Digital Violin Tutor: An Integrated System for Beginning Violin Learners

Jun Yin, Ye Wang and David Hsu

School of Computing
National University of Singapore
{yinjun, wangye, dyhsu}@comp.nus.edu.sg

ABSTRACT

Prompt feedback is essential for beginning violin learners; however, most amateur learners can only meet with teachers and receive feedback once or twice a week. To help such learners, we have attempted an initial design of Digital Violin Tutor (DVT), an integrated system that provides the much-needed feedback when human teachers are not available. DVT combines violin audio transcription with visualization. Our transcription method is fast, accurate, and robust again noise for violin audio recorded in home environments. The visualization is designed to be intuitive and easily understandable by people with little music knowledge. The different visualization modalities—video, 2D fingerboard animation, 3D avatar animation—help learners to practice and learn more effectively. The entire system has been implemented with off-the-shelf hardware and shown to be practical in home environments. In our user study, the system has received very positive evaluation.

Categories and Subject Descriptors

H.5.1 **[Multimedia Information Systems]** Animations

H.5.5 **[Sound and Music Computing]** Signal analysis, synthesis, and processing, Systems

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Music transcription, vibrato detection, score alignment, animation, violin tuner

1. INTRODUCTION

For beginners learning to play a musical instrument, prompt feedback is essential. However, constrained by time and financial resources, most amateur learners can only afford to meet with teachers and receive feedback once or twice a week. In the rest of the time, they must practice on their own. "Practice" becomes a dirty word: the lack of feedback makes practice boring and frustrating, which leads to slow progress. This issue is accentuated for children, who need more guidance, while their parents, who often do not play the instrument, lack the knowledge to help.

To address this critical issue, we have attempted the initial design of a system called Digital Violin Tutor (DVT) to help beginning violin learners, in particular, young children, along with their parents. In the absence of human teachers, DVT performs basic, yet key activities of a violin lesson, including helping to tune the violin, pointing out the mistakes in learner's play, and demonstrating the correct play. This is achieved by combining audio, video, and 3D animation to form an effective feedback loop (Figure 1). By exploiting the synergy of a multi-modal system, DVT attempts to offer beginning violin learners a stimulating learning environment that increases their interest.

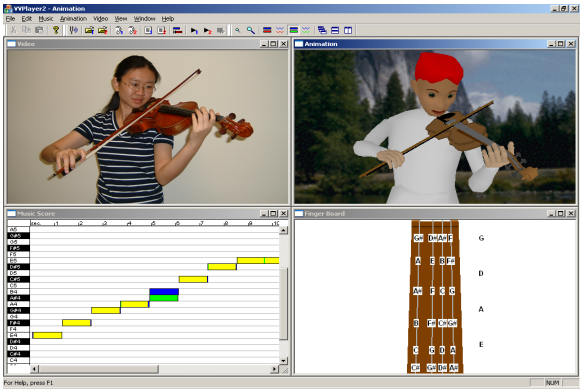


Figure 1: The user interface of Digital Violin Tutor. (*a*) Video of teacher's play. (*b*) 3D avatar animation. (*c*) Visualization of mistakes in learner's play. (*d*) 2D fingerboard animation.

The key design objective of DVT is effective and immediate feedback to beginning violin learners practicing at home. To achieve this, first, we have designed a system module for transcribing violin audio with high accuracy and speed. It is also robust against noise, which is common in audio recorded with low-quality microphones in home environments. Second, to identify mistakes in learner's play, another system module compares the transcribed music notes against the score or teacher's play, and the results are given to the learner through effective visualization. The visualization of mistakes is designed to be intuitive and easily understandable by people with little music knowledge, *e.g.,* parents who wish to help their children. Our experience with the system indicates that this combination of audio analysis and mistake visualization is natural and effective. Finally, the entire system is implemented with off-the-shelf

MM'05, November 6–11, 2005, Hilton, Singapore.

hardware—basically, a PC and a cheap microphone—so that it can be easily made available in home environments. In our user study, the system has received very positive evaluation.

DVT has several advantages over existing educational tools for beginning learners, such as audio or video recordings of teacher's play. Most importantly, recordings do not provide any explicit feedback. Without feedback, beginning learners often lack the knowledge to find their own mistakes. Even for the simple task of tuning the violin, although the traditional aid, a tuning whistle, generates a sound at the correct pitch for each string, many beginners cannot hear the difference between the correct pitch and that produced by the string being tuned. Visualization of the difference greatly simplifies the task.

We would like to emphasize that DVT is not intended to replace human teachers. Instead, it fulfills a critical need of beginning violin learners by providing basic, yet effective feedback, when human teachers are not available.

In this work, we have chosen violin, because it is an important and popular instrument; however, most of the previous work [3][10][13] is devoted to keyboard instruments, in particular, piano. We are not aware of similar systems for violin. Although we focus on violin, the main design considerations apply to other related string instruments as well.

The paper is organized as follows. Section 2 reviews previous work. Section 3 gives an overview of DVT. Sections 4–7 describe the main system modules. Section 8 presents the system evaluation based on a user study. Section 9 concludes with possible improvements on the system in the future.

2. PREVIOUS WORK

Many computer systems have been proposed to assist music education. Piano Tutor [3] is a comprehensive system for piano instruction. It performs score-tracking and error analysis based on MIDI input and draws upon an expert system to coordinate the presentation of lessons from a database. PianoFORTE [13] is another piano instruction system, which aims at enhancing the learner's ability in music interpretation. It displays the progression of learner's play on a music score. This display format assumes that the learner is proficient in reading music notations and is thus only suitable for more advanced learners and not for beginners or parents who have little music knowledge. A more recent system, Family Ensemble [10], tries to increase children's interest in practicing piano by providing duo play. It uses score-tracking to generate accompaniment automatically and helps a parent to "play" a MIDI keyboard together with a child.

