Identifying Master Violinists Using Note-level Audio Features

Yudong Zhao
Queen Mary University of London
yudong.zhao@qmul.ac.uk

György Fazekas
Queen Mary University of London
g.fazekas@qmul.ac.uk

Mark Sandler
Queen Mary University of London
mark.sandler@qmul.ac.uk

ABSTRACT

The same piece of music can be performed in various styles by different performers. Vibrato plays an important role in violin players' emotional expression, and it is an important factor of playing style while execution shows great diversity. Expressive timing is also an important factor to reflect individual play styles. In our study, we construct a novel dataset, which contains 15 concertos performed by 9 master violinists. Four vibrato features and one timing feature are extracted from the data, and we present a method based on the similarity of feature distribution to identify violinists using each feature alone and fusion of features. The result shows that vibrato features are helpful for the identification, but the timing feature performs better, yielding a precision of 0.751. In addition, although the accuracy obtained from fused features are lower than using timing alone, discrimination for each performer is improved.

1. INTRODUCTION

Music performance consists of two interdependent factors: musical form and structures established by composer and the interpretation by the performer [1]. Probably the most common structural characteristics of music are pitch and rhythm, which are typically defined by the composer. The factor that makes musical performance more colourful, attractive and unique is the interpretation. For example, although rhythm is defined by the composer, the tempo can be sped up or slowed down by the performer in a fairly flexible manner. There are other influential factors of musical performance including loudness, articulation, variation of timbre using different playing techniques as well as vibrato [2]. These techniques can be applied to varying extents and used differently among performers, resulting in different emotions perceived by, or evoked in listeners by different performances. Jung [3] analysed playing styles of three famous violinists, and argued for instance that Heifetz can be described as "unemotional" and "cold", whereas Oistrakh's performances always make listeners feel "warm" and rich in emotion. Therefore, expressive factors may have a great impact on the appraisal and appreciation of music performances.

The characteristic playing style developed by master violinists yields performance features, many of which can be observed in the audio signal. Other music representations such as scores are either void of these features or contain only minimal notation, which may still be interpreted differently by performers. Therefore signal processing and modeling methods are crucial in studying how characteristic playing styles are formed and in understanding which acoustic features are influenced by them the most. Applications of this knowledge include performer identification and helping music students to mimic the playing style of master violinists.

There are prior works on violin expression analysis and violinist classification. Pei-Ching Li and Li Su [4] developed a dataset containing 11 expressional items, then selected duration, dynamic and vibrato features to classify expressions using Support Vector Machine [5]. Ramirez and Maestre et al. [6] built a Celtic violinist classifier using machine learning. They extracted pitch, timing and amplitude features representing note-level characteristics and broader musical context. Molina et al. [7] proposed an approach for identifying violinists in monophonic audio recordings using a musical trend-based model. Chi-Ching Shih, Pei-Ching Li [8] used articulation and energy features to compare different playing styles of Heifetz and Oistrakh, arguably the most talented violinists in the world. To our knowledge however, most previous works attempted violinist identification using features of pitch, timing or energy. Such features are generally considered important for classification, while vibrato features are seldom used in this task. However, detailed vibrato features can be considered more important in understanding master violinists' playing styles and difference among individual expressions, and can be fused with timing features to get a better identification result.

In this paper, we compare classification methods for nine leading violinists using vibrato and timing features separately, followed by a feature fusion method to consider the features in combination. The structure of the performer identification method proposed in this paper and related experiments is summarised in the flow chart shown in Figure 1.

The rest of the paper is organised as follows: Section 2 introduces a novel dataset consisting of 162 solo vibrato notes and 3796 note onset times for each performer. Note-level excerpts from famous violin concertos played by all violinists were annotated manually. Section 3 presents the data pre-processing, feature extraction and selection method. Section 4 discusses the classification experiments and results. The overall conclusions and possible future developments are outlined in Section 5.

Copyright: © 2020 Yudong Zhao et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

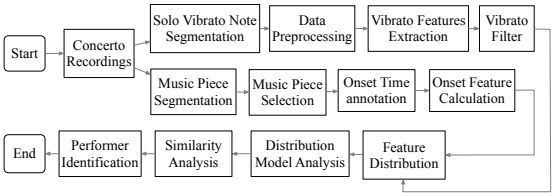


Figure 1. Flow chart of annotation and violinist classification.

