DETECTION OF COMMON MISTAKES IN NOVICE VIOLIN PLAYING

Yin-Jyun Luo[1,2] **Li Su[2]** **Yi-Hsuan Yang[2]** **Tai-Shih Chi[1]**

[1]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
fredom.smt02g@nctu.edu.tw, lisu@citi.sinica.edu.tw,
yang@citi.sinica.edu.tw, tschi@mail.nctu.edu.tw

ABSTRACT

Analyzing and modeling playing mistakes are essential parts of computer-aided education tools in learning musical instruments. In this paper, we present a system for identifying four types of mistakes commonly made by novice violin players. We construct a new dataset comprising of 981 legato notes played by 10 players across different skill levels, and have violin experts annotate all possible mistakes associated with each note by listening to the recordings. Five feature representations are generated from the same feature set with different scales, including two note-level representations and three segment-level representations of the onset, sustain and offset, and are tested for automatically identifying playing mistakes. Performance is evaluated under the framework of using the Fisher score for feature selection and the support vector machine for classification. Results show that the F-measures using different feature representations can vary up to 20% for two types of playing mistakes. It demonstrates the different sensitivities of each feature representation to different mistakes. Moreover, our results suggest that the standard audio features such as MFCCs are not good enough and more advanced feature design may be needed.

1. INTRODUCTION

With advances in music technology, the development of computer-aided music learning and automatic scoring systems has attracted wide attention. Such systems provide self-learning experiences to users through computer-aided platforms. Despite numerous efforts have been made, however, the performance of current systems still leaves plenty of space for improvement. A review of the music learning system and the main challenge can be found in [1].

For a novice player, three common basic aspects, intonation, rhythm and timbre, are often used to evaluate his/her performance [2]. Intonation refers to the pitch of the tone, rhythm specifies the duration of the tone, and timbre characterizes the overall quality of the tone. Conventionally, a novice player uses a tuner for correcting the intonation and a metronome for following the rhythm during the practice. In traditional music education, there is no hardware device capable of automatically evaluating the timbre quality.

Up to date, most of the computer-aided music learning systems also focus on intonation and rhythm only [1]. These studies mainly coped with learning intonation and rhythm in music in the context of automatic music transcription (AMT). For example, the pitch played by the violin learner was automatically detected and visually presented to evaluate the pitch intonation [3]. A fusion of audio and video cues improved the onset detection of non-percussive instruments, such as violin, and thereby enhanced the performance of AMT [4]. Automatic singing quality assessment is achieved by measuring the dissimilarity between singing voices of beginners and of trained singers [5]. Besides intonation and rhythm, timbre plays an essential role in identifying the skill (or proficiency) level of a player but has not attracted much attention in computer-aided music learning platforms. Some timbre-related research studies considered instrumental expression to recognize the techniques in playing musical notes by violin [6] and by electric bass guitar [7], respectively. Other studies aimed to evaluate the played notes, for example, using spectral parameters from long tones to evaluate the technical level of saxophone players [8]. Recently, a hierarchical approach combining deterministic signal processing and deep learning was employed to identify different common mistakes made by novice flute players [9]. Machine learning techniques were also adopted to distinguish good trumpet tones from bad ones [10]. The first attempt to detect bad violin playing in [11] is the most relevant work to the proposed study. One of the two tasks conducted in [11] classifies violin tones into binary clusters, i.e., good or bad, using k-nearest neighborhood algorithm. The other task examined the prominent feature sets for detecting individual playing mistakes. Similarly, in this paper, we explore the capability of timbre in detecting playing mistakes produced by novice violin players during practice. However, since the dataset and the algorithm codes in [11] are not publicly available, it is difficult to compare our approach with the approach in [11].