Most of the existing systems, including the ones described above, are devoted to keyboard instruments, *e.g.,* piano. They usually use a MIDI keyboard for input and circumvent the difficult problem of music transcription. Unfortunately, this approach does not work for string instruments such as violin: the equivalent of a MIDI keyboard does not seem to exist for violin, especially in a home environment. To solve this problem, our DVT system uses a fast and accurate music transcription method to deal with acoustic music input directly. In comparison with Piano Tutor, DVT is not a comprehensive instruction system. It targets beginners, especially children, along their parents, by providing effective visual feedback. The visualization is easily understandable by people who have little music knowledge.

3. SYSTEM OVERVIEW

After consulting with a professional violin teacher and our collaborators from our university's Conservatory of Music, we determined three main functions for the initial design of DVT: (1) helping to tune the violin, (2) pointing out the mistakes in learner's play, and (3) demonstrating the correct play. These are deemed the most basic and most important activities during a violin lesson for beginners.

To perform these functions, DVT consists of several interconnected system modules (Figure 2):

- The *transcriber* transcribes audio input from learner's play into a note table. Each table entry corresponds to a note with information on its pitch, loudness, onset, and duration. An entry may also contain information on playing styles, such as vibrato, *i.e.*, small cyclic change in pitch.
- The *performance evaluator* tries to identify the mistakes in learner's play by comparing the transcribed notes with either the score or transcribed teacher's play. It then visualizes the mistakes and provides feedback to the learner.
- The *tuner* uses the transcriber and simple visualization to help the learner tune the violin.
- The *animator* demonstrates the correct play through both 2D fingerboard animation and 3D avatar animation. The animation is driven directly by a score or by a note table transcribed from teacher's play.

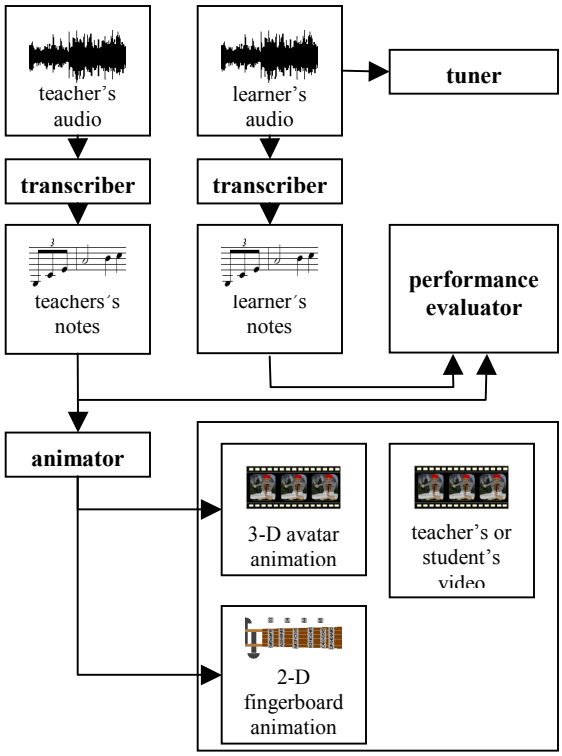


Figure 2: System diagram.

DVT can also show the teacher's demonstration video along with the animations. The different modalities—video, 2D fingerboard

animation, and 3D avatar animation—help the learner to practice and learn more effectively.

4. TRANSCRIBER

The transcriber is an essential component of our system. It enables DVT to deal with acoustic signals from microphones or CD recordings, rather than symbolic signals from MIDI keyboards, which do not seem to exist for violin.

4.1 Overview

Music transcription is difficult. Despite decades of research and significant progress (*e.g.,*. [8][9][11]), reliable transcription remains a challenge in practical applications. Recent work on transcription focuses on transcribing commercial CD music, which is recorded professionally. However, DVT must process audio input recorded with low-quality microphones in home environments. Such audio input is typically *noisy*. Furthermore, beginners tend to make a variety of mistakes, resulting in audio signal with irregular patterns. All these make it challenging to develop a fast and accurate transcriber for our system.

For our particular application, we have focused on three design objectives for the transcriber: *accuracy*, *robustness* against noise, and *speed*. Accuracy is important, because an inaccurate transcriber cannot be effective in providing feedback to learners. Robustness against noise is important, because sound recorded with low-quality microphones in a home environment is usually noisy. Speed is important, because learners are unlikely to be willing to wait long. To be useful, the feedback must be almost instantaneous.

A block diagram of our transcriber is shown in Figure 3. The input is PCM-encoded acoustic signal, and the output is a note table which contains information on the pitch, loudness, onset, and duration of the notes played. The upper path in Figure 3 shows the basic transcription process. In each time frame, the transcriber first converts the PCM signal to the frequency domain using FFT. It then maps the FFT spectrum to an energy-based *semitone spectrum* [15] (Figure 4) and estimates the pitch and loudness based on the harmonic structure of violin sound. The transcriber then performs onset detection and duration estimation. An entry of the note table is then generated for each note detected. The lower path in Figure 3 shows the steps of our high-resolution pitch estimation in DVT. More details of our transcriber are described in the following subsections.

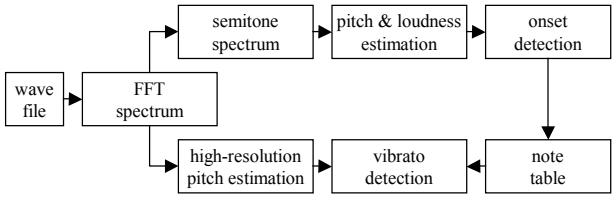


Figure 3: Block diagram of the DVT transcriber.

4.2 Accurate, Robust, and Fast Transcription

Our transcriber uses the idea of harmonic-structure tracking, similar to the algorithm described in [9], but extends the idea by taking advantage of special characteristics of violin sound. These special characteristics are the key to solving our transcription problem accurately and efficiently. Our transcriber makes three assumptions. First, it limits the pitch range between G3 and G6. Second, the harmonic structure of violin sound has the property that the energy of the sound is concentrated at the fundamental and the first five harmonics at 12, 19, 24, 28 and 31 semitones away from the fundamental. Finally, we limit the transcription to monophonic violin sound only. In general, these assumptions hold well for beginning learners.