2. DATASET

The concerto is a musical work that focuses on a solo instrument, such as the violin or piano, accompanied by an orchestra. It is paramount in the repertoire of master violinists who perform concertos very individually. Among other characteristics of music performance, tempo, intensity and vibratos are presented in a variety of different ways from person to person, since every violinist brings an individual style to the performance. For example, Heifetz plays the Beethoven D major violin concerto (III) faster than any other performer which, perhaps by intent, brings a feeling of "unemotional and cold" to listeners and differentiates his performances.

Most concertos contain a solo cadenza part. Performers can play this without concern to the coordination with the orchestra or obeying the global tempo. Violinists therefore often exhibit their unique playing style most expressively during the cadenza. Paying special attention to the cadenza is therefore very useful for our research aiming to understand how to model differences in individual playing style. In addition, we do not have to address the influence of accompaniment and can focus on features that may be extracted from the solo performance.

We select five concertos written by five well-known composers: Beethoven, Brahms, Mendelssohn, Tchaikovsky and Sibelius. These pieces have all been performed by nine violinists: Jascha Heifetz, Anne Sophie Mutter, David Oistrakh, Itzhak Perlman, Pinchas Zukerman, Isaac Stern, Salvatore Accardo, Yehudi Menuhin and Maxim Vengerov, who are all leading master violin players. We introduce the data annotation methods for two different kinds of features in the following two subsections. Further details of the recordings and the amount of annotated data are listed in Table 1.

2.1 Vibrato note data annotation

To reduce the influence of accompaniment on our features, we use solo notes that contain vibrato. We also sidestep the influence of variation between music pieces, therefore the same excerpts from each concerto for every performer are annotated. This way we can focus on the differences of vibrato characteristics between performers.

All vibratos were manually annotated at the note-level, including onset and offset times for note segmentation. Sonic Visualizer [9] was used for manual annotation, together with the Match Vamp plugin [10] to align the performances, guiding and improving the annotation performance. Figure 2 shows an example of the interface used for annotation and an excerpt of the data with several notes. In this plot, darker (purple) segments correspond to solo vibrato notes. Vibrato note onset and offset times are shown as dark purple vertical lines around segment boundaries.

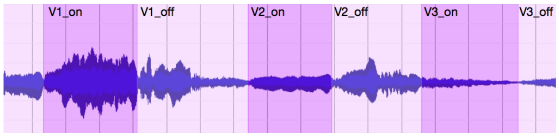


Figure 2. Vibrato notes segmentation (excerpt)

2.2 Note onset time annotation

The second data annotation task consists of labeling the note onset times for every music piece. In this task, the selection of music pieces are the same as above (see Section 2.1). Since we want to analyse the deviation in onset time among different performers while they play the same note, the onset time label of each note must be accurate. Although there are many existing automatic onset detectors, the accuracy on violin recordings is not high enough for our purpose, therefore we label onset times manually. Because there are not many violin solos in the 'prelude' or 'interlude' (sections that are performed by orchestra alone), we cut out the parts of the music without violin or where the violin cannot be heard clearly. Hence the commercial recordings are divided into several pieces before the data labeling. The overall procedure consists of the following three steps: music piece selection, music pieces alignment, and onset time labelling.

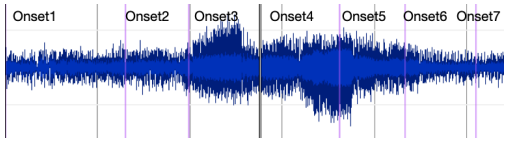


Figure 3. Note onset time annotations (excerpt)

The first step is to select the music pieces. We consider the impact of pieces themselves as well as note types. The speed of the performance and the onset time deviations can be very different for different parts of a concerto movement. For example, the start of a movement is always soft and slow, while the middle part is more varied and the ending is usually passionate. In addition, different note types such as semibreve, minim, crotchet, quaver, dotted note, etc. will also become factors affecting onset times. Therefore, we selected at least 3 different parts with different speeds and cover as many note types as possible from each movement to ensure the diversity of the data. To make the feature extraction easier, we align the music pieces and the start of the first note is considered as 0 seconds for each. Labelling short notes with correct onset times is a substantial challenge. For example, it is not easy to label a performance with a quick succession of very short sixteenth notes. To solve this, we slow down the music using the appropriate function provided in Sonic Visualizer [9].

Figure 3 shows the interface used for onset time annotation. The vertical line indicates the position of onset times. The number of annotated onset times from each movement is listed in Table 1.

