ANALYSIS OF EXPRESSIVE MUSICAL TERMS IN VIOLIN USING SCORE-INFORMED AND EXPRESSION-BASED AUDIO FEATURES

Pei-Ching Li[1] **Li Su**[2] **Yi-Hsuan Yang**[2] **Alvin W. Y. Su**[1]
[1] SCREAM Lab., Department of CSIE, National Cheng-Kung University, Taiwan
[2] MAC Lab., CITI, Academia Sinica, Taiwan
p78021015@mail.ncku.edu.tw, lisu@citi.sinica.edu.tw
yang@citi.sinica.edu.tw, alvinsu@mail.ncku.edu.tw

ABSTRACT

The manipulation of different interpretational factors, including dynamics, duration, and vibrato, constitutes the realization of different expressions in music. Therefore, a deeper understanding of the workings of these factors is critical for advanced expressive synthesis and computer-aided music education. In this paper, we propose the novel task of automatic expressive musical term classification as a direct means to study the interpretational factors. Specifically, we consider up to 10 expressive musical terms, such as *Scherzando* and *Tranquillo*, and compile a new dataset of solo violin excerpts featuring the realization of different expressive terms by different musicians for the same set of classical music pieces. Under a score-informed scheme, we design and evaluate a number of note-level features characterizing the interpretational aspects of music for the classification task. Our evaluation shows that the proposed features lead to significantly higher classification accuracy than a baseline feature set commonly used in music information retrieval tasks. Moreover, taking the contrast of feature values between an expressive and its corresponding non-expressive version (if given) of a music piece greatly improves the accuracy in classifying the presented expressive one. We also draw insights from analyzing the feature relevance and the class-wise accuracy of the prediction.

1. INTRODUCTION

The expressive meaning of music is generally related to two inter-dependent factors: the *structure* established by the composer (e.g., mode, pitch, or dissonance) and the *interpretation* of the performer (e.g., expression) [21]. Glenn Gould could phrase the *trills* in a way different from other pianists. Mozart's *Grazioso* should be interpreted unalike to Brahms'. Although the interplay between the structural and interpretational factors makes it difficult to characterize musical expressiveness from audio signals, it has been pointed out that such analysis is valuable in emerging applications such as automatic music transcription, computer-aided music education, or expressive music synthesis [2, 4, 7, 19]. Accordingly, computational analysis of the interpretational aspects in music expression has been studied for a while. For example, Bresin *et al.* analyzed the statistical behaviors of *legato* and *staccato* played with 9 expressive adjectives (not expressive musical terms) [3]. Grachten *et al.* made both predictive and explanatory modeling on the dynamic markings (e.g., *f*, *p*, *fz*, and *crescendo*) [10]. Ramirez *et al.* considered an approach of evolutionary computing for general timing and energy expressiveness [18]. Marchini *et al.* analyzed the performance of string quartets by the following three terms: *mechanical*, *normal* and *exaggerated* [14]. Recently, Rodà *et al.* further considered expressive constants as affective dimensions of music [20]. Related works also include the identification of performers, singers and instrument playing techniques in the context of musical expression [1, 6, 12, 15].

To model specific aspects of the complicated music expression quantitatively, a machine learning based approach is usually taken. Given an audio input, features are extracted to characterize the interpretational aspects of music, such as the dynamics, tempo and vibrato [3, 9, 12, 14]. [1] If the symbolic or score data such as the MIDI or MusicXML are available, one can further introduce more structural aspects including tonality, pitch, note duration and measure, amongst others [10, 15, 16]. In [14], the synchronized audio, score and even motion data are utilized to generate 4 sets of features, including sound level, note lengthening, vibrato extent and bow velocity, in an attempt to reveal human behaviors while playing the instrument or indicate the structural information of music. This way, the features investigated have music meanings, and can be adopted for specific applications such as the prediction and the generation of expressive performances [10, 18].

Among all the objects of music expression, we notice that the *expressive musical terms* (EMT) [2] have garnered less attention in the literature, although they have been

[1] Here we assume that any real-world interpretation of an expressive musical term performed by a musician can be "atomized" into several (independent) factors such as dynamics, tempo, and vibrato.

[2] In this paper, the expressive musical term is defined as the Italian musical term which describes an emotion, feeling, image or metaphor, rather than merely an indication of tempo or dynamics. It includes, but not limited to the emotional terms (see Table 1).