For pitch and loudness estimation, the transcriber first converts the FFT spectrum of violin sound into the semitone spectrum. The conversion has several advantages. In contrast to the linear scale of the FFT spectrum, the semitone spectrum is a logarithmic scale, which is commonly used in musical notation and matches closely with human pitch perception. The conversion reduces the search space for pitch estimation, resulting in faster transcription speed. It also increases noise robustness, because the energy in semitone bands is far more stable than the individual magnitude in the FFT domain. After the conversion, we perform the pitch and loudness estimation in the semitone domain. Our testing shows that the semitone band energy spectrogram is robust in the presence of noise and other hostile factors, such as missing fundamental or unstable harmonic structure, both of which are typical for violin sound.

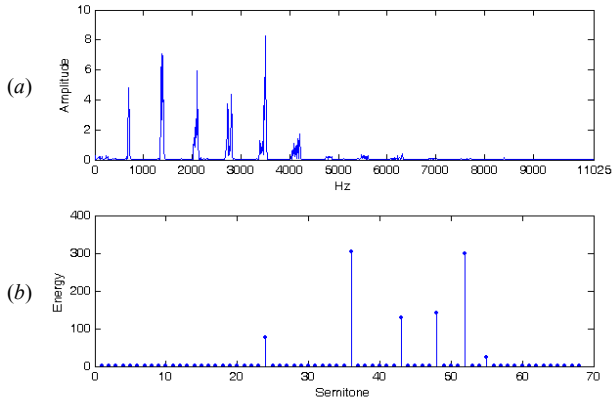


Figure 4: (*a*) FFT spectrum. (*b*) Semitone band energy spectrum.

After pitch and loudness estimation, the transcriber performs onset detection. Onset detection for violin sound is more difficult than that for other instruments (*e.g.,* percussion instruments). The onset of violin sound varies significantly from a soft one to a relatively hard one. To tackle this problem, we first normalize the loudness to account for variations in loudness of sound recorded in different environments and then use three criteria for onset detection. If any of the criteria is met, we consider it a valid candidate for an onset.

- The first criterion detects hard onset. It estimates the derivative of the loudness function using a finite difference method. It finds an onset if the derivative is above a constant threshold.
- The second criterion detects soft onset, which occurs frequently in violin sound. It finds an onset if the loudness is above a constant threshold.
- The third criterion checks the loudness change in a small surrounding window to distinguish two consecutive notes of the same pitch.

These three situations are illustrated in Figure 5.

Finally, a note table is generated. See Table 1 for an example.

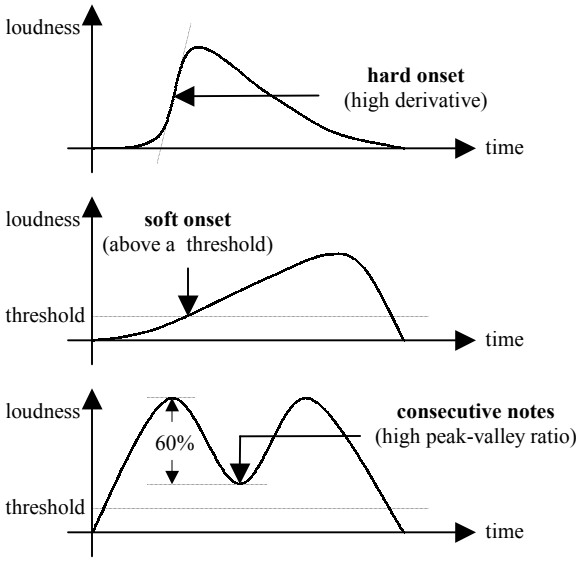


Figure 5: Three criteria for onset detection.

Onset (O)	Duration (D)	Pitch (P)	Loudness (L)
42	113	24	0.799
160	23	20	1.237
182	22	15	2.898
…	…	…	…

Table 1: An example basic note table.

4.3 Performance of the DVT Transcriber

We tested our transcriber with 15 violin pieces, of which 3 are professional recordings and 12 are ordinary home recordings. Three criteria are used to evaluate the accuracy of our transcriber:

$$P = \text{precision} = \frac{\#\text{ of correctly detected notes}}{\#\text{ of total detected notes}}$$

$$R = \text{recall} = \frac{\#\text{ of correctly detected notes}}{\#\text{ of notes in the input music}}$$

$$F_1 = \frac{2PR}{P+R}$$

$$0 \le F_1 \le 1$$

The test results (Table 2) show that by taking advantage of the specific characteristics of violin sound, our transcriber performs much better than commercial software for general-purpose music transcription.

Transcriber	Precision	Recall	F_1
DVT	0.97	0.99	0.98
Amazing MIDI v1.70 [16]	0.23	0.96	0.37
intelliScore v5.1 [17]	0.29	0.94	0.44

Table 2: Transcription accuracy.

Our transcriber is not only accurate, but also fast. The combination of searching in the semitone domain and tracking only the first six most significant frequency components enables our transcriber to achieve real-time performance. The transcription speed is shown in Figure 6.

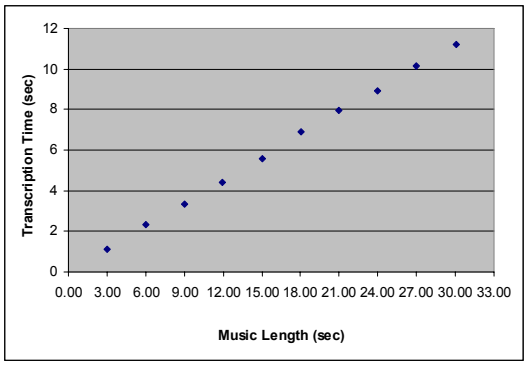


Figure 6: Transcription speed.

4.4 High-Resolution Pitch Estimation and Vibrato Detection

Vibrato is a distinct characteristic of violin performance. It is not always marked in music scores. Learners must learn by listening to teacher's play. To help beginners to easily find vibrato in teacher's play, DVT uses high-resolution pitch estimation to detect vibrato and visualize the results.