The first contribution of this paper is to build and release a new dataset[1] for such a research problem. To be as realistic as possible, we recorded four successive legato notes, which require smooth, round and continuous flow of tones [12], as a unit and then trimmed it into individual notes rather than simply recording single note at a time as in [11]. The resulting dataset comprises of 981 individual legato notes played by several players across different skill levels. The playing mistakes associated with each note were annotated by violin experts using the following four pre-defined classes: *scratching bow*, *crooked bow*, *bouncing bow*, and *inappropriate arm height*. More details of the annotations and dataset are elaborated in Sections 2 and 3, respectively.

The second contribution of this paper is to evaluate a number of features capturing the acoustic characteristics of different segments of a musical note for the task of automatic playing mistake classification. A set of spectral features is extracted from either the whole note or the segment of onset, sustain and offset, partitioned using the output of an optical sensor installed on the violin. The approach leads to five different feature representations: *Note, Onset*, *Sustain*, *Offset* and *Cascade*. They refer to features extracted from the whole note, from the corresponding segments only and the concatenation of segment-level features, respectively. More details about the partitioning method and feature extraction can be found in Sections 3 and 4, respectively.

In our approach, the Fisher score is used for feature selection and the support vector machine (SVM) is used for classification. Feature selection is done as a preprocessing step of classification. The performance of classification is assessed in terms of the F-measure. Experimental details are presented in Section 5. Exploration of insights to link specific feature representations to playing mistakes is presented in Section 6, before we conclude the paper in Section 7.

2. VIOLIN PLAYING MISTAKES

We defined four common playing mistakes made by violin novices. These mistakes are mainly related to the bow arm and the bow hand which dominantly control violin timbre for novice players and cause most of the trouble for violinists [2].

2.1 Scratching Bow (SB)

The pressure of the bow applied on the string can either come from the weight of the bow, arm and hand, from controlled muscular action, or from a combination of these factors [2]. Excessive bowing pressure without enough bowing speed to complement with can hinder the vibrations of the string and produce coarse sound with inferior quality. Without the support of bowing speed, extreme pressure of the bow on the string results in sound with scratching effect.

2.2 Crooked Bow (CB)

Drawing a straight bow from the frog to the tip is the foundation of the bowing technique [2]. If the bow is crooked, not parallel to the bridge, the sound quality will vary due to change of the contact position of the bow on the string. Severe inclination even causes sudden displacement of the bow from the bridge and produces sound with skating effect.

2.3 Bouncing Bow (BB)

Lack of muscular control of either the bow arm or the bow grip reduces strength to the bow. It might prevent the bow from properly laying on the string, thereby the bow bounces naturally due to its elasticity.

2.4 Inappropriate Arm Height (IAH)

Appropriate tilt of the arm relative to the bow is required in order to play on each string without touching the other strings. With inconsistent height or tilt of the arm when drawing the bow across the string, pitch produced by adjacent string might be heard.

3. DATASET

All notes in the dataset were played by ten players across different skill levels using the same violin in a semi-anechoic chamber. Four players are relatively more experienced in violin or similar string instruments such as cello, while the other six players have learned to play violin for less than one month. Each player was asked to play four successive notes as a clip at the speed of 60 beat-per-minute (BPM). Each clip was directed to start with down-bow and end with up-bow. In total, 26 clips containing 104 legato notes were played by each player. This style of successive playing is more similar to actual practicing than the style of playing an individual note at once. In our recordings, analysis of transition between notes is also feasible though we leave it as future work. We limit the study to consider legato notes only because legato is the essence of all cantabile playing [12] and one can hardly master other advanced techniques before playing it well.

Segmentation between notes and within each note was achieved using a photo resistor and four rings of surface-mounted light-emitting diodes (SMD LEDs) installed respectively underneath the violin bridge and on the bow stick. Two of the four rings were installed at the positions close to the frog and the tip on the bow stick, while the other two were placed at both ends of the middle of the bow. Segmenting a violin note can benefit the analysis, as the time domain signal varies in characteristic over a bow draw. The purpose of installing the optical sensor was to segment the time domain signal in a more direct way rather than the approach in [13]. When a legato note was played, the optical sensor was capable of marking the time instants, at which those ring-located

[1] The audio clips and annotations of playing mistakes can be found in http://perception.cm.nctu.edu.tw/sound-demo/.

positions of the bow stick passed through the violin bridge, without influencing playing. As our main purpose was to simply divide the bow draw into three segments, we can tolerate the small accuracy errors of the sensors on the longitudinal bow position [13].

