A VIOLIN MUSIC TRANSCRIBER FOR PERSONALIZED LEARNING

Wei Jie Jonathan Boo, Ye Wang, Alex Loscos

Department of Computer Science, School of Computing
National University of Singapore, Singapore 117543
{booweiji, wangye, loscos}@comp.nus.edu.sg

ABSTRACT

This paper presents a new version of our violin music transcriber [1] to support personalized learning. The proposed method is designed to detect duo-pitch (two strings being bowed at the same time) from real-world violin audio signals recorded in a home environment. Our method uses a semitone band spectrogram, a signal spectral representation with direct musical relevance. We exploit constraints of violin sound to improve the transcription performance and speed in comparison with existing methods. We have carried out rigorous evaluations using (a) single pitch notes and duo-phonic pitch samples within the violin's playing range (G3-B6), and (b) music excerpts. For pitch and duo-pitch samples our method can achieve a transcription precision score of 93.1% and recall score of 96.7% respectively. For music excerpts, an average of 95% of all notes could be found (recall), and 93% of notes transcribed correctly (precision).

1. INTRODUCTION

Automatic transcription of music refers to the automatic analysis and extraction of parameters from a music signal that would sufficiently describe it. Despite attempts dating back to the 1970s [2, 3] and recent progresses, it remains a challenging research problem. Its nontrivial nature is reflected in the variety of proposed methods, ranging from techniques emphasizing computation efficiency to those emulating theoretical models of human music perception. Recent techniques include Kashino's integration of psychoacoustics processing with the Bayesian probability network [4], and Martin's blackboard architecture system utilizing musical rules to transcribe four-voice piano compositions [5]. Another system that imposes some constraints is Goto's partial transcription of melody and bass line [6]. Klapuri, who has actively contributed to music transcription [7, 8], recently proposed a method transcribing pitched notes without restricting the original signal by modeling note events, silence and musicology [9].

Systems with more significant success typically involve techniques that incorporate some constraints; hence, the challenge is to build a system where restrictions, if any, make perfect 'musical' sense. For our intended applications in personalized violin education [1], it would be sufficient if the transcriber can extract the pitches and other relevant audio features specific to the targeted instrument. This reasoning has motivated our current work on violin music transcription. The violin has been chosen because it is the most common instrument used in classical music of various forms. It is also the second most popular solo instrument after the piano for music learners. It is commonly recognized that violin is much harder than piano to start with and it may take a beginning violin learner years to simply produce the correct pitch.

This paper presents a significantly improved version of our earlier transcriber which works only for monophonic violin signal [1]. The proposed transcriber works now for the entire playing range of the violin (G3-B6), and is capable of duo phonic transcription – the maximum music theoretical polyphony for the violin [10]. Sections 2 detail our transcription system. Section 3 presents our evaluation results. We conclude with a summary of this work and possible future improvements in Section 4.

2. SYSTEM DESCRIPTION

The proposed violin transcriber has two main building blocks, namely note segmentation detection and pitch estimation (see Figure 1).