Vibrato is defined as small oscillations in pitch and/or loudness of a musical tone [1]. The oscillations in pitch are usually smaller than half of a semitone away from the central pitch. In order to detect the oscillations and thus the vibrato, semitone level resolution pitch estimation described in Section 4.2 is not adequate. Pitch estimation for detecting the oscillations requires the resolution at the *cent* level, defined as 1/100 of a semitone.

There are several methods to increase pitch estimation resolution, *e.g.,* enlarging FFT window and peak picking [6]. However, a large FFT window reduces not only the transcription speed, but also the time resolution. Peak picking is found to be unreliable, due to the irregularity of violin harmonic structure, especially for sound recorded in a home environment.

DVT uses a dynamic linear interpolation method, which offers good tradeoff between increased resolution in pitch estimation and transcription speed. Denote $Y[0..4095]$ as the 4096-point FFT amplitude spectrum obtained during our semitone level pitch estimation. Denote $Y(c)$ as a function that returns the amplitude of frequency at cent c in the spectrum. If c is non-integer, $Y(c)$ returns the linear interpolated value from $Y[0..4095]$. We estimate the cent of the pitch by the formula below:

$$c = \arg\max_i \sum_{k=1}^{6} Y(k \cdot i) \quad c_{\min} \le i \le c_{\max}.$$

For each cent from the lowest to the highest, we compute the sum of the amplitudes for its fundamental and the first five harmonics, and choose the one with largest sum as the estimated pitch. The principle of our algorithm is illustrated in Figure 7. Again, we have exploited the harmonic structure of violin sound in our hig-resolution estimation.

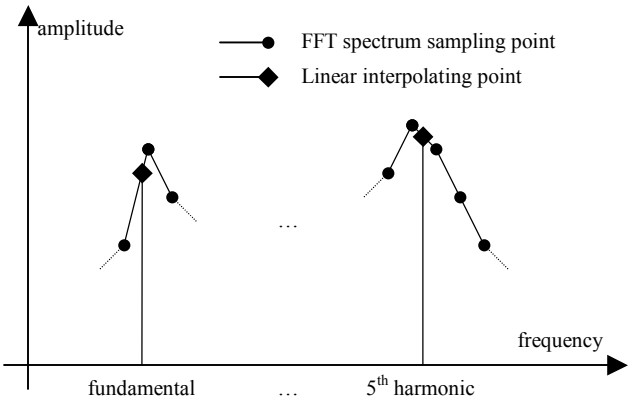


Figure 7: Linear interpolation for the fundamental and first fifth harmonics.

The estimated cents of all frames of a note form a pitch trajectory. We then use minima-maxima detection [12] to find vibrato. In our implementation, vibrato is considered detected if the following two criteria are both satisfied:

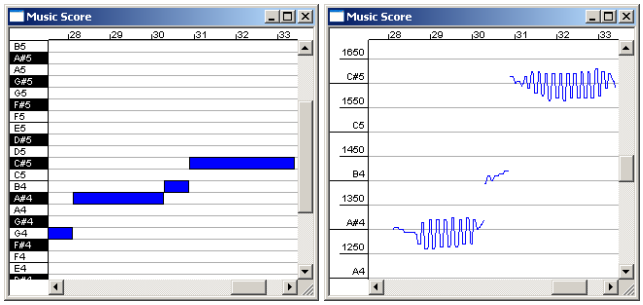
- a minimum of three oscillation cycles;
- a minimum average oscillation amplitude.

In our current system, we assume at most one vibrato for each note. This results in four parameters for each detected vibrato: starting frame, duration, strength (frequency shift), and vibrato frequency (the number of oscillation cycles in a unit time). Our vibrato transcriber outputs a note table with vibrato parameters. See Table 3 for an example.

DVT visualize the notes and vibratos (Figure 8) in distinct ways, so that learners can easily see what and how notes are played.

O	D	P	L	Vibrato Start	Vibrato Length	Vibrato Strength	Vibrato Frequency
…	…	…	…	296	50	51	0.140000
…	…	…	…	0	0	0	0.000000
…	…	…	…	419	64	51	0.140625
…	…	…	…	0	0	0	0.000000

Table 3: An example vibrato note table.



(*a*) (*b*)

Figure 8: (*a*) Visualized notes. (*b*) Visualized vibratos.

5. TUNER

The DVT tuner uses the transcriber and simple visualization to help the learner tune the violin. It detects the pitch of the violin sound from a microphone and displays it in real time. The learner chooses a string to be tuned from the user interface and bows the corresponding string near a microphone. The tuner samples the audio from the microphone at the rate of 22 kHz and creates the corresponding spectrum using non-overlapping 4096-sample FFT in order to satisfy the real-time requirement. The pitch is then estimated with the high-resolution method described in Section 4.4. The tuner displays the detected pitch continuously (Figure 9). The darker color (blue) indicates that the pitch is too high. The lighter color (green) indicates that the pitch is too low.

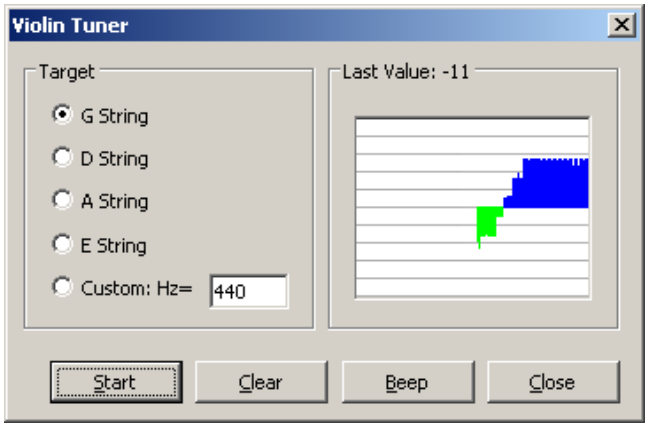


Figure 9: The tuner user interface.

Our pitch visualization differs from traditional techniques in a simple, but important way. While most traditional techniques, which are needle-based, show only the *instantaneous* estimated pitch, our visualization shows a short *history* of the estimated pitch as well. This simple feature, somewhat unexpectedly, seemed to reduce tuning time and received enthusiastic welcome in our user study. Our conjecture is that the pitch history is used by the learner to predict the future pitch and avoid overshooting the target. This is similar to the idea of look-ahead in control theory. We plan to carry out controlled experiments in the future to confirm this observation.