Based on the marked time instants, we divided each clip into four individual notes and segmented each note into three different segments, i.e., onset, sustain and offset, as playing mistakes can occur at any instant of the drawing. The two ends of each clip, the start of the first note and the end of the fourth note, were manually determined by an energy threshold. The edges between successive notes and within each individual note were automatically defined by the marked time instants. At the end, we collected 981 notes in total and the corresponding segments after discarding notes containing accidently made distinct noise during the recording.

We employed the hardware-assisted approach instead of automatic approaches proposed in the literature [14, 15] because automatic approaches usually segment an individual musical note according to the temporal evolution of amplitude envelope and spectral centroid [14, 15]. Since we are dealing with notes produced by the violin novices, the algorithms developed for well-played musical notes are not applicable in our case. For instance, simply dividing the note into three equal-duration segments would not produce same results as our hardware-assisted approach since novice players cannot draw the bow with a constant speed. Therefore, without the assistance of the optical sensors, automatic segmentation of violin notes performed by novice players should be a difficult task, which is beyond the scope of this paper and deserves further research in the future.

The notes were then annotated by violin experts using the four pre-defined mistakes. Note that a single note could possess multiple playing mistakes. Fig. 1 shows the duration distributions of notes and the corresponding segments. One can observe that although players were asked to play each note at the speed of 60 BPM, beginners, especially those who lack of musical background, weren't necessarily able to perform accurately. Table 1 summarizes the numbers of instances of the playing mistakes in the first row. Dividing the first row by the total number of the collected notes gives the percentages in the second row.

4. METHOD

4.1 Preprocessing

All of the notes in the dataset were resampled to 44.1 kHz and saved in the mono-channel WAV format. Before feature extraction, each time domain signal was first normalized to zero mean and unit variance and then divided into three segments as described in Section 3 for further analysis.

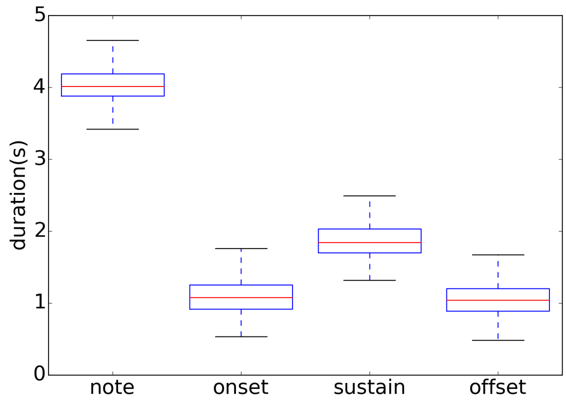

Figure 1. Duration distributions of 981 notes (the first column), the corresponding onset, sustain and offset segments (the second to the last column).

	SB	CB	BB	IAH
Numbers	265	133	154	53
Percentage	27.0%	13.6%	15.7%	5.4%

Table 2. The number of instances of each mistake and the corresponding percentage.

4.2 Feature Extraction

A set of 30 frame-level spectral features, including high frequency content (HFC) [16], 13 Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral crest, spectral flatness, spectral flux, spectral roll-off, descriptors of spectral distribution (i.e., spectral variance, skewness and kurtosis), tristimulus [17], odd-to-even harmonic energy ratio (OER) [18], the estimated pitch, zero crossing rate and the instant power, were extracted from either the waveform or the spectrum using the ESSENTIA open-source library (version 2.0.1) [19]. The feature extraction was performed in each Hanning-windowed frame with the frame duration of 46 ms and the frame shift of 50%. These features are capable of characterizing timbre and regularly employed in audio signal processing applications [20]. The six temporal functionals, including mean, variance, skewness, kurtosis, mean and variance of the derivative, of all the frame-level features were derived to generate clip- or segment-level features. The outcome of the feature extraction stage is a feature vector of 180 dimensions.

