HIERARCHICAL APPROACH TO DETECT COMMON MISTAKES OF BEGINNER FLUTE PLAYERS

Yoonchang Han, Kyogu Lee
Music and Audio Research Group
Seoul National University, Seoul, Republic of Korea
{yoonchanghan,kglee}@snu.ac.kr

ABSTRACT

Music lessons are a repetitive process of giving feedback on a student's performance techniques. The manner in which performance skills are improved depends on the particular instrument, and therefore, it is important to consider the unique characteristics of the target instrument. In this paper, we investigate the common mistakes of beginner flute players and propose a hierarchical approach to detect such mistakes. We first examine the structure and mechanism of the flute, and define several types of common mistakes that can be caused by incorrect assembly, poor blowing skills, or mis-fingering. We propose tailored algorithms for detecting each case by combining deterministic signal processing and deep learning, to quantify the quality of a flute sound. The system is structured hierarchically, as mis-fingering detection requires the input sound to be correctly assembled and blown to discriminate minor sound difference. Experimental results show that it is possible to identify different mistakes in flute performance using our proposed algorithms.

1. INTRODUCTION

The most important part of a music lesson is giving a student feedback on his or her performance, posture, and playing skills so that the student can play the sound correctly. Music lesson methods vary depending on the instrument being learned; therefore, audio signal processing for music education should make extensive use of prior knowledge regarding playing style, common mistakes, unique characteristics, and constraints of the target instrument. However, most existing music signal analysis techniques use a general-purpose model, and relatively little attention is paid to an instrument-specific approach. A general-purpose model is advantageous because it can be applied to various types of instruments. However, this model lacks the capability to capture instrument-specific sound characteristics. There are always common mistakes that beginners make, but little is known about how to detect these automatically.

The goal of this paper is to investigate common beginner's mistakes when playing a specific instrument—the flute, in this case—and to analyze the spectral characteristic of each case to give the student appropriate feedback on his or her performance. Because the sound of a musical instrument is affected by numerous factors, in our work, we first divide the factors that usually lead beginners to play the wrong sound into three parts: incorrect flute assembly, blowing skill, and fingering.

The rest of the paper is organized as follows: We briefly present existing works related to our proposed idea. Then, we investigate possible mistakes in flute performance by examining the structure and mechanism of the flute, and several types of common mistakes and the resulting sounds are explained. Next, we present an overall system structure to distinguish each mistake, along with a detail explanation of each proposed algorithm. We then present the experimental results to demonstrate the feasibility of the proposed system, followed by our conclusion and directions for future work.

2. RELATED WORK

The characteristics of musical instruments depend on their sound production mechanism. The characteristics of one instrument can greatly differ from those of others, and each instrument's characteristics may not be captured equally well as another even when using the same computational model [2]. However, there has been minimal research regarding an instrument-specific model. Some examples of instrument-specific approaches involve the use of a violin [8, 14-16], guitar [1], bells [9], and tabla [5]. For instance, the violin transcription system in [8] makes use of characteristics such as highest and lowest pitch, possible play style (e.g., upper octave duophony), vibrato, and loudness. The training system in [14] uses a common envelope style of violin sound for note segmentation prior to real-time pitch detection, and [9] uses the acoustic characteristics of a church bell, as well as the rules of a bell charming performance, for transcription and estimating the number of bells. In addition, a chord transcription system designed for guitar in [1] outperforms the non-guitar-specific method.

As shown above, using prior knowledge of the characteristics of a target instrument creates new possibilities in music signal processing, and can also improve the per-

formance of the system. However, there are still many instruments to be studied, and the flute is one of them.

3. COMMON MISTAKES OF A FLUTE PLAYER

3.1 Assembling the Flute

Like most woodwind instruments, the flute needs to be assembled before it is played. The flute consists of a head joint, body joint, and foot joint, as shown in Figure 1. The connecting part between the body and foot joint is very short, while the connecting part between the head joint and body joint is a few centimeters long. This intentionally designed adjustable part is called the tuning slide, and it can be used for changing the total length of the flute to various sizes, which affects the overall pitch of the flute. For instance, if the head joint is placed very deep into the tuning slide of the body, the pitch will be increased for every note. By contrast, if the head joint is pulled out too far, the overall pitch will drop owing to the longer wavelength.