6. PERFORMANCE EVALUATOR

After transcribing learner's play into a note table, DVT compares it with a note table based on a score or one transcribed from teacher's play, in order to find the learner's mistakes.

6.1 Note Alignment

Our note table is basically an annotated list of notes. It is well known that simple Euclidian distance is not useful for comparing two lists of notes, as the learner may play at a different starting time and with a different tempo. To find the learner's mistakes in a meaningful way, we must first *align* the notes in the two tables first.

Our note alignment is based on the dynamic programming (DP) algorithm, similar to those used for score tracking [2][10]. It performs an initial alignment using only pitch information and then adjusts the tempo to produce the final alignment.

In the DP algorithm, we define three types of mistakes in learner's play:

- *Mismatch*. The pitch of the note that the learner plays is different from that of the note that the teacher plays.
- *Insertion*. The learner plays an additional note that should not be played.
- *Deletion*. The learner misses a note that should be played.

Each type of mistake is given a fixed cost.

Define $S[1..m]$ as the pitch list of the learner's notes and $T[1..n]$ as the pitch list of the teacher's notes. A DP alignment matches notes in $S[1..m]$ with those in $T[1..n]$. Define $V[i,j]$ as the cost for an optimal alignment between $S[1..i]$ and $T[1..i]$, one with the minimum cost. To initialize DP, we set

$$V[0,0] = 0$$
$$V[i,0] = V[i-1,0] + \text{cost}(insertion)$$
$$V[0,j] = V[0,j-1] + \text{cost}(deletion)$$

The DP recurrence is given by

$$V[i,j] = \min\begin{cases} V[i-1,j-1] & \text{if } S[i-1] = T[j-1] \\ V[i-1,j-1] + \text{cost}(mismatch) & \text{if } S[i-1] \neq T[j-1] \\ V[i-1,j] + \text{cost}(insertion) & \\ V[i,j-1] + \text{cost}(deletion) & \end{cases}$$

After computing $V[m,n]$, we obtain an optimal alignment between the learner's and the teacher's pitch lists, and the pitch mistakes in the learner's play can thus be easily identified.

Next, we further align the two note tables using timing information to find mistakes in tempo. From the DP alignment, we obtain a list of notes played at the correct pitch and use the onset information of these notes as "anchors" for temporal alignment. Let $X[1..l]$ and $Y[1..l]$ be the onset lists of the correctly played notes in the learner's and the teacher's note tables, respectively. We try to find the tempo ratio a and starting time shift b so that the sum of square error for onset e is minimized:

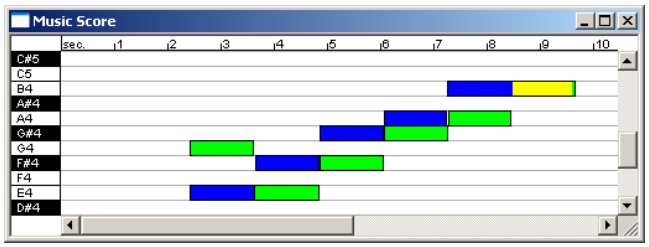
$$e = \sum_{i=1}^{l} (Y[i] - (a \cdot X[i] + b))^2$$

This can be obtained by standard linear regression:

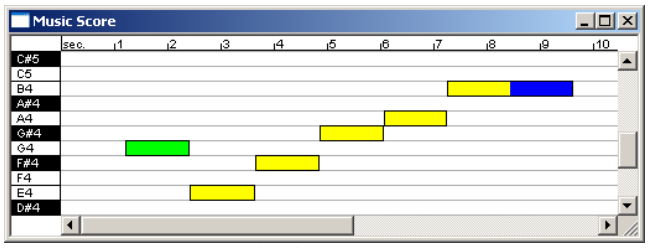
$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{l} X[i]^2 & \sum_{i=1}^{l} X[i] \\ \sum_{i=1}^{l} X[i] & l \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{l} X[i] \cdot Y[i] \\ \sum_{i=1}^{l} Y[i] \end{bmatrix}$$

Finally, we stretch the learner's play by a and shift it by b to produce the final alignment.

The alignment result is visualized in an intuitive way (Figure 10). Every note played is marked by a short bar. The vertical position of the bar indicates the pitch. The horizontal position and the length of the bar indicate the start time and the duration of the note, respectively. The dark color (blue) corresponds to teacher's play. The medium color (green) corresponds to learner's play. The overlapping parts, marked by a light color (yellow), correspond to correct play by the learner. In our user study, we have found that this display format is easily understandable by people who cannot read the usual music notation. Figure 10 shows an example. The teacher plays five notes in the E major scale with the last one doubled in length. The learner plays six notes of same length. The alignment result shows that the learner plays an addition note, the first one and that the last note is wrong in duration.



before the alignment



after the alignment

Figure 10: An example alignment result.

6.2 Summary Evaluation of Learner's Play

After the alignment, we have three aspects of information on the learner's play:

- Pitch error rate, including mismatch, insertion and deletion.
- Tempo ratio b and starting time offset a.
- Onset offset of $E[i] = Y[i] - (a \cdot X[i] + b)$ for each note i.

For the first criterion, we use precision, recall, and F_1 measure to evaluate the note error rate. For the second criterion, starting time offset a is not critical, because it is often due to silence at the beginning of a recording. However, tempo ratio b must be considered, because the learner should not play the piece too slowly or too fast. A Gaussian function is used to score the tempo ratio. The third criterion, notes onset offset, reflects the steadiness of tempo. A Gaussian function of average note square error is used. The overall performance score is defined as a weighted sum of the above three component scores.

7. ANIMATOR

The DVT animator attempts to demonstrate the correct play to the learner in an interactive and entertaining way. Traditionally, in the absence of human teachers, the learner usually relies on video recordings, which offer no feedback and interactivity. The learner must watch the video from a fixed viewing angle, the one that is recorded, and cannot zoom in to watch more closely, *e.g.,* the finger movements. In comparison, the DVT animator is driven by a note table. The learner can choose any part of the table to be played and zoom in to watch more closely different aspects of the play, *e.g.*, fingering or bowing.