The feature extraction process was done on different segments of notes resulting in five feature representations: *Note*, *Onset*, *Sustain*, *Offset* and *Cascade*. The *Note* representation was extracted from each intact note while the *Onset*, *Sustain* and *Offset* representations were extracted from corresponding segments of each note. These four representations consist of feature vectors of 180 dimensions. The *Cascade* representation was produced by concatenating the *Onset*, *Sustain* and *Offset* representations to give a 540-dimentional feature vector for each note. All feature representations were derived from 981 recorded notes and used for the task of playing mistake classification.

4.3 Feature Selection and Classification

The Fisher score was considered for selecting prominent features in a pre-processing step prior to classification to reduce amounts of computation [21]. It is defined as [22]

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n^{(+)}-1}\sum_{k=1}^{n_+}\left(\bar{x}_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n^{(-)}-1}\sum_{k=1}^{n_+}\left(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2},$$

where $n^{(+)}$ and $n^{(-)}$ are the numbers of positive and negative instances, respectively; $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the averages of the ith feature over the whole, positive, and negative instances, respectively; $\bar{x}_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $\bar{x}_{k,i}^{(-)}$ is the ith feature of the kth negative instance.

We followed the framework in [22] which selects features with high Fisher scores and uses the support vector machine (SVM), implemented by LIBSVM [23], for classification. The performance was evaluated in terms of the averaged F-measure, which is the harmonic mean of precision and recall, for each mistake using each feature representation with 100 repetitions of stratified five-fold cross-validation (CV).

5. EXPERIMENTS

The goal of the experiments is to investigate the capability of used features to detect playing mistakes and bridge the relation between playing mistakes and feature representations from different segments of notes.

Detection experiments were carried out through all feature representations after completing feature extraction. Following the procedures in [22], we first adopted a nested stratified five-fold CV to find the best percentage threshold to retain features based on Fisher scores, and then used the selected features for grid searching the optimized hyper-parameters C and γ, from the choices of $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ and $\{1, 10, 100, 1000\}$ respectively, of the radial-basis function (RBF) kernel based SVM. Finally, the selected threshold and the hyper-parameters were fed into another stratified five-fold CV. The overall performance was evaluated by averaging the F-measures of 100 repetitions of the final CV. Note that the above experiments were conducted for each feature representation and for each mistake. In other words, we trained M binary SVMs on each feature representation, where M is the number of types of playing mistakes.

To further have subset analysis, the experiments were conducted using three sets of data: *All*, *Down-bow*, and *Up-bow*, which respectively refer to the full data set of 981 notes, the set of 480 notes played with down-bow and the set of 481 notes played with up-bow. Moreover, we performed the same experiment on the 570 notes recorded by the six beginners who have played violin less than one month. Experiment results on these subset data and related discussions will be given in the next section.

6. RESULTS

The averaged F-measure using each feature representation for identifying each playing mistake in the *All* dataset is shown in Fig. 2. One can see that *Cascade* performs slightly better than *Note* in terms of the F-measure across all the mistakes, which is verified by the two-tailed t-test (p<0.01). It is probably because *Cascade* contains more detailed information of each individual segment. Except for the BB mistake, *Cascade* performs better than each of its constituents, i.e., *Onset*, *Sustain* and *Offset*. Note that the F-measures of the playing mistakes by the random guess would be 35.0%, 21.3%, 23.7% and 9.7%, respectively, equivalent to the prior probabilities $p(m)$ of the mistake m as shown in the second row of Table 1 divided by $p(m)$ + 0.5. It is because we preserved the prior distribution of the dataset in all partitions during the stratified five-fold CV procedures for each playing mistake. For comparison, we show in Fig 3 the performance of using the original 180 features without feature selection. Similar results between Figs. 2 and 3 suggest that the selected features sufficiently capture information embedded in the original 180 features for our experiments.