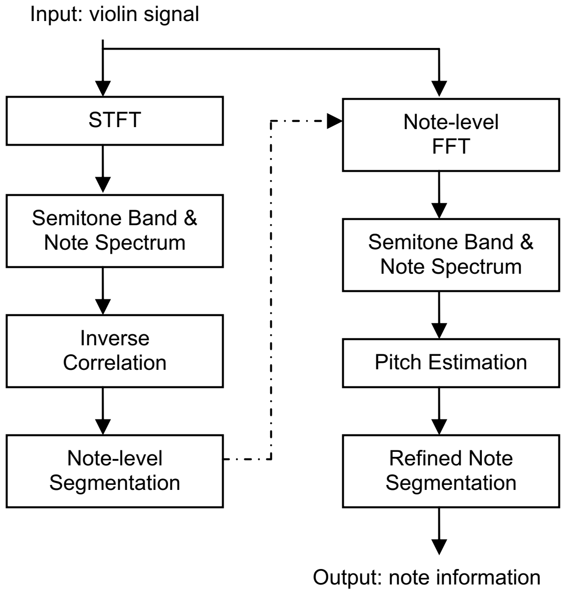

Figure 1: block diagram of proposed system

We have relaxed the requirement of system complexity in order to overcome an important limitation of our earlier system: In addition to a fairly low time resolution determined by the analysis window length, our earlier system could not be easily scaled up to transcribe duo-pitch violin sound. To solve this problem, we perform a STFT with shorter fixed length FFT (2048 PCM samples) and 75% overlap. This is followed by the creation of a note spectrogram via summation of relevant frequency components in a semitone band spectrogram. Initial note segmentation onset detection is carried out by means of an inverse autocorrelation function of adjacent frames note spectrograms. We then use an adaptive window length determined by subsequent pairs of note boundary points for pitch estimation through note spectrogram. In order to assure note detection in cases such as two fast consecutive notes with the same pitch, a step to refine note segmentation is carried out before the note information is given out. Note information includes note pitch, note start time, note duration, and note strength.

2.1. Semitone Band and Note Spectrogram Creation

Fast Fourier Transform (FFT) is first performed to convert the time domain signals into a frequency domain representation. The linear frequency representation is then mapped to an equal temperament western musical scale to create a semitone band spectrogram, which models the human perception of pitch, [11, 1].

In a semitone band spectrum, the second to the fifth harmonics lie in the semitone bands 12, 19, 24 and 28 semitones away from the fundamental. Denoting *Z(n)* as the semitone band spectrum, *A[n]* is the estimated amplitude level of a musical note with semitone index *n.* In our earlier system [1], *A[n]* was obtained by adding up the estimated amplitude levels of harmonics which fall on specific semitone indexes *n+k* where *k = 0, 12, 19, 24 and 28*:

$$A[n] = \sum Z[n+k] \qquad (1)$$

One of the problems of above method is that it can cause sub-harmonics of the signals to be amplified. To mitigate the problem, we carry out harmonic summation using the following expression [11]:

$$A[n] = \sum \min(\alpha Z[n], Z[n+k]) \qquad (2)$$

In the expression, α ensures that the power of the harmonics added does not exceed the power at the fundamental by more than a factor of α. We choose $\alpha = 5$ for our method because this value seems to give the best transcription performance. By replacing (1) with (2), we can improve our transcription performance by 7%. Repeating the above operation on musical notes ranging from G3 to B6, we map the semitone band spectrogram to a note spectrogram that shows the distribution of amplitude levels for each note. We have chosen the range because of the following: G3 is the lowest note that our pitch detector would need to detect. All violins are expected to have a frequency range of at least three octaves where the ‘practical’ pitch range reaches G6 [12]. Hence, it is sufficient for the range of our note spectrogram to be set at slightly more than three octaves for onset and pitch detection.

2.2. Note-level segmentation using Note Spectrum Inverse Correlation

Note level segmentation is very much related with onset detection since note onsets define the start time of a new musical event. While onsets may include changes in expressive features and timbre, for our purpose, an onset is defined in this paper as the time instance when, from a current note, there is (a) a pitch change to the next note, or (b) a transition to a subsequent note of the same pitch.

Previous onset detection techniques include the use of high frequency content, spectral difference and phase deviation amongst many others. An extensive coverage is provided in [13]. Correlation coefficients, numbers between –1.0 and 1.0, give us a measure statistically of how strong pairs of variables are related. The larger the coefficient is, the greater the variables are related. Our experiments have shown that calculating the correlation coefficient between subsequent frames of a note spectrogram is reliable for detecting pitch changes. This is extremely useful for the violin because unlike instruments such as the piano or guitar, onsets may occur without strong energy changes.