We, however, do recognize the limitation of 3D animation. Despite the significant progress in the last decade, fully automatic 3D animation is still not good enough to generate natural-looking human motion. As a result, DVT provides an option to show video recordings, when needed.

7.1 3D Avatar Animator

While playing, a violinist executes intricate motions of fingers, hands, and arms. The left hand shifts among different positions along the neck of the violin; the fingers press the strings and control the pitch of the sound produced. Simultaneously the right arm moves the bow back and forth on the strings to produce the sound. Producing this coordinated motion realistically is a challenging task, due to the complexity of human anatomy. From the shoulder to the tips of fingers, each limb consists of many joints with nearly 30 degrees of freedom (dofs), some of which are interdependent (Figure 11). To play the desired music notes, all the joints have to move in a coordinated fashion to place the fingertips at the intended positions on the strings.

The difficulty of synthesizing human motion has attracted strong interest (see, *e.g.*, the references in [5]). There is also recent work on animating guitar and violin playing. Observing the complexity of motion in playing music instruments, previous work attempts to capture this complexity using sophisticated models such as neural networks [7] and minimization of a global cost function [4].

In contrast, we believe that despite the intricacy required of violin playing, the range of motion executed is fairly restricted. A simple procedural model can capture the motion effectively without much sacrifice in visual realism. The key observation is that to a reasonable approximation, every note has a unique posture of the left hand and arm for producing the note. Given the pitch of a note to be played, we can determine the string for playing the note, the finger, and the fingertip position. This one-to-one relationship is sufficient for beginners and allows us to pre-compute a database of hand postures for all the notes by solving inverse kinematics (IK) and interpolate between the pre-computed postures to synthesize the continuous motion. The bowing motion of the right arm is synthesized in a similar way, but the arm postures are computed on the fly rather than pre-computed and stored in a database. Finally, to render the 3D avatar on the screen, the fingers, hands, and arms are all controlled by a list of joint-angle parameters on a skeletal model. See [14] for details.

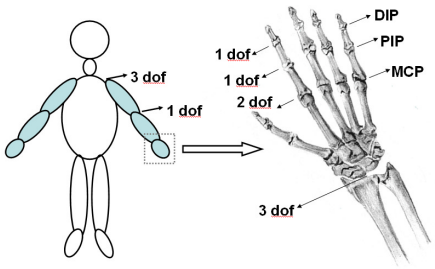


Figure 11: Modeling the dofs in the arm and the hand.

7.2 2D Fingerboard Animator

Although 3D avatar animation is vivid and entertaining, it is sometimes more effective to show an iconography of the fingerboard (Figure 12). At the suggestion of our collaborators from the conservatory, we have incorporated in DVT a module for 2D fingerboard animation, which clearly demonstrates when, which string, and where on the string a finger should press. The animation is again driven by a note table. During the animation, we change the color of the string and highlight the finger position on the string for the note being played. An optional Suzuki band can also be displayed to imitate the fingerboard of violin often used by beginners.

The 2D animation and 3D animation are synchronized by a note table and shown simultaneously (Figure 12).

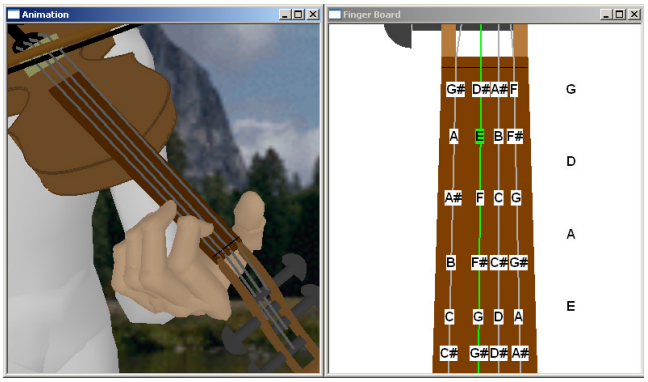


Figure 12: 3D avatar animation (left) and 2D fingerboard animation (right).

8. SYSTEM EVALUATION

DVT is implemented in Microsoft Visual C++, and OpenGL is used for 3D graphics rendering. It runs on any Win32 operating system. In all of our tests, the DVT transcriber achieves real-time performance (Figure 6). The performance evaluator takes very little time compared with the transcriber and returns the result almost instantaneously. The 3D animator generates animation at the rate of 25 frames/second using an ordinary graphics card.

8.1 System Evaluation Method

We conducted a pilot user study and used 12 subjects to evaluate our DVT system.

- Children: we have three child violin beginners. The first one is 5 years old with 7 months learning experience. The second one is 11 years old with 2.5 years experience. The third one is 7 years old with 1 year experience.
- Parents: we have three parents, the first child's mother, the second and third child's fathers.
- Amateurs: we have three amateur violin players. They are about 22 years old with an average 2 years of violin experience.
- Conservatory students: we have two violin students from the Conservatory of Music with 11 and 14 years experience, respectively.
- Teacher: we have one violin teacher with 8 years of teaching experience, who is currently teaching 20-30 students.

The subjects used their own acoustic violin. Since DVT is intended for home use, we used a PC with a Pentium III processor and a cheap microphone in a ordinary room for the study.

To evaluate DVT, we designed a questionnaire (Table 4). Each subject goes through the procedure outlined below, while the parent of the subject (if available) observes the process how his/her child uses the system.

1) We give a brief introduction of the system to the subject.

2) We ask the subject to use the tuner to tune the four strings of the violin.

3) We play the 3D avatar and 2D fingerboard animations of a prerecorded piece of violin music to the subject. We demonstrate the ability to view the 3D avatar from different angles and to zoom and shift 2D fingerboard.

4) We ask the subject to play a piece that he likes. DVT transcribes the play.

5) The subject looks at the raw transcription result (before note alignment), in both note mode and cent mode. We ask the subject if he can find any of his own mistakes.