Table 1. Details of recording selection and data annotation

Composer	Concerto	Movement	No. of vibratos	No. of onsets
L. V. Beethoven	V.Concerto D major, Op.61	I	21	572
		II	26	271
		III	4	328
J.Brahms	V.Concerto D major, Op.77	I	11	214
		II	6	196
		III	4	203
F.Mendelssohn	V.Concerto E minor, Op.64	I	13	195
		II	N/A	196
		III	3	211
P.I. Tchaikovsky	V.Concerto D major, Op.35	I	26	189
		II	7	191
		III	18	209
J.Sibelius	V.Concerto D minor, Op.47	I	23	188
		II	N/A	191
		III	N/A	209

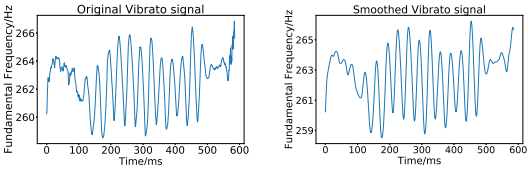
3. METHODOLOGY

As mentioned in Section 1, the flow of the performer identification methods proposed in this paper is shown in Figure 1. There are two branches after the selection of "Concerto recordings", which represent two different feature extraction methods. These are summarised first while details of each steps are given in the following subsections. lin

In the first method, monophonic vibrato notes are annotated manually from the original recordings. However, we cannot extract vibrato features from audio waveform directly. Therefore, we use the PYIN [11] algorithm to obtain the fundamental frequency of each annotated note, so that the change of the pitch within every note can be observed. Since all vibrato features are extracted from the pitch curve of each note, fundamental frequency estimation is the first step. To avoid the interference of noise and make all vibrato features extracted from relevant vibrato data, the frequency signal is smoothed before feature extraction. These two steps are denoted *preprocessing* and detailed in Sec. 3.1. We introduce specific vibrato features in Sec. 3.2. Similar structure can be found in the second method, which is also shown in the flow chart in Fig. 1. However, there are some differences in the data annotation and feature extraction process, which will be explained in Sec. 3.2. Finally, we find differences of individual vibrato and timing characteristics by comparing their feature distributions. The classification process using these features is introduced in Sec. 3.3.

3.1 Preprocessing

Vibrato is a phenomenon of oscillating pitch that is related to fundamental frequency. After vibrato note segmentation, we obtain waveforms of each note from the original audio ($F_S = 44.1kHz$). In this research, a smoothed pitch track is calculated using the PYIN algorithm [11] with a frame size and hop size of 2048 and 256 respectively. We thus obtain an array of instantaneous frequency values. However, this array exhibits some noise and artefacts near note boundaries. Figure4(a) shows the fundamental frequency (F_0) estimation curve of a note with noise around



(a) Original vibrato pitch curve (b) Smoothed pitch curve

Figure 4. Pitch curve before smoothed and after

the location of onset and offset times. This is smoothed to obtain reliable vibrato features using a zero-phase Butterworth low-pass filter to avoid phase delay. This filter is designed in the scipy package [12]. The smoothed signal is shown in Figure 4(b). Compared to the original pitch curve, the clarity is improved but some small fluctuations remain around the boundaries. This issue is addressed in the following steps.

3.2 Feature Extraction

We design four note-level vibrato features and an onset feature. To characterise vibrato, we extract average vibrato extent (AE), average vibrato rate (AR), standard deviation of vibrato extent (SE) and standard deviation of vibrato rate (SR). All features are computed from the fundamental frequency estimates. Additionally, we calculate the onset time deviation (OTD) as a feature related to expressive timing. The feature extraction process for these are described in Section 3.2.3

3.2.1 Vibrato extent

In every period of the pitch curve, the instantaneous vibrato extent is considered to be the distance between an adjacent peak and trough. The average and standard deviation of vibrato extent is calculated from all instant vibrato extent values within a note. First we find the location of every peak and trough contained in the pitch curve by locating maxima and minima in the smoothed F_0 array. We then calculate the absolute frequency distance between successive peaks and troughs to obtain the instantaneous vibrato extent. The collection of note-level instant vibrato extents are used to calculate the average and standard deviation of vibrato extent for all annotated notes.