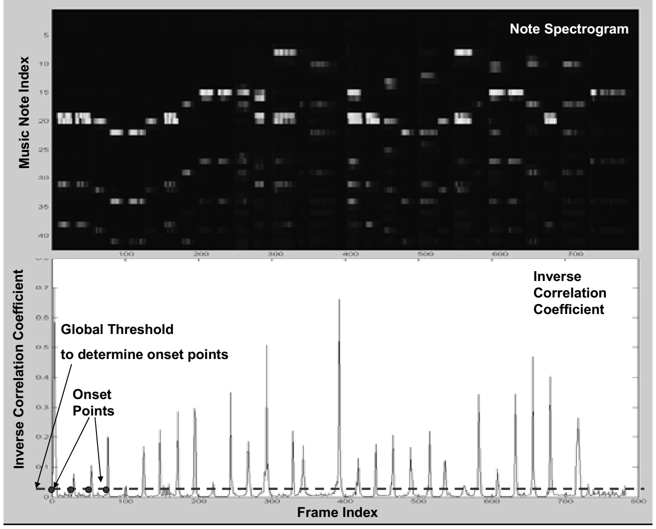

Figure 2: analysis of a music excerpt showing (top) note spectrogram as derived from semitone spectrogram, and (bottom) inverse correlation of note spectrum and resulting onset instants

The principle of our onset detector is illustrated in Figure 2. A global threshold is set where onsets are determined to have occurred when the inverse correlation coefficients exceed the threshold. In regions where several successive coefficients exceed the threshold, the first point is taken as the onset point. The processed audio signal is segmented based on the onset points where pitch estimation is performed on each segment.

2.3. Pitch Estimation

Violin monophonic transcription is not sufficient as the violin is capable of producing two notes simultaneously [10]. To perform duo-phonic detection, we extended the semitone band spectrum pitch detection algorithm described in [1]. If a detected dominant pitch P_1 with $P_1 = max(A[n])$ is reduced, a second pitch P_2, if present, would now have the maximum $A[n]$, and thus could be identified. This is similar to the principle used in [9], where a detected sound is removed from the mixture, and pitch detection performed on the residue. In our system, we reduce the estimated amplitude levels of the harmonics (semitone band indexes) that form P_1 to one-fifth its original value. This is the best reduction level based on our experiments. At this level, pitch P_2, if present, may be detected on the residue semitone band spectrum.

Octave errors occur where instead of the actual pitch, a pitch an octave away is detected. Traditional spectral-location based approaches are prone to pitch halving, and spectral-interval based approaches to errors in pitch doubling [8]. Through our improved harmonic summation, we have also implemented a reliable octave error detector. With $\alpha = 5$, if a pitch P is detected and $2*A[P-12] > A[P]$, we can conclude an octave error has occurred, and we set P to be one octave lower i.e., $P = P-12$.

2.4. Refined Note segmentation

For violin sound, onset detection, especially the detection of subsequently played notes of the same pitch, is significantly more difficult compared to other instruments (*e.g.,* percussion instruments). This is because notes amplitude related attributes change significantly from a soft to a relatively hard performing style.

To tackle this problem, we use three criteria for onset detection:

1. There is a pitch change.
2. The amplitude tops a predetermined loudness threshold.
3. There is an amplitude valley beyond a predetermined threshold.

If any of the criteria is fulfilled, we consider the candidate valid for an onset.

While the onset detector in Section 2.2 could detect pitch changes (criteria 1) reliably, a step of refined note segmentation is required to check criteria 2 and 3. The last criterion, in particular, is targeted at finding soft onsets that occur when subsequent notes of the same pitch are played. An improvement in this paper over [1] is the evaluation of criteria 2 and 3 with direct relevance to human pitch perception by expressing the total estimate amplitude level of a note *A[n]* in a decibel scale.

From experimentation, for criteria 2, we have set loudness threshold to 10dB, and for criteria 3, we request amplitude valley to be at least 10dB. These levels work well across a wide variety of our test musical excerpts.

3. EVALUATION

Evaluation comprises two phases; we test the system's transcription ability with: (a) monophonic and duo-phonic violin samples, and (b) solo violin excerpts.

