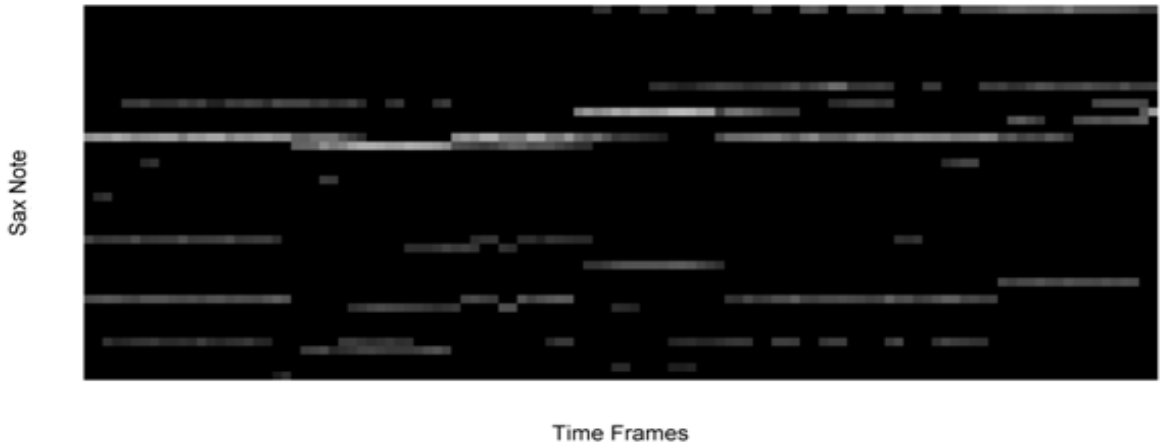

Figure 4: *Grayscale image obtained after the Image Processing, and Error Correction stages. The intensity values represent the amplitudes of the spectral peaks.*

A deep analysis of the saxophone's repertoire shows that due to the saxophone's transposing nature there is marked tendency to use keys with sharps in the key signature. To account for this fact, the choice not to smooth the results across frequencies was made. Details of this likelihood data is forthcoming and interested readers are welcome to contact the authors in the meantime. Fig. 4 shows a much cleaner representation of the tone candidates where the continuity of the tones becomes more apparent after these various data have been applied. After the Tone Formation stage, the two dimensional array of tones has been weighted with information from the spectral peaks' amplitude and the Note Likelihoods.

2.1.3. Linking Tones

The final stage in the MLD algorithm is the Linking Tones stage as shown in Fig. 2. The purpose of this stage is to identify and create longer tones and to select the final sequence of tones that corresponds to the melody path of the soloist saxophone. By using simple logical operations in the pixels of individual rows – saxophone notes - the exact frames where every tone begins and ends are detected. Now the likelihood of each formed tone, as a whole, is incorporated into the process by using their spectral amplitudes and Note Likelihoods. The tone likelihoods are normalized over tone length to avoid bias towards longer tones.

To create the final melody path, a Starters Detection algorithm was implemented with the purpose of determining the first tone within the melody path. Although we assume that the saxophone is playing at all times, we give the algorithm some flexibility within the assumption by searching for a starter note within the first three frames of the audio clip. A maximum of 3 starter tones are kept after an initial detection and starters are eliminated based on their Tone Likelihoods so that only the ones with the highest Tone Likelihoods are kept.

After the starter tones are detected the sequence of tones is defined by means of a local search. Tones completely overlapping with the starter tones were eliminated as tone candidates. As this algorithm is meant to deal with audio clips from commercial recordings, reverberant conditions had to be accounted for. We use a simple approach allowing successive melody tones to have up to 50% of overlap, and this proves to considerably increase the algorithm's performance. We also incorporate more saxophone-specific information at this stage. As in the Note Likelihood analysis, the 4 selected pieces from the saxophone's repertoire were analyzed for the frequency interval occurrence. The goal once again is to obtain numerical values for instrument specific Interval Likelihoods. The Interval Likelihoods are used to weigh transitions between tones. To avoid broken melody lines, tones shorter than 4 frames are under-weighted to reduce their likelihoods. While this approach prevents some detection errors, it also allows for the algorithm to deal with faster melody lines. The local search continues until no tones are left and are either included in the melody line or are eliminated as possible tone candidates. As a maximum of 3 starter tones are allowed, a maximum of 3 melody paths are obtained after the local search. At this point, the melody path that exhibits the maximum likelihood is selected as the final melody line.