3.2.2 Vibrato rate

After obtaining the locations of every peak and trough in the pitch curve of a note, the vibrato rate features can be calculated easily. We find the time when every peak and trough appears in the pitch curve. The interval between adjacent peaks and troughs can be considered a half period (t_h), then the rough instant vibrato rate can be calculated using the formula in Eq. 1.

$$Vibratorate = \frac{1}{2t_h} \tag{1}$$

However, as shown in Figure 4(b), although the pitch curve has been smoothed before the feature extraction, there are still low amplitude oscillations present that do not

correspond to the player's vibrato. We consider a heuristic to eliminate the effect of this. In general, the range of vibrato rate is 2 Hz to 15 Hz, and the range of vibrato extent is between 9 cents and 50 cents. After extracting the rough instant vibrato extent and rate at the note-level, we discard values outside these ranges.

3.2.3 Onset time deviation

After the segmentation of music pieces, as well as alignment and onset time annotation (see Sec. 2.2), we calculate the onset time deviation (OTD). The first step is to obtain the reference onset time of each note. There are two viable methods to consider: the first is using score based note onset time as reference, the second is using the mean note onset time of each note across all performances in the dataset. In this paper, we use the latter approach, since we can assume that averaging removes most of the expressive timing and individual interpretation of the performer, except for a generally accepted interpretation of the piece, where such interpretation exists. This approach also avoids the need for audio to score alignment.

As Figure 5 shows, the first note of every piece is aligned in time. Using the alignment between individual performances, the corresponding notes are identified first, then the mean onset time of each note is calculated from all violinists' performances as the reference time. This is followed by the calculation of onset time deviations from this reference for each performer to characterise expressive timing. For example, a score is shown at the top of Figure 5. The different vertical bars indicate note duration from different performers. The vertical dashed lines are the average onset times of each notes in this piece, which are treated as the "reference onset time" for the rest of this paper. Then the distance between each actual onset time and the reference time is the onset time deviation, with some examples indicated in Fig. 5 as well.

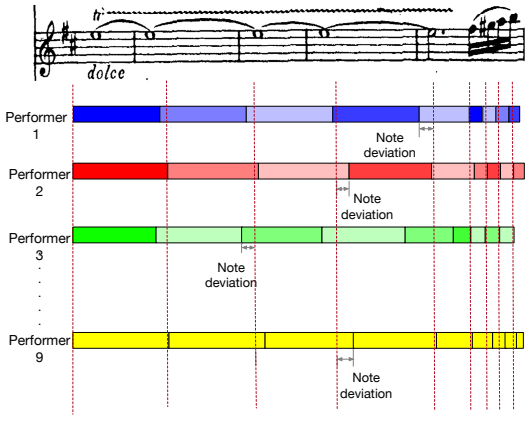
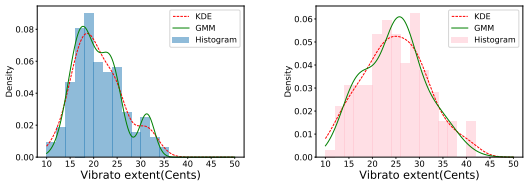


Figure 5. Expressive timing feature extraction

There are other features of violin performance currently ignored by our method. This includes phrasing as well as the dependence of the note onset times and vibrato technique on motion [13] and differences in the characteristics of the actual instrument. Addressing these issues constitute future work. For example the dependence on phrasing and previous onsets may be addressed using Bayesian techniques. Instrument specific features may also be developed, for instance, detecting the amount of sympathetic resonance has been shown to be useful for characterising instrumental gestures in [14].



(a) Heifetz (b) Mutter

Figure 6. Distribution of two performer's average vibrato extent

3.3 Violinist Classification

In this section, we describe our violinist identification method, consisting of the estimation of feature distributions, similarity calculation and classification. All methods are first presented using the features separately. A method using feature fusion is discussed in Section 3.3.3.

3.3.1 Classification based on vibrato features

When different performers play the same music piece, they typically use different vibrato rates and extents in their respective performances. Therefore, we model the vibrato characteristics of each performer using the distribution of these features. In this paper, we calculate histograms, kernel densities (KDE) and Gaussian Mixture Models (GMM) separately to model these distributions, assuming that these provide compact representations of the violinists' style, which we can use later for identification. Figure 6 shows how the global distribution of average vibrato extent for Heifetz and Mutter differs for example. We can easily see that the highest density of the vibrato extent distribution appears between 15 cents and 20 cents for Heifetz, but it is 20 cents to 25 cents as well as 29 to 30 cents in Mutter's performances. In addition, there is no vibrato greater than 35 cents in Heifetz's performances, whereas the maximum vibrato extent reaches above 40 cents in Mutter's. This shows that Heifetz prefers to use the vibrato in a smaller scale, but Mutter's vibrato extents are broader. Based on similar observations for several performers, we can assume that the feature reflects an important aspect of the vibrato characteristics for every performer.