Violin pieces	Measure	Expressions
W. A. Mozart - *Variationen*	1-24	*None, Scherzando, Tranquillo, Con Brio, Maestoso, Risoluto*
T. A. Vitali - *Chaconne*	1-9	*None, Scherzando, Affettuoso, Con Brio, Agitato, Cantabile*
G. Faure - *Elegie*	2-9	*None, Scherzando, Grazioso, Agitato, Espressivo, Cantabile*
P. I. Tchaikovsky - *String Quartet, No. 1, Mov. II*	1-16	*None, Affettuoso, Tranquillo, Con Brio, Cantabile, Risoluto*
M. Bruch - *Violin Concerto, No. 1, Mov. I*	6, 10 (solo. ad lib.)	*None, Affettuoso, Tranquillo, Agitato, Maestoso, Cantabile*
A. Vivaldi - *La primavera, Mov. I*	1-13	*None, Scherzando, Affettuoso, Grazioso, Con Brio, Risoluto*
A. Vivaldi - *La primavera, Mov. II*	2-11	*None, Grazioso, Agitato, Espressivo, Maestoso, Cantabile*
E. Elgar - *Salut d'Amour*	3-17	*None, Affettuoso, Grazioso, Agitato, Espressivo, Maestoso*
A. Vivaldi - *L'autunno, Mov. I*	1-13	*None, Tranquillo, Grazioso, Con Brio, Espressivo, Risoluto*
A. Vivaldi - *L'autunno, Mov. III*	1-29	*None, Scherzando, Tranquillo, Espressivo, Maestoso, Risoluto*

Table 1: The proposed dataset contains 10 different classical music pieces and each with 6 distinct expressions.

widely used in specifying expressions of classical music for hundreds of years. How the interpretational factors (dynamics, duration or vibrato) are taken for a musician to interpret the terms is still not well understood. This might be due to the lack of a dataset containing various interpretations for a fixed set of classical music pieces.

In this paper we address these issues, and particularly, focus on the classification of expressive musical terms in violin solo music. We compile a new dataset of solo violin excerpts featuring the realization of 10 expressive terms and 1 non-expressive term (e.g., no expression) by 11 different musicians for 10 classical music pieces (Section 2). After collecting the MIDI and MusicXML data for the music pieces, we design a number of dynamic-, duration- and vibrato-based features under a score-informed scheme (Section 3.2). Moreover, we also consider a baseline feature set comprising of standard audio features that can be computed without score information, such as the Mel-frequency cepstral coefficients (MFCCs), spectral flux, spectral centroid, and the zero-crossing rate (Section 3.1). As such features have been widely used in music information retrieval tasks like the classification of mood, genre or instruments [25], we want to know whether they are also useful for classifying the expressive musical terms. However, we should note that many of the baseline features do not bear clear music meanings as the proposed features do. In our experiments, we will evaluate the performance of these features for expressive musical term classification, and analyze the importance of such features (Section 4).

The dataset is referred to as the SCREAM-MAC-EMT dataset. For reproducibility and for calling more attention to this research problem, we have made the audio files of the recordings publicly available online. [3]

2. THE SCREAM-MAC-EMT DATASET

To find out how a violinist interprets the expressive musical terms, the scope of the music data, the difference in personal interpretation, and the suitability between the music piece and the musical term are all considered. We started by listing 20 typical violin pieces ranging across the Baroque, Classical, and Romantic eras, such as Vivaldi's *The Four Seasons*, Beethoven's *Spring*, and Schubert's *Ave Maria*, to name a few. Then, we consulted with 3 professional violinists, who are active in classical music performance, to select 10 pieces from the list and assign 5 suitable expressive musical terms for each of them. The major criterion of selecting the music pieces, as it turns out, requires that an excerpt has a simple melody that can be effectively manipulated to exhibit different characteristics when being interpreted with different expressions.

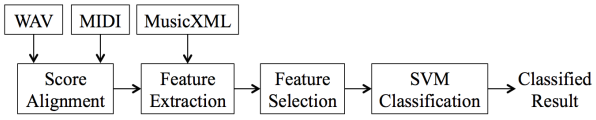

Figure 1: Flowchart of the proposed system.