2.2. Source Separation

After the melody path has been detected, the source separation implementation uses the information to remove the soloist from the audio track. For the source separation implementation an Overlap and Add system is used where the audio track is segmented using a Sine window 3072 samples long with a 50% overlap between consecutive frames. The magnitude spectrum and the melody path are used to modify the spectrum in such a way that the best representation of the accompaniment spectrum is obtained. The Inverse Fourier Transform (IFFT) produces a time domain signal from the modified spectrum.

At first, the algorithm determines the number of tones present in every time-frequency frame. This stage is necessary because some time-frequency frames have two simultaneous tones, one corresponding to the actual tone in the melody line and the other one corresponding to a reverberant tail from the previous tone. A TONE/TAIL flag is used to classify each tone either as a current saxophone tone or as a reverberant tail. The next stage in the algorithm consists of building a saxophone spectrum that best represents the saxophone in every tone. Three things needed to be considered before a representation of the saxophone spectrum could be obtained: (1) total number of harmonics considered, (2) location of each harmonic component and (3) amplitude of every partial. As far as the number of harmonics included, results suggest a minimum of seven harmonics are necessary to accurately represent the saxophone's spectrum. Building a spectrum with fewer harmonics results in audible artifacts and some tonal components left after the source separation process.

The location of the harmonic components was obtained assuming perfect harmonicity as a starting point and refining their location by searching the spectrum for local maxima. A well known characteristic of conical bore instruments is the flattening of upper resonances in relation to the fundamental component due to open end corrections in the tone hole lattice. For this reason refinement of the locations of the harmonic components was performed searching in lower frequency bins than the calculated for perfect harmonicity.

The amplitude of the harmonics was determined based on the assumption that 80% of the peak's spectral amplitude is produced by the soloist and the remaining 20% belongs to the accompaniment. This percentage showed to provide good separation results for most signals. To represent and capture the variations of the saxophone's spectral envelope, the amplitude of the first three partials was kept to be 80% of the spectral peak's amplitude and the remaining components were weighted with a frequency dependant decay as described in [11].

The saxophone spectrum representing each tone is built and by means of a spectral subtraction implementation, the accompaniment spectrum is obtained. For tones classified as reverberant tails, only 70% of the tone's amplitude is removed.

3. RESULTS

For the purpose of testing the algorithm, six different audio clips were taken from commercial recordings. Three of them were piano and saxophone recordings and the other three were orchestra and

saxophone. Each clip was processed manually and the notes being played by the saxophone were transcribed and used to compare the results delivered by the algorithm. The performance of the algorithm was tested on a frame by frame basis and results are shown in percentages of correct frames detected. To assess the contribution of each module to the overall performance of the algorithm, the different processing stages were temporarily removed from the system and the same clips were used to test performance. The modules were removed one at a time and performance was tested for following stages: Perceptual Pruning, Error Correction, Note likelihood, Interval Likelihood and Reverberant Conditions. Results are shown in Fig. 5 and exhibit that each one of the modules contributes to improving the performance of the system. In all cases the performance of the system decreases when removing modules. It is particularly important the contribution of the Error Correction module where performance decreases 18.79% when removing the stage. Similarly removal of the Reverberant Conditions module brings a 14.45% decrease of the system's performance.

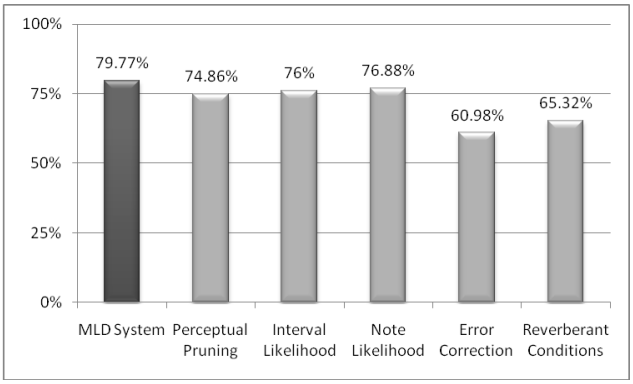

Figure 5: *Performance of the MLD algorithm*.

For the MLD tests described the six clips used were in average 2.0 seconds long. Some preliminary tests were performed using longer clips up to 8 seconds long. The performance of the algorithm considerably decreased and results suggest that a segmentation algorithm might be needed to obtain shorter clips from the original tracks. These results also suggest a strong dependence of the algorithm on the assumption that the saxophone is continuously playing. The Linking Tones module benefits from melodic continuity, but in longer clips where silent frames are frequent, performance considerably decreases.