In Fig. 6, the red line shows the Gaussian kernel to estimate kernel density of average vibrato extent data from Heifetz and Mutter as well, the curve of the two distributions show similar properties to histograms. We also train a 3-component Gaussian Mixture Model to estimate the distribution of the data. Their PDF curves are shown in Fig. 6 too using continuous (green) line. The number of components in these models is selected using empirical observation, i.e., the distributions do not generally exceed three

modes so the GMMs represent the histograms and kernel densities well. Given these curves, we can observe the continuous distributions of features for each performer, and their differences should reflect individual characteristics.

In order to quantify these differences, we calculate the similarity of distributions of each given feature for all performers using the Kullback-Leibler (KL) divergence [15] shown in Equation 2. This corresponds to the likelihood ratio between two distributions and tells us how well the probability distribution Q approximates the probability distribution P by computing the cross-entropy minus the entropy. The KL divergence between kernel densities is estimated using the approach proposed in [16]. Since the KL divergence between GMMs is not analytically tractable, we use variational Bayes approximation following the implementation in [17].

$$D_{KL}(P\,||\,Q)) = \sum_i P(i) log(\frac{P(i)}{Q(i)}) \tag{2}$$

For classification, the KL divergence can be calculated between vibrato feature distributions of an unknown performer and every known performer in the dataset. Finding the minimum divergence between an unknown and known performer should help to identify the unknown performer.

3.3.2 Classification based on onset time deviation

The playing style of each performer has a great impact on the expressive timing. For example, Heifetz always plays fast no matter which piece he is playing. To characterise expressive timing statistically, we compute the distribution of all timing deviations using the same three statistical approaches used for vibrato features: Histogram, KDE and GMM. Then we measure the similarity between training data and a test set corresponding to a performer using the KL divergence. The minimum divergence provides the identity of the performer.

3.3.3 Fusion method

In this research, we use linear combination with equal weights to fuse similarity estimates for the distributions of different features summarised in Table 2. During the evaluation, leave one group out cross validation with 8 folds is used to calculate the KL divergence between training set and testing set for every group of data. The similarity estimates of feature distributions in every fold are combined for the different kinds of features using the approach shown in Equation 3:

$$KL_{overall} = \sum_{n=1}^{|\Theta|} w_n KL_{\Theta_n}, \tag{3}$$

where $\Theta = \{V_1, V_2, V_3, V_4, T_1\}$ with $V_1, ..., V_4$ denoting the sets of statistical models corresponding to four kinds of vibrato features (AE,AR,SE,SR) computed separately, while T_1 corresponds to the OTD (see Sect. 3.2.3). All corresponding weights w_n are set to one in the current implementation, however, feature importance may be investigated in future work using methods discussed in [18] in the audio context. Moreover, the way how the features are fused is not unique. We can combine for instance any 3 or 4 features together to compute the overall KL divergence. Next, we validate if this mechanism works for violinist identification and test how accurate the classification results are for different performers. The design of the experiment and the results are discussed in Section 4.

4. EXPERIMENTS AND RESULTS

We test the effect of proposed identification mechanism using leave one group out cross validation and show the classification result (F-measure) and confusion matrix for all performers in our dataset. In this section, we will show the violinist identification results based on different features: vibrato features only, timing features only and the combination of features using the fusion method described in Section 3.3.3. A summary of the features used in this paper and their abbreviations are listed in Table 2 for clarity.

Table 2. Summary of features and abbreviations

Original Feature Name	**Shortened Name**
Average Vibrato Extent	AE
Average Vibrato Rate	AR
Standard Deviation Vibrato Extent	SE
Standard Deviation Vibrato Rate	SR
Combination of all Vibrato features	VC
Onset Time Deviation	OTD

4.1 Classification result based on vibrato features

There are 12 music pieces each containing different amounts of annotated solo vibrato notes ranging from 4 to 20. To ensure there is sufficient data in every test set given the small size of our dataset, we perform eight folds Leave one group out cross validation to assess the different features and statistical representations.

In the experiment, we first designate one group of data from a random performer as test data.We then compute the distribution of four vibrato features based on the test data and every other group in the training set separately. The number and size of histogram bins, as well as the KDE kernel and GMM hyper parameters are kept constant.