The following 10 expressive terms are considered: *Tranquillo* (calm), *Grazioso* (graceful), *Scherzando* (playful), *Risoluto* (rigid), *Maestoso* (majestic), *Affettuoso* (affectionate), *Espressivo* (expressive), *Agitato* (agitated), *Con Brio* (bright), and *Cantabile* (like singing). [4] In order to have a balanced dataset, we require that each expressive musical term is associated with 5 pieces. This is not easy, because not all of the 20 pieces can be interpreted with diverse expressions. Eventually, some compromises have to be made. For example, we chose *Maestoso* instead of *Cantabile* for Elgar's *Salut d'Amour*, although the former is somewhat awkward for this music piece. The resulting selection of the music pieces and the assigned expressions is shown in Table 1.

After selecting the music pieces, we recruited 11 professional violinists to perform them one by one in a real-world environment. In addition to the 5 assigned terms, every musician performed a non-expressive (denoted as *None*) version for each piece. Here, *None* means mechanical interpretation [14] by which the music is of constant dynamics, constant tempo and no vibrato. The dataset therefore contains 660 excerpts as there are 10 classical music pieces and each piece is interpreted by 6 different versions by all the 11 violists. We have 110 excerpts of *None*, and 55 excerpts for each of the 10 expressions.

3. METHOD

Figure 1 shows the proposed system diagram. At the first stage of the system, the input audio signal is aligned with

[3] https://sites.google.com/site/pclipatty/scream-mac-emt-dataset

[4] For more information, see http://www.musictheory.org.uk/res-musical-terms/italian-musical-terms.php

Name	Abbreviation	Note-level aggregation	Song-level aggregation
Dynamics	D	M, Max, maxPos	M, S, $\mathrm{C_M}$
Duration	ND, 1MD, 2MD, 4MD	—	M, S, $\mathrm{C_M}$
	FPD	—	—, $\mathrm{C_M}$
Vibrato rate	VR	M, S, MΔ, SΔ, Max, Min, Diff	M, S, $\mathrm{C_M}$
Vibrato extent	VE	M, S, MΔ, SΔ, Max, Min, Diff	M, S, $\mathrm{C_M}$
Global vibrato extent	GVE	—	M, S, $\mathrm{C_M}$
Vibrato ratio	vibRatio	—	—

Table 2: Proposed features, the note-level and song-level aggregation methods.

its corresponding MIDI file in order to find the onset and offset positions and the pitch of each note in the audio signal. To do this, we adopt a chromagram-based audio-score alignment algorithm proposed in [23]. The positions of the bar lines are extracted from the MusicXML-formatted score sheets by using an XML parser. [5] Then, to better characterize the attributes of the basic temporal elements (note or bar) of music, frame-level features are aggregated over time to generate note-level or bar-level features according to the desired segmentation. Furthermore, the note-level and bar-level features are aggregated again into a song-level representation, which allows us to map a variable-length sequence into a fixed-size feature vector that can be fed into a classifier. Finally, in the classification stage, we use radial-basis function (RBF) kernel Support Vector Machine (SVM) implemented by LIBSVM [5].

For the feature aggregation process from note-level (or bar-level) to song-level, we consider 3 different ways: (1) taking mean value over all notes in the excerpt (M), (2) taking standard deviation over all notes in the excerpt (S), and (3) taking the contrast of M between the expressive and its corresponding non-expressive version ($\mathrm{C_M}$): $\mathrm{C_M} = \mathrm{M_{expressive}}/\mathrm{M_{None}}$. $\mathrm{C_M}$ here is designed to "calibrate" the effect of *None*, which can be regarded as a baseline for the other 10 expressive musical terms. That is, $\mathrm{C_M}$ can somehow tell how different the expressive feature is from its non-expressive version. For the feature aggregation methods from frame-level to note-level, we will introduce them separately since they are different for each feature.

We introduce below the baseline feature set and the proposed feature set.

3.1 Baseline Features

The baseline features are a rich set of audio features covering dynamics, rhythm, tonal, and timbre. In particular, the baseline features are a rich set of temporal, spectral, cepstral and harmonic descriptors. It contains the mean and standard deviation of spectral centroid, brightness, spread, skewness, kurtosis, roll-off, entropy, irregularity, flatness, roughness, inharmonicity, flux, zero-crossing rate, low energy ratio, attack time, attack slope, dynamics and the mean and standard deviation of first-order temporal difference for all the above features, totaling 4×17=68 features. Besides, it involves the mean of fluctuation peak and centroid, tempo, pulse clarity and event density, generating 5 features; the mean and standard deviation of mode and key clarity, resulting 4 features. Furthermore, it includes the mean and standard deviation of the 40-D MFCCs, ΔMFCCs (first-order temporal difference) and ΔΔMFCCs (second-order temporal difference), totaling 2×120=240 features. In sum, we have 317 features extracted by the MIRtoolbox (version 1.3.4) [13].