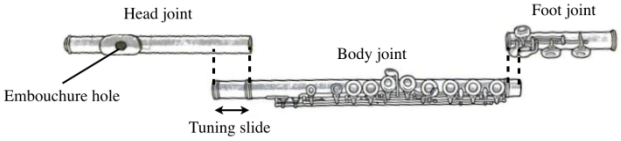

Figure 1. Flute consists of head joint, body joint, and foot joint (modified after [11]).

Another method of pitch tuning is adjusting the cork part of the head joint, as shown in Figure 2. This can be adjusted by a screw. Pushing the cork will raise the pitch of all notes. However, this is beyond the scope of this paper, as this screw is normally not adjusted by flute performers but by flute technicians.

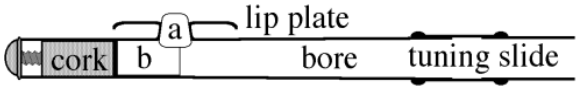

Figure 2. Schematic of a flute head joint [13].

Trained performers use this variable tuning slide for pitch tuning. The pitch of the flute is sensitive to the conditions of the surrounding environment, such as humidity and temperature. However, novice flutists are not sensitive to minor pitch shifting, and they may play the flute in the wrong overall pitch without recognizing it.

3.2 Blowing Embouchure

The flute generates sound by blowing a rapid air jet across the embouchure[1] hole, as shown in Figure 3. Hence, the quality of the generated sound is highly dependent on the blowing skill of the performer. Blowing skill involves lip position and the thickness/stability of the air jet. Clear tone production is challenging for beginners because the method of tone production for the flute is not supported by mechanical parts; rather, it depends only on the player's blowing skill [4].

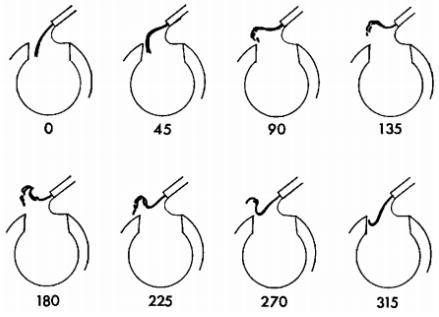

Figure 3. Airstream oscillation of the flute embouchure hole. The labels indicate the phase angles of the acoustic current at the hole [3].

Tone quality and octave of the sound are related to blowing skill. The flute has a range of three octaves, starting from middle C (C4), with several less-used notes in octaves 3 and 7. The blowing pressure determines the octave of the sound, as shown in Figure 4. Greater blowing pressure can be achieved by blowing a narrower and stronger air jet. To generate a stable and clean sound, it is important to keep this blowing pressure reasonably steady. Failure to do this will result in fluctuating sound and noise, which is highly unpleasant and typically the first hurdle for beginners to overcome in their training.

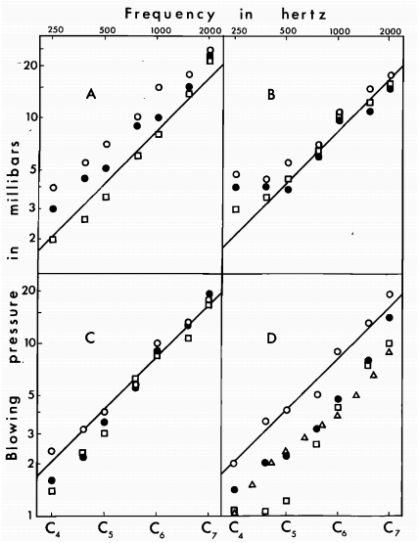

Figure 4. Air jet blowing pressure has a roughly linear relationship to fundamental frequency. A, B, C, and D are different performers, and different shapes represent different dynamics [12].