A common cause of error in the Melody Detection algorithm was octave confusion. Due to the usual harmonic relation between the accompaniment and the soloist line, it was a common mistake for the algorithm to detect the melody line an octave below the actual F0 played by the soloist. The algorithm was tested using the magnitude spectrum raised to the second power instead of using the spectral compression stage. The number of frames where the algorithm detects the melody in the wrong octave was compared and results show that the octave error percentage decreased 30% with spectral compression.

It is important to mention that after thorough evaluation of the results obtained with each of the six clips tested, in most of the frames where the algorithm had detected the wrong note the correct note had been also selected as a tone candidate but was not included in the final melody path. This result shows that Frequency Analysis, Error Correction and both pruning stages provide solid and reliable information of the most relevant tones within the track.

The MLD algorithm decreases in performance in audio clips with faster tempos and faster rhythmic structures; however, the analysis of the tone candidates in each frame shows once again a tendency to include the right note as a tone candidate but not to incorporate it in the final melody path. Further work needs to be done to guarantee a melody path that while avoiding spurious detections and "jumpy" melody lines, achieves a better performance in faster tempos.

In most of the obtained accompaniment tracks hints of the saxophone could still be perceived. The biggest difficulty in the Source Separation stage is to accurately determine how much of the spectral content belongs to the soloist and how much is part of the accompaniment. The number of simultaneous sources is not known and harmonic collisions of the different instruments are hard to predict but too frequent in nature as to be neglected. The percentage used represented a good tradeoff between the amount of soloist removed and accompaniment information left, but this tradeoff could be improved as well.

Obtaining piano tracks showed to be a more complicated task than obtaining orchestra tracks. This was to be expected as with more instruments playing in the orchestra, imperfections caused by spectral subtraction are not so evident. The saxophone sound left in the accompaniment tracks is much more noticeable for piano tracks too. Results show that it might be convenient for future work to use different parameters in the source separation algorithm when piano tracks are used. Slightly wider bands surrounding the spectral peaks proved to deliver better results in piano tracks.

4. CONCLUSIONS

This project has shown that the extraction of solid information from the melody line allows for a successful blind source separation task. The use of frequency analysis techniques and the inclusion of specific information about the saxophone, musical lines and intervals is critical to the algorithm's performance. Furthermore, the treatment of reverberant conditions and the implementation of an error correction stage considerably increase the system's performance. In the Source Separation stage, accompaniment tracks that preserve the nature of the accompaniment sound with limited saxophone sound remaining and reduced amount of audible artifacts were obtained. Determining the amplitude of the saxophone spectral content within the track is still a challenge as collisions between partials are expected to happen but are not easily predicted. The approach taken delivers acceptable separation results and represents a good tradeoff between saxophone sound removed and accompaniment sound left.

5. REFERENCES

[1] M. Goto, "A predominant FO estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models," *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP)*, pp. V-3365-3368, 2001.

[2] R. P. Paiva,T. Mendes and A. Cardoso,"An auditory model based approach for melody detection in polyphonic musical recordings," *Computer Music Modeling and Retrieval (CMMR): II International Symposium*, pp. 21-40. Denmark, 2004.

[3] Eggink and G. J. Brown, "Extracting melody lines from complex audio". *International Conference on Music Information Retrieval* , pp. 84-91, 2004.

[4] T. Virtanen, and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 1757-1760. Orlando, 2002.

[5] J. Woodruff, B. Pardo and R. Dannenberg, "Remixing stereo music with score-informed source separation," *International Conference on Music Information Retrieval (ISMIR)*, pp. 314-319. Victoria, 2006.

[6] M. Bay and & W. J. Beauchamp, "Harmonic source separation using prestored spectra," In *Lecture notes on computer science*, Vol. 3889, pp. 561-568. Berlin: Springer, 2006.

[7] L. Fritts, (1997). *The University of Iowa Electronic Music Studios. Musical instrument samples*. Retrieved January 10, 2009, from http://theremin.music.uiowa.edu/MIS.html

[8] M. Karjalainen, and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing.* Phoenix, 1999.

[9] T. Painter, and A. Spanias, "Perceptual coding of digital audio," *Proceeding of the IEEE , 88* (4), pp. 451-515, 2000.

[10] International Organization for Standardization (ISO). (2003). Acoustics. Normal equal- loudness-level contours.

[11] A. H. Benade and S. J. Lutgen,"The saxophone spectrum". *Journal of the Acoustical Society of America , 83* (5), pp. 1900-1907, 1988.