To explore more connections between playing mistakes and feature representations, one can re-arrange the F-measures of *Onset*, *Sustain* and *Offset* against mistakes as in Table 2. Results in Table 2 show that *Onset* has advantage in detecting SB over the others. It means that the onset segment is more sensitive for detecting SB, which somehow implies that the 10 players tended to have excessive bow pressure at the beginning of the bow draw. In contrast, *Sustain* surpasses the others in both CB and BB by up to 8% and 20%, respectively, which suggests CB and BB have higher chance to emerge during the middle of a drawing bow. Lastly, *Offset* dominates the IAH mistake. Such "favor" of a specific playing mistake in a particular segment of a note reveals the tendency of players to make that mistake at certain moment of a bow draw. This kind of information is helpful to novice players during their practice.

As shown in Table 2, SB and BB are prone to happen in the onset segment and sustain segment, respectively. Furthermore, it is commented by violin experts that such "favor" of SB and BB would be even more obvious in down-bow notes based on their teaching experiences. Figs. 4 and 5 compare the F-measures between the *Up-bow* and *Down-bow* subsets for the SB and BB mistakes, respectively. Obviously, these two figures indicate the down-bow notes are more associated with the mistakes than the up-bow notes, which is consistent with experiences of the violin experts.

Moreover, Fig. 6 shows the results on notes only played by the six beginners. It shows better overall performance than results in Fig. 2, which suggests notes played by the beginners reveal more obvious characteristics of mistakes than the ones played by experienced players. In other words, the adopted features might be incapable of capturing slight mistakes made by experienced players.

The inferior performance in classifying the IAH mistake, as shown in Figs. 2 and 6, might result from the severe imbalance of the dataset. In addition, pitch-related features are overwhelmed by timbre-related features in our adopted feature set. If more pitch features are considered, it is possible to further improve the performance for IAH detection, since it is about the mistake of playing undesired pitch.

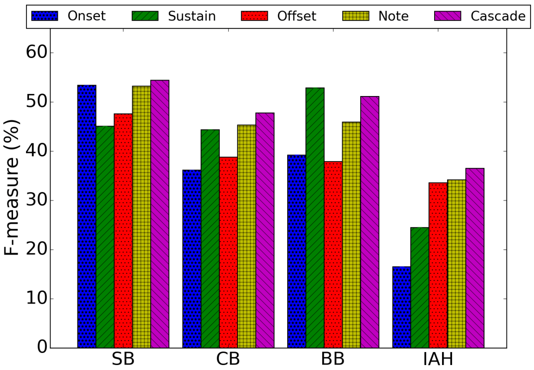

Figure 2. Averaged F-measures of playing mistake classification on all recorded notes using different feature representations.

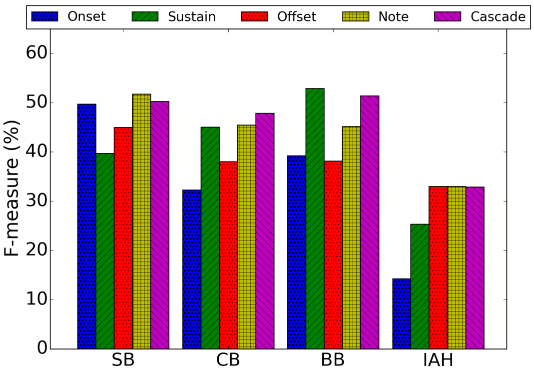

Figure 3. Average F-measures of playing mistake classification on all recorded notes using different feature representations from the original 180 features.

	SB	CB	BB	IAH
Onset	**53.4**	36.1	39.2	16.5
Sustain	45.0	**44.3**	**52.9**	24.4
Offset	47.6	38.8	37.9	**33.5**

Table 2. Averaged F-measures (in %) of *Onset*, *Sustain* and *Offset*. The feature representation with the highest F-measure for each mistake is highlighted.