As mentioned above, we have separated each performer's data into 8 groups, so that for each test player, we can get eight group-level distributions for each feature analysis. Furthermore, since there are four vibrato features in all, 32 distributions for one test performer can be obtained to reflect his vibrato characteristics. Then, we compute the KL divergence between each feature's distribution from test performer and the same features for every performer in the training data. The similarity results for vibrato characteristics based on four features can be obtained between the test performer and every performer in the training set. The smaller the KL divergence the greater the similarity, therefore we treat the performer that corresponds to the minimum value as the identified performer with each feature. Finally, we also compute the mean of normalized KL divergences. These are donoted as 'combination' features (VC).

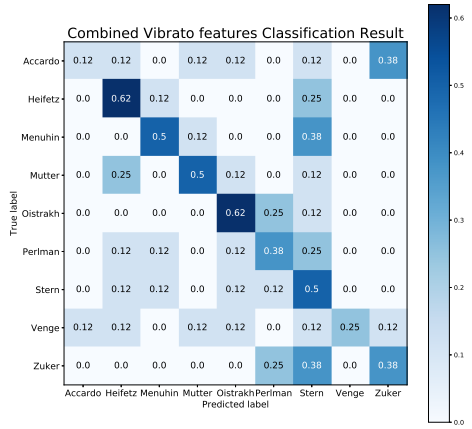


Figure 7. Violinists classification using combined vibrato features

From the cross validation, we get the similarity of vibrato features between every two performers in the dataset and performer identification based on one feature or the combination feature can be obtained. Table 3 shows the precision of violinist identification using three distributions separately. The combination feature performs better than any single ones, and the histogram model yields the best result overall. Fig. 7 shows the normalised confusion matrix of violinist identification using the combined vibrato features and histogram distribution. To avoid overlap of performer names in the X-axis of the confusion matrix, we abbreviate 'Vengerov' as 'Venge'; 'Zukerman' as 'Zuker' in the confusion matrix plots.

Table 3. Violinist identification result

Precision / Feature / Model	VC	AE	AR	SE	SR	OTD
Histogram	0.513	0.243	0.385	0.136	0.114	0.697
KDE	0.392	0.252	0.220	0.107	0.014	0.751
GMM	0.417	0.189	0.352	0.148	0.007	0.734

Table 4. Violinist identification result using histogram

Feature	Precision	Recall	F-score
VC	0.513	0.430	0.425
AE	0.243	0.264	0.242
AR	0.385	0.306	0.292
SE	0.136	0.148	0.132
SR	0.114	0.097	0.090

4.2 Classification result based on timing features

There are 3739 annotated notes extracted from each violinist's performance, and the same amount of onset time deviations are calculated from these notes. In this experiment, a similar approach is taken to the assessment of vibrato features. Again, to ensure there are enough data in every test set, and also a reasonably high number of cross validation folds, we perform eight folds Leave One Group Out cross validation to test the feature and the models. We first split each performer's feature data into eight groups on average, which means there are 467 notes in each of the first 7 groups, and 470 notes in the last group. Then we select a random test performer and designate one group as test data while the rest of the groups from all performers is considered training data. Then, the distribution of test data and every other group in the training set are obtained by using the histogram, KDE and GMM separately. The number and size of histogram bins, as well as the KDE kernel and GMM hyper parameters are kept constant. Finally, we measure the similarity between the test performer and every performer in the training set by calculating the KL divergence, so that the minimum value identifies the performer. Fig. 8 shows the normalised confusion matrix of violinist identification using the timing features and KDE distribution, the comparison of different distributions is listed in Table 4 as well.

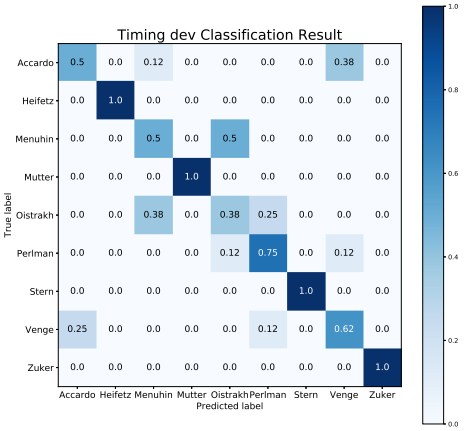


Figure 8. Violinists classification using Onset time deviation

4.3 Classification result using the fusion of features

As mentioned previously, there are 5 note-level features extracted from the music signal. According to the identification results based on vibrato features, AE and AR perform much better than the other two features. Therefore it is sensible to assess features in different combinations. Due to the low number of features we sidestep the use of complex feature selection methods. In addition, in table 3, it is obvious to find that histogram performs better while we use vibrato features to classify violinists. However, KDE performs better while we use onset feature. Therefore, in this experiment, the histogram is used to present vibrato features, whereas KDE is used as onset feature dis-