6) The subject chooses to play one of three pieces, *E-Major Scale*, *Twinkle Twinkle Little Star*, and *Lightly Row*. DVT transcribes the subject's play and compares it with pre-transcribed teacher's play. DVT aligns the transcribed note tables and displays the mistakes in the subject's play.

7) The subject plays the same piece for two or three more times and observes whether the system gives a higher score, if the subjects improves the performance.

Q1	Is it easy to use the tuner to tune the violin? (1-very difficult...5-very easy)
Q2	Do you think 3D avatar is useful and effective? (1-disagree...5-agree)
Q3	Is the 2D finger board helpful to memorize the finger position sequence? (1-useless...5-very helpful)
Q4	Is DVT able to find out the mistakes you make during practice? (1-not at all...5-yes)
Q5	Is the score given by performance evaluator reasonable? (1-no...5-yes)
Q6	Does DVT look fun and cool to play with? (1-boring...5-fascinating)
Q7	Overall, do you think DVT is useful? (1-no...5-yes)
Q8	Among the 4 modules (tuner, score, finger board, avatar) in the system, which is the most important and useful? What's their rank? Why? (explanatory)
Q9	How do you wish to improve DVT? Why? (explanatory)

Table 4: Questionnaire for the user study.

8.2 Evaluation Results

8.2.1 Overall Results

The results of the first seven questions are listed in Table 5:

Subject	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Conservatory 1	5	2	5	4	3	4	5
Conservatory 2	4	4	5	4	3	3	4
Amateur 1	5	3.5	4.5	5	3	3	4
Amateur 2	5	1	4	5	4	3	4
Amateur 3	4	1	2	4	4	4	4
Child 1	2	2	4	5	1	5	3
Parent 1	4	3	3	4	3	5	5
Child 2	5	5	5	5	5	5	5
Parent 2	5	3	5	5	5	4	5
Child 3	5	5	5	5	5	5	5
Parent 3	5	5	5	5	5	5	5
Teacher 1	4	4	4	4	3	4	5
Average	4.42	3.21	4.29	4.58	3.67	4.17	4.50

Table 5: Result of the first seven questions.

The response to Q1 shows that the DVT tuner is effective and easy to use. Both learners and parents who have little music knowledge find it easy and convenient to use the tuner. Our history-based visualization is particularly welcomed. It enabled a five years old child, who does not know how to tune the violin, to successfully tune his violin for the first time by himself. Previously, he relied solely on the teacher to tune the violin once a week, during violin lessons. Therefore, in most of the time during the week between two lessons, the child practices on the untuned violin, which adversely affects the learner's progress. DVT tuner is of great value in these situations.

From Q4, we observe that DVT is capable of identifying the mistakes in learner's play. The intuitive visualization can clearly show the note progression. By comparing the learner's play with the teacher's, the mistakes are visualized clearly. This provides effective feedback to the learner so that he or she can take corrective actions. This feature is particularly welcomed by parents who cannot read musical notation.

Q5 gets an average evaluation of 3.67. In our step 7 of the experiment, we let the subject play the same pieces several times. If his performance improves, the overall score given by the system does increase as well. The numerical score, however, does not seem to be very attractive to learners. Further investigation is needed to reconsider the relevance of the three criteria used for scoring and how they are combined.

The response to Q2 seems to suggest that our 3D avatar animation is not as useful as we expected. Most subjects indicated that the 3D avatar cannot demonstrate the finger positions clearly. It seems that our 3D avatar implementation needs to be improved to achieve the desired realism for effective learning.

The 2D fingerboard animation, despite its simplicity, is quite effective. It highlights the finger position as well as the active string and has received very positive responses from the users. See the responses to Q3.

We also had a case that an amateur learner stopped playing violin for several years. With the help of our fingerboard animation, he managed to pick up a designated piece in 30 minutes.

Q6 and Q7 are overall scores. Q7 shows that the overall evaluation of the DVT is well above the average of the evaluations for the individual components (Q1–Q5). This suggests that what the DVT performs well is the most important ones for users. Our system is perceived to be fun and cool, especially by the younger users (see responses to Q6). These results show that DVT can not only provide effective feedback to improve the learning process, but also stimulate the learner's interest in practicing and potentially increase their average practicing time.

8.2.2 Classified Analysis

Here we try to further analyze the relationships among the different types of learners and their opinions on the system.

Most subjects think that the tuner and the mistake visualization are useful, regardless of their backgrounds. Younger subjects seem to like the avatar animation more, while older ones consider the fingerboard animation is sufficient and more effective. Almost all the users believe that DVT is able to identify and visualize the mistakes that they make in the practice.

DVT is a computer-based system. Those who do not know how to use computers, often demonstrate little interest in DVT. In our study, the little 5 years old child is an example. However, older children who like computers or computer games are particularly interested in DVT.

For conservatory students, their violin performance level is already very high. As a result, they find DVT is not particularly helpful. They also find violin playing performance should be evaluated in more ways than the three criteria in our current system.

The violin teacher and parents are very enthusiastic about DVT. They master the usage of DVT very quickly. They also believe that it is a very useful tool to help their students or children learning to play the violin.

As a whole, we conclude that DVT system is best suited for violin beginners at the age of 10 or above, who know some familiarity with computers. DVT is also good for parents who little knowledge in music or violin, but wish to help their children.

8.2.3 Wish List

We got some valuable feedback from the explanatory question Q9. They are grouped and listed below.

The first group of suggestions is about the graphic display of scores:

1) To reflect the pitch accuracy in the note level score
2) To have a real music score like sheet music

The second group of suggestions is about the avatar:

1) To include more avatar characters (such as spider man) into the system, so that the user can choose his favorite avatar

Finally, we also have another important suggestion from a parent. Most modules of the system provide visual feedback to audio inputs. However, it is also important to provide audio feedback to the learners, to let them listen to the difference between teacher's playing and their playing. For example, after aligning the teacher's score and the learner's one, the system can generate MIDI notes and play them back along with the learner's playing. MIDI notes are very accurate in pitch. By comparing the difference in sound of MIDI notes and the actual notes the learner played, the learner can gradually learn how to hear the difference in pitch, and thus improve the ability to adjust the finger positions just by hearing the sound.