3.3 Fingering

Novice flutists frequently make mistakes in fingering owing to their lack of familiarity with the irregular fingering rules of the flute. High-octave fingering is comparatively more complex than low-octave fingering [4], which is the reason why flute lessons usually start with the lowest octave and move step-by-step to higher octaves. Hence, we

[1] Mouthpiece of a musical instrument.

focus on octaves 4 and 5, which are the octaves that beginner flute players initially study.

Most of the octave 5 fingerings are identical to those of octave 4, as shown in Figure 3. However, the fingering for C and D, as well as the sharps of these notes, require different fingerings than those of octave 4. These notes can be played with octave 4 fingering using a faster and sharper air jet, but this results in a slightly airy timbre, compared to the sound when the flute is correctly fingered. As this airy timbre is not significantly noticeable, and most of the notes in octaves 4 and 5 share the same fingering, many beginners do not notice that they used octave 4 fingerings to play octave 5, unless the instructor spots it.

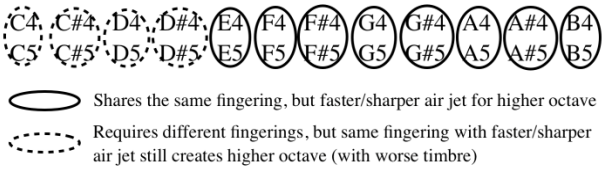

Figure 5. Fingering of octave 4 and 5 flute notes. Note that C, D, and sharps of these require different fingering, unlike E, F, G, A, and B.

Another fingering-related problem is the proper positioning of the fingers. The open-hole flute requires that the flutist use his or her fingers to block the holes in the keys. Most professional flutists prefer the open-hole flute owing to its advantages in tone production and intonation adjustment [4]. However, this is not considered in our system because beginners who have trouble with blocking open-hole keys can avoid this problem by putting plastic plugs in the holes until they get used to playing the open-hole flute.

4. PROPOSED SYSTEM & METRICS

The overall system comprises several steps. In the first step, the system determines whether the flute is assembled correctly using entire input audio. Next, once the flute sound is detected as coming from a correctly assembled flute, the system measures if sound of the each note is a clear, correctly blown sound or an airy-timbered sound. Finally, the properly blown sound is identified as sound generated from either correct fingering or incorrect fingering. The system is hierarchically structured, because mis-fingering detection does not work well for fluctuated sound or head joint pushed/pulled sound as it requires discriminating minor sound difference. The input audio is resampled to 16 kHz first, and he system architecture is shown in Figure 6.

4.1 Assembling Error Detection

Some mistakes can cause modifications to the overall pitch, and some mistakes result in poor timbre. The assembling error affects only the overall pitch of the generated sound. As mentioned in 3.1, the distance the head joint is pushed in or pulled out from the tuning slide of the body joint determines the overall pitch. To this end, a quantized chromagram from Harte and Sandler is used to detect the tuning center [6].

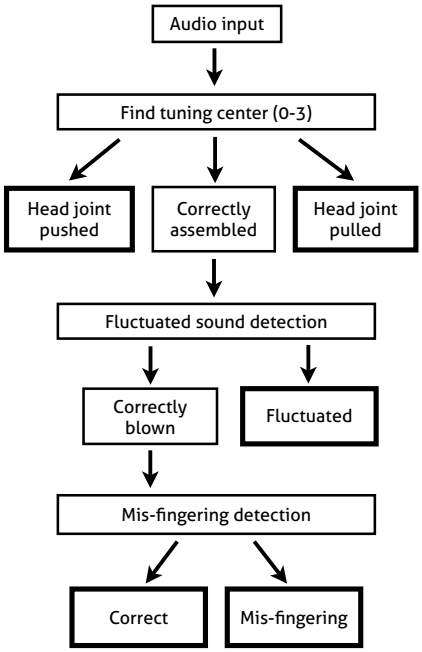

Figure 6. Flow diagram of the overall system. The bold box indicates where the system sends feedback to the user.