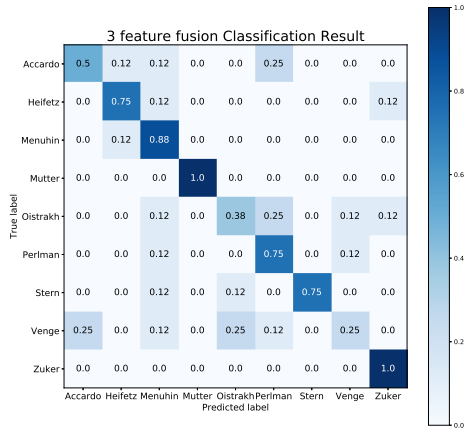


Figure 9. Violinists classification using the fusion of 3 features

tribution. We firstly combine timing feature with AE and AR together. The result is shown as '3 Feature Fusion ($3FF$)' in Table 5. Then, we fuse four features in two ways: timing feature fused with AE, AR, SE, which is presented as $4FF_1$; timing feature fused with AE, AR, SR, this is denoted $4FF_2$. Finally we fuse the timing feature with all vibrato features, whose results are denoted $5FF$ in Table 5 as well. We can observe that $3FF$ performs best in terms of precision. The corresponding confusion matrix is shown in Figure 9.

Table 5. Violinist identification using different fusion features

Feature	Precision	Recall	F-score
3 FF	0.700	0.694	0.681
4 FF_1	0.651	0.652	0.632
4 FF_2	0.640	0.639	0.626
5 FF	0.670	0.639	0.635

5. DISCUSSIONS AND CONCLUSIONS

Given the results obtained from the vibrato features only, Figure 7 shows that the discrimination for Heifetz and Oistrakh are best (0.62), while the identification for Menuhin, Mutter, Perlman, Stern and Zukerman are also good. Their performances can be identified correctly most of the time. However, the identification for Vengerov and Zukerman are less reliable. Their performances are confused with some of other performers leading to incorrect classification. The worst result is obtained for Accardo. His performance is confused with Zukerman's, i.e., we cannot obtain the correct identification from his vibrato characteristics, probably due to his limited use of the technique. We also compute the macro F-score based on single features and the combination feature. The result is shown in Table 4. The best performing feature is *average vibrato rate*, with a macro F-score of 0.385, and the worst is *Standard deviation of vibrato rate*.

No matter which distribution model is used, timing features perform clearly better than vibrato features. This might partly be due to the smaller size of the vibrato data. The highest precision is obtained using KDE distribution, which is 0.751; whereas histogram performs worse for this feature. However, although the overall result is the best among all experiments, there are still some misclassifications. From Figure 8, it is easy to discover that Menuhin's performance is confused with Oistrakh's, which means they have similar timing feature distributions. Using the KDE-based model, it can be observed that the shape of their pdf contours are more similar to each other. We can thus conclude that this timing feature based method is promising as it works very well for most performer identifications in our dataset, but it may still yield some confusion between certain performers.

To address this problem, we tried to classify performers using different fusion features. As explained above, we used 4 ways to fuse features. The $4FF_2$ performs worse in Table 9, which shows that SR contributes mainly noise for this identification task, and we can remove this feature in the future. Meanwhile, the F-score and precision of 3 features fusion performs best. According to Figure 9, although the overall F-measure results is a bit lower than using timing feature only, the discrimination for every performer is higher than using timing feature only. So there is an obvious improvement in discrimination using the combination of features.

In future work, we may fuse different features using different weights or design more features based on spectral characteristics, dynamics, or features that correlate with timbre or timbre changes during vibrato playing. We found that some vibrato features work better than others. We can therefore combine different vibrato features using different weights instead of a uniform weight. Furthermore, we may test other classification mechanisms including SVMs, Decision Trees or Neural Networks.