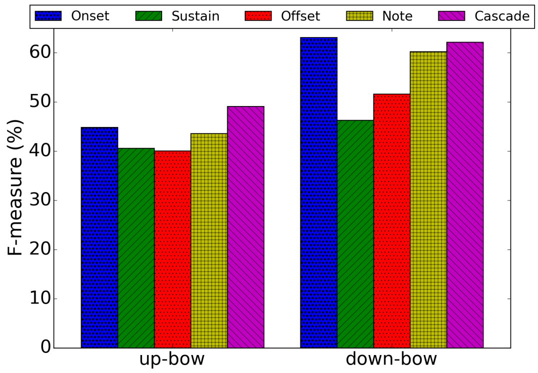

Figure 4. Averaged F-measures of the playing mistake 'scratching bow' (SB) using different feature representations within the *up-bow* and *down-bow* subsets.

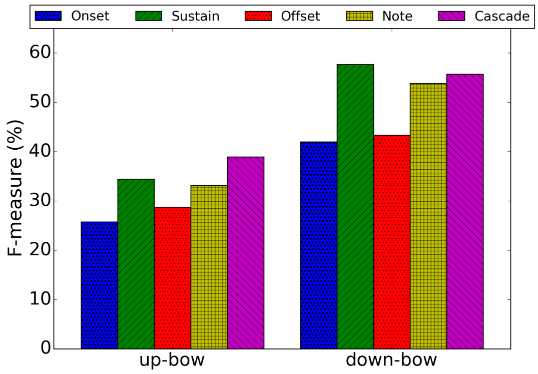

Figure 5. Averaged F-measures of the playing mistake 'bouncing bow' (BB) using different feature representations within the *up-bow* and *down-bow* subsets.

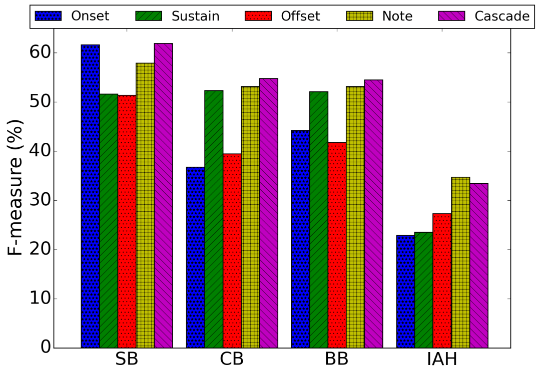

Figure 6. Averaged F-measures of playing mistake classification on notes played by beginners using different feature representations.

7. CONCLUSION AND FUTURE WORK

In this study, we first recorded a new dataset of violin legato notes played by novice players. Then we defined four common playing mistakes mainly made by bow arm and performed automatic playing mistake classification using spectral and temporal features extracted from different segments of the notes.

Our evaluation on different feature representations suggests concatenation of segment-level features provides more information than the note-level features in identifying playing mistakes. Furthermore, by exploring connections between playing mistakes and feature representations, we found SB, CB, BB, and IAH mistakes are prone to happen in the onset, sustain, sustain, and offset segments, respectively. These findings would serve pedagogical purpose and benefit novice violin players. Our future work will focus on improving the overall classification performance by enriching the dataset and seeking more relevant features, using either feature design or feature learning techniques [24, 25].

8. ACKNOWLEDGEMENTS

This research is supported by the National Science Council, Taiwan under Grant No NSC 102-2220-E-009-049, the Biomedical Electronics Translational Research Center, NCTU, and the Academia Sinica Career Development Award.

9. REFERENCE

[1] C. Dittmar, E. Cano, J. Abeßer and S. Grollmisch: “Music information retrieval meets music education,” *Multimodal Music Processing*. Dagstuhl Follow-Ups M. Müller, M. Goto and M. Schedl, Eds., vol. 3, pp. 95–120, 2012.