[5] For more details about MusicXML, please refer to `http://www.musicxml.com/`

3.2 Proposed Features

3.2.1 Dynamic Features

The dynamics of each note is estimated from the short-time Fourier transform (STFT). Given a segmented note $x(n)$ and the Hanning window function $w(n)$, the STFT is represented as $X^w(n,k) = M^w(n,k)\, e^{j\Phi^w(n,k)}$, where $M^w(n,k)$ is the magnitude part, $\Phi^w(n,k)$ is the phase part, n is the time index, and k is the frequency index. The dynamic level function $D(n)$ is computed by the summation of the magnitude spectrogram over the frequency bins and is expressed in dB scale:

$$D(n) = 20\ \log_{10}\left(\sum_k M(n,k)\right). \qquad (1)$$

Three note-level dynamic features are computed from $D(n)$. Each of them are the mean value of $D(n)$ (D-M), the maximal value of $D(n)$ (D-Max) and the proportion of the maximum position to the note length (D-maxPos):

$$\mathrm{maxPos} = \frac{\arg\max_n D(n)}{\mathrm{length}\,(D(k))} \times 100\%\,. \qquad (2)$$

D-maxPos therefore measures the time a note reaches its maximal energy from its beginning, normalized to the length of the note. All of these three note-level features are then aggregated to song-level by M, S, and $\mathrm{C_M}$, totaling 9 features (see the second row of Table 2). For the $D(n)$ calculation, frames of 23ms (1014 samples) with an 82% overlap (832 samples), as used in [14], are adopted.

3.2.2 Duration Features

After score alignment and note segmentation, we take the following values as the features: the duration of every single note (ND), measure (1MD), two-measure segment (2MD), four-measure segment (4MD), and the full piece (FPD) (see the third row of Table 2). We expect that these features can capture the interpretation of local tempo variations measured by single notes, downbeats, and phrases. We take M and S on ND, 1MD, 2MD and 4MD to obtain song-level features. FPD itself is already a song-level feature so no aggregation is needed. Moreover, all of these

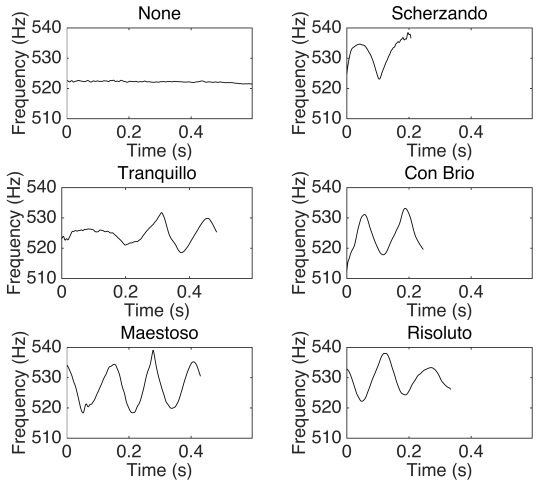

Figure 2: Pitch contours of the first crotchet (C5) of Mozart's *Variationen* with 6 expressions: *None*, *Scherzando*, *Tranquillo*, *Con Brio*, *Maestoso* and *Risoluto*.

five features are processed by $\mathrm{C_M}$. Figure 2 shows examples where the same note (a crotchet C5) is interpreted in different ND for distinct expressions.

There are some more implementation details about the duration features. If a music piece has an incomplete measure in the beginning (e.g., Vivaldi's *La primavera Mov. I*) then the incomplete measure is merged into the next one and features are computed starting from the first complete measure. If the length of a phrase is not the multiple of the 2 or 4 measures then the remainders are combined as a group. Bruch's *Violin Concerto No. 1 Mov. I* (the 5^{th} piece) is an unusual instance that has two *ad libitum* measures. In this case, 4MD is set at zero. In the parser process, a special part is to eliminate rests and ties because they do not have a unique sound. The former means an interval of silence and the latter has a curved line connecting to its previous note of the same pitch, indicating that they should be played as a single note.