A potentially interesting future direction is the use of a "semantic space" describing differences in timbral characteristics resulting from different playing techniques. Semantic descriptors may be obtained using a combination of acoustic features and machine learning models as in [19], where the authors used machine learning to model the associations between semantic terms used by violin makers and acoustic features produced by the instruments described by them. Similar techniques may be useful to learn relationships between expressive performance descriptors and the sound produced by different performers. This poses interesting future challenges in differentiating between the effects of the acoustic environment, the specific instrument, different playing techniques and the individual player. Emerging technologies, such as the work of Engel et.al. [20], may enable us to model sound production mechanisms, playing techniques and the effects of the environment separately.

Using other data modalities, such as motion data obtained

from sensors as in [21,22] may also be useful in certain application contexts, such as estimating how well a beginner approximates a master player. Our current method benefits from the non-intrusive nature of using only the audio, but this may also be seen as a limitation when complementary information in motion data and instrumental gestures are considered.

Finally, we also aim to enlarge the size of vibrato dataset. It would be beneficial to test the method with more than 500 annotated notes for each performer and 20 or more performers.

In summary, we first construct a novel dataset from nine master violinists' performance. Four kinds of vibrato features are proposed and extracted along with one feature related to expressive timing. We propose a method to identify different violinists using the distributions of each feature. The results show that our proposed method works reasonably well for the identification of master players in our dataset.

6. REFERENCES

[1] S. Vieillard, M. Roy, and I. Peretz, "Expressiveness in musical emotions," *Psychological research*, vol. 76, no. 5, pp. 641–653, 2012.

[2] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological bulletin*, vol. 129, no. 5, p. 770, 2003.

[3] J. W. N. Jung, "Jascha Heifetz, David Oistrakh, Joseph Szigeti," 2007.

[4] P.-C. Li, L. Su, Y.-h. Yang, A. W. Su *et al.*, "Analysis of expressive musical terms in violin using score-informed and expression-based audio features." in *ISMIR*, 2015, pp. 809–815.

[5] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA, MIT Press, 2002.

[6] R. Ramirez, E. Maestre, A. Perez, and X. Serra, "Automatic performer identification in celtic violin audio recordings," *Journal of New Music Research*, vol. 40, no. 2, pp. 165–174, 2011.

[7] M. Molina-Solana, J. Lluís Arcos, and E. Gomez, "Identifying violin performers by their expressive trends," *Intelligent Data Analysis*, vol. 14, no. 5, pp. 555–571, 2010.

[8] C.-C. Shih, P.-C. Li, Y.-J. Lin, A. Su, L. Su, and Y. Yang, "Analysis and synthesis of the violin playing styles of Heifetz and Oistrakh," in *Proc. Int. Conf. Digital Audio Effects*.

[9] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1467–1468.

[10] S. Dixon and G. Widmer, "Match: A music alignment tool chest." in *ISMIR*, 2005, pp. 492–497.

[11] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.

[12] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[13] A. Friberg and J. Sundberg, "Does Music Performance Allude to Locomotion? a Model of Final Ritardandi Derived From Measurements of Stopping Runners," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1469–84.

[14] B. Liang, G. Fazekas, and M. Sandler, "Piano legato-pedal onset detection based on a sympathetic resonance measure," in *26th European Signal Processing Conference (EUSIPCO), 3-7 Sept., Rome, Italy*, 2018, pp. 2484–2488.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[16] H. Sasaki, Y.-K. Noh, and M. Sugiyama, "Direct density-derivative estimation and its application in kl-divergence approximation," in *Artificial Intelligence and Statistics*, 2015, pp. 809–818.

[17] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proceedig of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[18] D. Sheng and G. Fazekas, "Feature selection for dynamic range compressor parameter estimation," in *144th Convention of the Audio Engineering Society, 23-26 May, Milan, Italy*, 2018.

[19] M. Zanoni, F. Setragno, F. Antonacci, A. Sarti, G. Fazekas, and M. Sandler, "Training-based semantic descriptors modeling for violin quality sound characterization," in *138th Audio Engineering Society Convention, 7-10 May, Warsaw, Poland*, 2015.

[20] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *International Conference on Learning Representations*, 2020.

[21] L. S. Pardue, C. Harte, and A. McPherson, "A low-cost real-time tracking system for violin," *Journal of New Music Research*, vol. 44, no. 4, pp. 305–323.

[22] B. Liang, G. Fazekas, and M. Sandler, "Measurement, Recognition and Visualisation of Piano Pedalling Gestures and Techniques," *JAES Special Issue on Participatory Sound And Music Interaction Using Semantic Audio*, vol. 66, no. 6, pp. 448–456.