3.1. Monophonic and Duo-Phonic Samples

The test data consists of all possible monophonic and duo phonic violin sound samples from G3 to B6. We evaluated every possible combination up to an interval of 16 semitones, which is the maximum interval a typical violin player can play given the fingers' physical stretching limitations. Overall,

	P.I	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
P.I		D4	D#4	E4	F4	F#4	G4	G#4	A4	A#4	B4	C5	C#5	D5	D#5	E5	F5	F#5	G5	G#5	A5	A#5	B5	C6	C#6	D6	D#6	E6	F6	F#6	G6	G#6	A6	A#6	B6
*	G3	*	0	*	0	0	╳	0	*	*	*																								
2	G#3	0	*	0	0	0	0	╳	0	0	0	0																							
3	A3	0	*	*	*	0	*	*	╳	0	*	*	0																						
4	A#4	0	0	*	0	0	*	0	0	╳	*	0	*	*																					
5	B3	0	*	0	*	0	*	*	*	0	╳	*	0	*	*																				
6	C4	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	*																			
7	C#4	0	*	*	*	0	*	0	*	*	*	*	╳	*	*	0	*																		
8	D4		*	*	*	*	*	*	0	*	*	*	*	╳	*	*	*	*																	
9	D#4			*	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	*																
10	E4				0	*	*	*	*	*	*	*	*	*	*	╳	0	*	*	*															
11	F4					*	*	*	*	*	0	*	*	*	*	0	╳	0	0	*	*														
12	F#4						*	0	0	*	*	*	*	*	*	*	0	╳	*	*	*	*													
13	G4							*	*	*	0	*	*	*	*	*	*	*	╳	*	*	*	*												
14	G#4								*	*	*	0	*	*	*	*	*	*	*	╳	*	*	0	*											
15	A4									*	*	0	*	*	*	*	*	*	*	*	╳	*	*	*	*										
16	A#4										*	*	0	*	0	*	*	*	*	*	*	╳	0	*	0	*									
17	B4											*	*	*	*	*	*	*	*	*	*	*	╳	*	0	*	*								
18	C5												*	*	0	*	*	*	*	*	*	*	*	╳	*	*	0	*							
19	C#5													*	0	0	0	*	*	*	*	*	*	*	╳	0	*	*	0						
20	D5														*	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	0					
21	D#5															*	*	*	*	*	*	*	*	*	*	*	╳	*	*	0	*				
22	E5																*	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	*			
23	F5																	*	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	*		
24	F#5																		*	*	*	*	0	*	*	*	*	*	*	╳	*	*	*	*	
25	G5																			*	*	*	*	*	*	*	*	*	*	*	╳	*	*	*	*
26	G#5																				0	*	*	*	0	*	*	*	*	*	*	╳	*	0	0
27	A5																					*	*	*	*	*	*	*	*	*	*	*	╳	0	*
28	A#5																						*	*	*	*	*	*	*	*	*	*	*	╳	0
29	B5																							*	*	*	*	*	*	*	*	*	*	*	╳
30	C6																								*	*	*	*	*	0	*	*	*	*	*
31	C#6																									*	*	*	0	*	*	*	*	*	*
32	D6																										0	0	*	0	*	0	*	*	*
33	D#6																											*	0	*	*	*	0	*	0
34	E6																												*	*	*	*	*		*
35	F6																													*	*	*	*	*	*
36	F#6																														*	0	*	*	*
37	G6																															*	*	*	0

P.I = Pitch Index
*=Correctly Detected
0=Mistake in detection
╳ Not Considered

Figure 3: performance of proposed system with pitch samples

our system *can achieve a precision score of 93.1% and recall score of 96.7%.* The test results are summarized in Figure 3.

3.2. Short Music Excerpts

The test data consists of short excerpts of solo violin pieces. The metrics of evaluation are:

$$precision = ({}^{N_c}\!/_{N_d}) \times 100\%$$

$$recall = ({}^{N_c}\!/_{N_{orig}}) \times 100\%$$

N_{orig} represents the numbers of notes to be detected in the music excerpt, and N_d the number of notes detected by the system. Out of the notes detected, N_c represents the notes correctly detected. Figure 4 excerpts our results, where an average of 95% of all notes were found (recall), and 93% of notes transcribed correctly (precision).