[2] I. Galamian: *Principals of Violin Playing and Teaching*, London, Prentice Hall, 1985.

[3] J. Wang, S. Wang, W. Chen, K. Chang and H. Chen: “Real-Time Pitch Training System for Violin Learners,” *Proc. IEEE Int. Conf. Multimedia and Expo Workshops*, pp. 163–168, 2012.

[4] B. Zhang and Y. Wang: *Automatic music transcription using audio-visual fusion for violin practice in home environment*, Tech. Report TRA7/09, School of Computing, National University of Singapore, 2009.

[5] E. Molina *et al.*: “Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* pp. 744-748, 2013.

[6] I. Barbancho, C. Bandera, A.M. Barbancho and L. J. Tarón: “Transcription and Expressiveness Detection System for Violin Music,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* pp. 189-192, 2013.

[7] J. Abeßer et al.: “Feature-based extraction of plucking and expression styles of the electric bass guitar,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* pp. 2290-2293, 2010.

[8] M. Robine and M. Lagrange: “Evaluation of the Technical Level of Saxophone Performers by Considering the Evolution of Spectral Parameters of the Sound,” *Proc. Int. Society for Music Information Retrieval Conference*, pp. 79-84, 2006.

[9] Y. Han and K. Lee: "Hierarchical approach to detect common mistakes of beginner flute players," *Proc. Int. Society for Music Information Retrieval Conference*, pp. 77-82, 2014.

[10] T. Knight, F. Upham, and I. Fujinaga: “The potential for automatic assessment of trumpet tone quality,” *Proc. Int. Society for Music Information Retrieval Conference*, pp. 573–578, 2011.

[11] J. Charles: *Playing Technique and Violin Timbre: Detecting Bad Playing*, Ph.D. dissertation, Dublin Institute of Technology, 2010.

[12] L. Auer: *Violin Playing As I Teach It*, Dover Publications Inc., New York, 1980.

[13] T. Grosshauser, and T. Gerhard: “Optical bow position, speed and contact point detection,” *Proc. ACM Int. Conf. Pervasive and Ubiquitous Computing Adjunct Publication*, 2013.

[14] M. Hajda: “A New Model for Segmenting the Envelope of Musical Signals: The Relative Salience of Steady State Versus Attack, Revisited,” *Audio Eng. Soc*, paper No. 4391, 1996.

[15] M. Caetano, J.J. Burred and X. Rodet: “Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues,” *Proc. Int. Conf. Digital Audio Effects*, 2010.

[16] P. Masri and A. Bateman: “Improved modellling of attack transients in music analysis-resynthesis,” *Proc. Int. Computer Music Conference*, pp. 100–103, 1996.

[17] G. Peeters: *A large set of audio features for sound description (similarity and classification) in the CUIDADO project,* CUIDADO I.S.T. Project Report, 2004.

[18] K. D. Martin and Y. E. Kim: “Musical instrument identification: A pattern-recognition approach,” *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1768–1768, 1998.

[19] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera and O. Mayor: “ESSENTIA: an audio analysis library for music information retrieval,”

Proc. Int. Society for Music Information Retrieval Conference, pp. 493-498, 2013.

[20] M. Müller et al: "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6,pp. 1088-1110, 2011.

[21] I. Guyon and A. Elisseeff: "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[22] Y.-W. Chen and C.-J. Lin: "Combining SVMs with various feature selection strategies," *Feature extraction foundations and applications*, pp. 315-324, 2006.

[23] C.-C. Chang and C.-J. Lin: "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011.

[24] P.-C. Li, L. Su, Y.-H. Yang and A. W. Y. Su: "Analysis of expressive musical terms in violin using score-informed and expression-based audio features," *Proc. Int. Society for Music Information Retrieval Conf.*, 2015.

[25] L. Su, L.-F. Yu and Y.-H. Yang: "Sparse cepstral and phase codes for guitar playing technique classification," *Proc. Int. Society for Music Information Retrieval Conf.*, pp. 9-14, 2014.