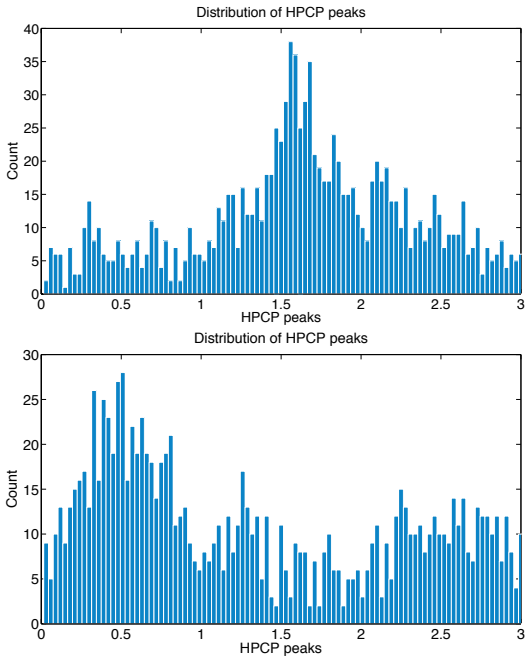

Figure 7. HPCP peaks histogram within a semitone for correctly (up) and loosely assembled (down) case.

To determine the tuning center, a spectrum of linear frequency spectra is Constant-Q transformed and summed across octaves to produce a harmonic pitch class profile (HPCP). A 36-bin quantized chromagram is used to determine the semitone center, and three bins were al-

located for each semitone. By observing the distribution of peak positions across the width of a semitone, as shown in Figure 7, it is possible to determine the tuning center of the instrument. Because three bins are allocated for each semitone, the tuning center of a perfectly tuned sound would ideally be 1.5. Therefore, the system will consider the input sound to be correctly tuned when the tuning center value is approximately 1.5. If the detected tuning center is too low (less than 1), the system sends feedback to the user that the head joint is too loosely assembled. Conversely, the system tells the user that the head joint is assembled more tightly than necessary when the tuning center is high (greater than 1).

4.2 Fluctuated Sound Detection

Incorrect lip position on the embouchure, along with an irregular stream of blown wind, results in a highly unpleasant and fluctuating tone. This sound contains many inharmonic partials in a spectrum, and it is clearly visible on a spectrogram. Performing binary masking on a spectrogram makes these inharmonic partials more obvious, as shown in the second row of Figure 8.

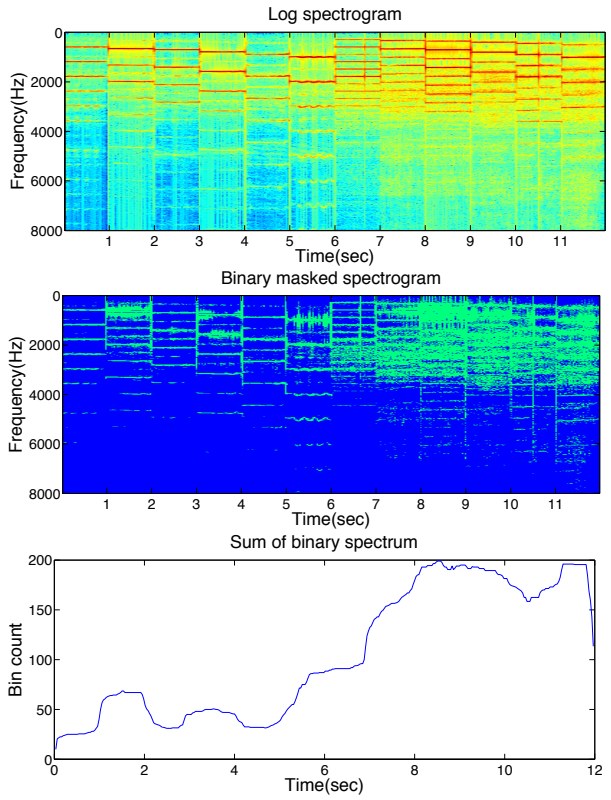

Figure 8. Log spectrogram, binary masked spectrogram, and sum of bins for each frame for D, E, F, G, A, and B of octave 5. Up to 6 second is correctly blown sound and from 6 to 12 second is fluctuated sound.