Title of Music Excerpt	Pitch Combination	N_{orig}	N_d	N_c	N_u	N_w	Precision (%)	Recall (%)
Traditional, Twinkle Twinkle Little Star (Arranged for solo violin)	Single Pitch	42	40	40	2	0	100	95.2
Beethoven, Symphony No. 9 in D minor, Op. 125, 4th Movement (Arranged for solo violin)	Single Pitch	62	56	56	6	0	100	90.3
Copland's Shaker Melody	Single Pitch	87	92	79	8	13	85.7	90.1
J.S Bach, Partita No. 3 in E, BWV 1006, 4-Minuet I	Single Pitch & Double Pitch	48	49	47	1	2	95.9	97.9
J.S Bach, Partita No. 3 in E, BWV 1006, 3-Gavotte en eondeau	Single Pitch & Double Pitch	53	56	53	0	3	94.6	100
Beethoven, Symphony No. 9 in D minor, Op. 125, 4th Movement (Arranged for solo violin with double stop improvising)	Single Pitch & Double Pitch	41	44	38	3	6	86.4	92.7
Traditional, Twinkle Twinkle Little Star (Arranged for solo violin with double stop improvising)	Single Pitch & Double Pitch	67	72	66	1	6	91.2	98.5

Figure 4: performance of proposed system with short music excerpts

4. CONCLUSION AND FUTURE WORK

We have presented a new violin sound transcriber designed for music education applications. It utilizes a semitone band spectrogram to perform violin transcription with high accuracy and speed. Such a signal representation is not only musically relevant; it also accounts for the fact that while string sounds belong to the class of harmonic sounds, higher-order partials suffer from the inharmonicity and tend to shift upwards slightly in frequency [8].

The system can be improved in several ways. First, its detection accuracy could be further enhanced. Second, the current system does not consider octave detection; hence, an immediate extension is for it to perform duo-phonic octave transcription. Another improvement is to extend the method to detect duo phonic violin vibrato. Finally, violin music playing is limited not only by instrument characteristics, but also human physical limitations. For example, certain note jumps may not be physically possible. Hence, incorporating such knowledge into a transcription system could significantly enhance its accuracy.

5. REFERENCES

[1] J. Yin, Y. Wang, D. Hsu, "Digital Violin Tutor: An Integrated System for Beginning Violin Learners" ACM Multimedia, Hilton, Singapore, 2005.

[2] J.A. Moorer, On the segmentation and analysis of continuous musical sound by digital computer. PhD thesis, CCRMA, Stanford University, 1975.

[3] M. Piszczalski and B. A. Galler, "Predicting musical pitch from component frequency ratios," J. Acoust. Soc. Am., 66 (3), 710–720, 1979.

[4] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism," in Proc. International Joint Conf. on Artificial Intelligence, Montréal, 1995.

[5] K. D. Martin, "Automatic transcription of simple polyphonic music: robust front end processing," MIT Media Laboratory Perceptual Computing Section Technical Report No. 399, 1996.

[6] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," Speech Communication, vol. 43, no. 4, pp. 311–329, 2004.

[7] Klapuri, A, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, 2004.

[8] Klapuri, A, "Automatic music transcription as we know it today," Journal of New Music Research, Vol. 33, No. 3, pp. 269-282, Sep. 2004.

[9] Ryynänen M., Klapuri A, "Polyphonic music transcription using note event modeling ," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, New York, Oct. 2005.

[10] E. Taylor, The AB Guide to Music Theory Part 2. London: The Associated Board of the Royal Schools of Music, 1991, ISBN: 1-85472-447-9.

[11] J. McNames, C. Crespo, M. Aboy, J. Bassale, L. Jenkins, B. Goldstein, "Harmonic spectrogram for the analysis of semi-periodic physiologic signals," In Proc. the Second IEEE Joint EMBS/BMES Conference, Houston, TX, pp. 143-144, 23-26 October 2002.

[12] William F. Lee, "Music Theory Dictionary: The Language of the Mechanics of Music", Charles Hansen Educational Music & Books, 1966, ASIN: B0007G6ALA.

[13] J. P. Bello, L. Daudet, S. Abadía, C. Duxbury, M. Davies and M. B. Sandler, "A tutorial on onset detection in music signals", IEEE Transactions on Speech and Audio Processing. September, 2005.