8.2.4 Discussion

In this pilot study, we took an anecdotal approach, and the results indicate the strong potential of DVT as a useful tool for beginning violin learners. However, systematic controlled experiments are needed to further validate these results. Such experiments are in the planning.

8.3 Comparison with Other Systems

Now we compare DVT with related and recent music educational systems, Family Ensemble [10] and pianoFORTE [13].

Family Ensemble is a music edutainment system, with which a parent and his/her child can play ensemble on a MIDI keyboard. Using a music database and score tracking techniques, wrong notes played by the parent can be substituted to the correct ones automatically by the system. Thus it enables parents with little or no experience in playing the piano to accompany the child during the practice.

PianoFORTE is a music education system which provides visual feedback of student's mistakes. Student's performance on a MIDI keyboard is recorded. Notes, tempo, dynamics and articulation are displayed on the computer screen with graphic notations, which enables the student and the teacher to review the performance and point out mistakes.

A detailed comparison of our system with the other two systems is listed in Table 6.

Item	DVT	Family Ensemble	PianoFORTE
Input	Acoustic Signal	MIDI	MIDI
Device Needed	Cheap Microphone	MIDI Keyboard	MIDI Keyboard
Score Database	Optional	Required	Not Required
Tuner	Yes	No	No
Score Visualization	Yes	No	Yes
Animation	Yes	No	No
Automatic Error Identification	Yes	Yes	No
Feedback	Multimodal Feedback	N/A	Limited

Table 6: System comparison.

Table 6 shows that the input of DVT is acoustic signal, not MIDI as in the other two systems. This makes our system advantageous due to the fact that it can be extended to other acoustic instruments, especially the strings, therefore is not restricted to keyboard instruments.

Family Ensemble requires a score database in order to track and correct the parent's playing. These scores must be pre-stored into the database to enable the system to work on these specific pieces. DVT does not require pre-stored database. The teacher and the learner can just record and transcribe to get an analyzable score. However optionally, we can include some teaching materials with standard scores and ship them together with DVT.

In comparison to the other two systems, a distinct feature of DVT is its multimodal feedback to learners.

The feedback of DVT is mainly visual, which is similar to PianoFORTE. We believe adding audio feedback is a useful approach as used in Family Ensemble.

9. CONCLUSION AND FUTURE WORK

We have presented the initial design of and experience with Digital Violin Tutor, an integrated system for helping beginning violin learners. DVT combines fast and accurate violin audio transcription with intuitive visualization to provide much-needed feedback to learners, when human teachers are not available. The

entire system has been implemented with off-the-shelf hardware and shown to be practical in home environments. Although we have chosen violin as the instrument for our system, we believe that the main design considerations generalize to other related string instruments.

DVT can be improved in many ways. Based on our user study, we have identified two most important ones as the immediate next step. First, incorrect bowing is a common mistake among beginners. It does not lead to wrong pitch or timing, which our current transcription method focuses on. To find bowing mistakes, we plan to extend our method to detect the difference in the *timbre* of the sound. Second, we plan to perform polyphonic transcription in order to make the system useful for more advanced learners. Another important issue is to deal with transcription or note alignment errors, which may lead to incorrect feedback to the learner. We believe that it is possible to build statistical models of transcription and note alignment errors. DVT can then use such models to provide the learner a confidence estimate of the performance evaluation.

10. REFERENCES

[1] D. Bendor, M. Sandler. Time Domain Extraction of Vibrato from Monophonic Instruments. *International Symposium on Music Information Retrieval*, 2000.

[2] R. B. Dannenberg. An online-line algorithm for real-time accompaniment. *Proc. ICMC 1984, ICMA (1984),* 193-198.

[3] R. B. Dannenberg, M. Sanchez, A. Joseph, P Capell, R. Joseph, R. Saul. A Computer-Based Multi-Media Tutor for Beginning Piano Students. *Journal of New Music Research,* 19(2-3), 1990.

[4] G. ElKoura, K. Singh. Handrix: Animating the Human Hand, *ACM Symp. Computer Animation,* pp. 110-119, 2003.

[5] P. Faloutsos, M. van de Panne, D. Terzopoulos. Composable Controllers for Physics-based Character Animation. *SIGGRAPH Conference Proceedings*, pp. 251-260, 2001.

[6] B. Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11): 1477-1493, 1986.

[7] J. Kim, F. Cordier, N.M. Thalmann. Neural Networks Based Violinist's Hand Animation. *Computer Graphics International*, pp. 37-41, 2000.

[8] A. P. Klapuri. Automatic Transcription of Music. *Proceedings of the Stockholm Music Acoustic Conference*. 2003.

[9] P. M. Martins, J. S. Ferreira. PCM to MIDI Transposition. *Audio Engineering Society 112th Convention Paper*, May 2002.

[10] C. Oshima, K. Nishimoto, M. Suzuki. Family Ensemble: A Collaborative Musical Edutainment System for Children and Parents. *ACM Multimedia 04*, 2004.

[11] M. Piszczalski, B.A. Galler. Automatic Music Transcription. *Computer Music Journal*, 1(4):24-31, 1977.

[12] S. Rossignol, X. Rodet, J. Soumagne, J. L. Collete, and P. Depalle. Automatic characterization of musical signals: feature extraction and temporal segmentation. *Journal of New Music Research*, 1999.

[13] S. W. Smoliar, J. A. Waterworth, P. R. Kellock. pianoFORTE: A System for Piano Education Beyond Notation Literary. *ACM Multimedia 95*, 1995.

[14] J. Yin, A. Dhanik, D. Hsu, Y. Wang. The Creation of a Music-Driven Digital Violinist. *ACM Multimedia 04*, 2004.

[15] J. Yin, T. Sim, Y. Wang, A Shenoy. Music Transcription Using an Instrument Model. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[16] Amazing MIDI, http://www.pluto.dti.ne.jp/~araki/amazingmidi/index.html

[17] intelliScore, http://www.intelliscore.ne