Binary masking is performed as follows:

$$X_b(k) = \begin{cases} 0 & X(k) < \theta \\ 1 & X(k) > \theta \end{cases} \quad (1)$$

where X is the log spectrum, X_b is the binary masked spectrum, and θ is the threshold constant. Empirically, a value between -20 and -30 works well for θ, depends on recording environment. Note that these values are obtained when natural log multiplied by 20 is used for the log spectrum. Using this binary masked spectrogram, the sum of the number of positive valued bins of each spectrum can be used as a measurement for determining how the sound fluctuates owing to poor blowing skill. This can be expressed as follows:

$$F(k) = \sum_{k=0}^{N-1} X_b(k) \quad (2)$$

where F is the amount of fluctuation. The third row of Figure 8 is F value obtained from (2) with 1 second median filtering, and it is possible to observe the value is much higher for fluctuated sound than correctly blown sound.

4.3 Mis-fingering detection

As mentioned in 3.3, for C5, C#5, D5, and D#5, using octave 4 fingering with a faster and sharper air jet still generates octave 5 pitches even without correct fingering, although the timbre is slightly airy. To detect this timbral difference, we decided to use both the Mel-frequency cepstral coefficient (MFCC)—a widely used, hand-designed feature—and sparse filtering (SF) [10]—a deep-layered, unsupervised feature learning method. SF works by optimizing the sparsity of feature distribution, and it works well on a range of data modalities without specific tuning. Both single- and double-layered sparse filtering were used with 200 units for each layer. The obtained feature was classified into two classes (correct/incorrect) using a random forest (RF) classifier, which exhibits better performance than a support vector machine or back-propagation neural network in a variety of cases [7]. The flow diagram for mis-fingering detection is shown below.

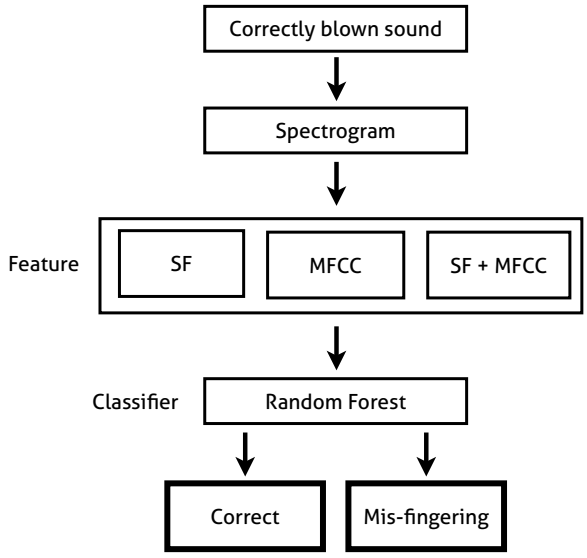

Figure 9. Flow diagram for mis-fingering detection.

5. EXPERIMENT

5.1 Objective & Procedure

The goal of our experiment was to explore whether the proposed system and algorithms work well for detecting the mistakes of beginner flutists. Flute sound samples were obtained from two intermediates (who have played the flute for one to two years) and one expert (who holds an exam score of Grade 8 with a Distinction). Flutes used for the experiment were a B foot joint with open holes, and a silver head with nickel body and foot. The correct flute sound, fluctuating sound, head joint pulled, and head joint pushed sound were recorded for octaves between 4 and 5. The length of the collected audio was 30 seconds for each semitone. The case of correct and incorrect fingering for C5, C#5, D5, and D#5 was recorded for 10 minutes each to obtain sufficient training data. The input audio was recorded at 44.1 kHz mono and downsampled to 18 kHz. Tuning center was calculated from whole target audio as it is not time-varying characteristics. Meanwhile, fluctuating and mis-fingering detection was performed framewise. Different window and hop size were used for each experiment, as each mistake detection algorithm requires different spectral resolution.

5.2 Results

The experimental results show that the system successfully distinguishes each mistake. To find tuning center, a 74 ms window and 18 ms hop size were used. As shown in Table 1, the tuning center of a correctly played sample is close to 1.5, which is the exact center. Also, tuning center values for the head joint when it is pushed and pulled fell into the expected range, which were (0–1) and (2–3), respectively.

Mistake cases	Tuning center value (0 to 3)		
	Player 1	Player 2	Player 3
Correct	1.68	1.59	1.62
Head joint pushed	2.55	2.49	2.43
Head joint pulled	0.48	0.45	0.75

Table 1. Tuning center values of correct, head joint pushed, and head joint pulled flute sound of three different flutists.

Next, Figure 10 is a framewise distribution of fluctuation measure (1) for correct and fluctuated flute sound. A 64 ms window and 32 ms hop size were used, with θ value of -25dB. The median value of the correct flute sound is 50, and most of the values fall between 63 and 37. The fluctuating flute sound has a median value of 167, and most of the values fall between 150 and 178. This means that these cases are clearly distinguishable using the proposed metric.

Finally, Table 2 shows the ten-fold cross-validation results of the proposed mis-fingering classification using single-layer SF, double-layer SF, and MFCC as a feature, and RF as a classifier. A 16 ms window and 10 ms hop size were used, and SF was used with 200 units per layer.

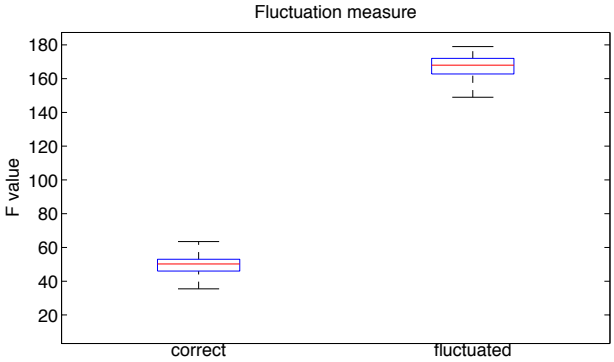

Figure 10. Box plot of fluctuation measurements. The central marks indicate the median, and the edges are the first and third quartiles.

Method	Accuracy (%)
Spectrogram + SF (single)	90.24
Spectrogram + SF (double)	90.02
MFCC	90.89
MFCC + SF (single)	90.33
MFCC + SF (double)	91.35

Table 2. Mis-fingering classification ten-fold cross-validation result using SF/ MFCC as a feature and RF as a classifier.

The result shows that the combination of the MFCC and double-layered SF performs the best; however, all of the approaches perform reasonably well within a not very meaningful margin. The result indicates that the MFCC, a handcrafted feature, is still useful in separating the timbral differences of the flute. Further, although SF is not designed for the purpose of timbre analysis, it works quite well without fine-tuning, as mentioned in [10]. In the experiment, single-layered SF worked better when the input is a spectrogram, but double-layered SF showed better performance when the input is MFCC.

6. CONCLUSION & FUTURE WORK

The objective of our work is to use audio signal analysis to give a student feedback on his or her flute performance to help fix mistakes, as a lesson teacher would do. To achieve this goal, we examined the mechanism and structure of the flute. We also investigated the common mistakes of beginner flute players. We determined several types of common mistakes and developed a hierarchical system to detect such cases by observing the tuning cen-

ter, fluctuation metric, and a mis-fingering detection algorithm. As a result, we have successfully identified common mistake cases from input audio, which can be used as feedback that would be provided by a lesson teacher-. Head-joint assembling errors were detected by determining the tuning center of the flute sound. Fluctuating sound caused by poor blowing skills was separated from the correct flute sound by measuring the amount of noisy harmonic contents. Finally, mis-fingering cases were detected by analyzing their timbre using MFCC and SF with an RF classifier.

There remain some problems to be tackled in this mistake detection algorithm for real-world user applications. First, the mis-fingering detection algorithm may be affected by the material or maker of the flute because the algorithm detects very minor changes in timbre. In the experiment, only two types of flute (silver head with nickel body, and foot) were used. However, the flute can be made of various types of metal, such as silver, gold, and platinum. Moreover, various flute makers have their own timbral characteristics, which may influence the classification results. Second, the experiment was done on the frame level, but the user perceives the score based on the note level. Hence, the system should be used along with appropriate onset-offset detection to give more user-friendly feedback.

We believe that this type of timbre-related and user-behavior-oriented feedback is highly important for the next-generation music transcription systems, especially those used for educational purposes. Playing the instrument with correct onset and pitch is not a very difficult part of being a good player, but making a beautiful timbre is what really takes time. This paper focuses only on the flute; however, our overall approach, including analyzing mistake cases and determining customized solutions, can be applied to various instruments in a similar way.

7. ACKNOWLEDGEMENTS

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency" (NIPA-2013-H0301-13-4005).

8. REFERENCES

[1] A. M. Barbancho, A. Klapuri, L. J. Tardon, and I. Barbancho: "Automatic Transcription of Guitar Chords and Fingering from Audio," IEEE TASLP, 20(3): 915–921, 2012.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchoff, and A. Klapuri: "Automatic Music Transcription: Breaking the Glass Ceiling," In *ISMIR*, 2012.

[3] J. W. Coltman: "Sounding Mechanism of the Flute and Organ Pipe." *The Journal of the Acoustical Society of America* 44, 1968.

[4] C. Delaney, *Teacher's Guide for the Flute*. Rev. 11/98, Selmer, 1969.

[5] O. Gillet and G. Richard: "Automatic Labelling of Tabla Signals," In *ISMIR*, 2003.

[6] C. Harte and M. Sandler: "Automatic Chord Identification using a Quantised Chromagram." In *Audio Engineering Society Convention 118*. 2005.

[7] M. Liu, M. Wang, J. Wang, and D. Li: "Comparison of Random Forest, Support Vector Machine and Back Propagation Neural Network for Electronic Tongue Data Classification: Application to the Recognition of Orange Beverage and Chinese Vinegar," Elsevier Sensors and Actuators, pp. 970–980, Vol. 177, 2013.

[8] A. Loscos, Y. Wang, and W. J. Boo: "Low Level Descriptors for Automatic Violin Transcription." In *ISMIR*, 2006.

[9] M. Marolt: "Automatic Transcription of Bell Chiming Recordings." In *IEEE TASLP*, 20(3): pp. 844–853, 2012.

[10] J. Ngiam, P. W. Koh, Z. Chen, S. Bhaskar and A. Y. Ng., "Sparse Filtering," In NIPS, 2011.

[11] H. Pinksterboer, *Tipbook Flute and Piccolo: The Complete Guide*, Hal Leonard, 2009.

[12] T. D. Rossing, F. Richard Moore, and P. A. Wheeler: *The Science of Sound*. Vol. 2. Massachusetts. Addison-Wesley, 1990.

[13] J. Smith, J. Wolfe, and M. Green: "Head Joint, Embouchure Hole and Filtering Effects on the Input Impedance of Flutes." In *Proc. of the Stockholm Music Acoustics Conference*, pp. 295–298. 2003.

[14] J. Wang, S. Wang, W. Chen, K. Chang, and H. Chen: "Real-Time Pitch Training System for Violin Learners," *Multimedia and Expo Workshops (ICMEW), IEEE*, 2012.

[15] R. S. Wilson: "First Steps Towards Violin Performance Extraction using Genetic Programming," In John R. Koza, editor, *Genetic Algorithms and Genetic programming*, pp. 253–262, 2002.

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