3.2.3 Vibrato Features

Vibrato is an expressive manipulation of pitch corresponding to a frequency modulation of F0 (fundamental frequency) [17]. Because the vibrato is characterized by the rate and extent of the frequency modulation of F0, a precise estimation of the instantaneous pitch contour is needed. Since the frequency resolution in the STFT representation may not be high enough to represent the instantaneous frequency, we compute the instantaneous frequency deviation (IFD) [11] to estimate the instantaneous frequency:

$$\mathrm{IFD}^{w}(n,k) = \frac{\partial \Phi^{w}}{\partial t} = \mathrm{Im}\left(\frac{X^{D^{w}}(n,k)}{X^{w}(n,k)}\right), \quad (3)$$

where $D^{w}(n) = w'(n)$. Given the pitch of each note from the score, instantaneous frequency is computed by summing the IFD and the bin frequency of the bin which is nearest to the pitch frequency. Figure 2 also sketches examples of the vibrato contours. We can see large differences in both duration and vibrato among them. For the IFD calculation, a window of 1025 samples at 44.1 kHz sampling rate and a hop size of 64 samples are applied.

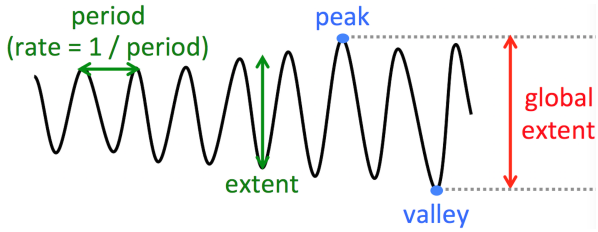

Figure 3: Illustration of vibrato rate, vibrato extent, and global vibrato extent in a single note.

After obtaining the vibrato contour of each note, we adopt a moving-average filter with length of one-hundredth of the note length to reduce the spurious variation of the pitch contour. The filter length is empirically set so as not to avoid much distortion and to remove high-frequent noise. Based on the smoothed pitch contour, we consider the *vibrato rate* (VR) and the *vibrato extent* (VE). The former means the reciprocal of the time duration of two consecutive peaks, while the latter means the frequency deviation between a peak and its nearby valley. Following [8], we require that a vibrato chain contains more than 3 points and VR is between 3 and 12 Hz; otherwise, the vibrato chain is excluded. For each note, we compute the mean, standard variation, mean of difference (MΔ), standard variation of difference (SΔ), maximum (Max), minimum (Min) and difference (Diff) between the maximal and minimal values of both VR and VE over all frames within a note [24]. These note-level features are also aggregated to song-level features by means of M, S, and $\mathrm{C_M}$.

In addition, we consider a note-level feature called global vibrato extent (GVE), meaning the difference of the maximal peak value and the minimal valley value within a vibrato note as shown in Figure 3. GVE is also aggregated to song-level features through M, S, and $\mathrm{C_M}$. Finally, we consider a song-level feature called vibrato ratio (vibRatio), defined as:

$$\mathrm{vibRatio} = \frac{\#\ \text{vibrato notes}}{\#\ \text{notes in a violin piece}} \times 100\%. \quad (4)$$

When no vibrato note is detected or the ND is shorter than 125ms [14], the vibrato features are set at zero.

3.3 Feature Selection and Classification

To evaluate the importance of the adopted features in our task, we perform feature selection on both the baseline and the proposed feature sets. Here, the ReliefF routine of the MATLAB statistics toolbox [6] is employed in the feature selection process [22]. In the training process, ReliefF sorts the features in descending order of relevance (importance). Then, the top-n' most relevant features are taken for SVM modeling. The optimal feature number n_{opt} which results in the best accuracy is obtained by brute-force searching.

[6] http://www.mathworks.com/products/

Baseline	n	n_{opt}	c	γ	ACC
	317	107	1	2^{-8}	0.473

Proposed	without C_M					with C_M				
	n	n_{opt}	c	γ	ACC	n	n_{opt}	c	γ	ACC
Dynamics	6	6	16	2^{-6}	0.318	9	9	16	2^{-8}	0.398
Duration	9	8	16	2^{-4}	0.331	14	13	64	2^{-6}	0.380
Vibrato	31	6	256	2^{-6}	0.264	46	15	64	2^{-6}	0.258
All	46	22	4	2^{-6}	0.425	69	36	1	2^{-6}	**0.531**
Fusion	363	148	1	2^{-8}	0.498	386	68	1	2^{-6}	**0.589**

Table 3: Performance of the baseline and the proposed feature sets. 'All' indicates the combination of dynamics, duration and vibrato; 'fusion' represents the combination of baseline and 'all.' n and n_{opt} are the original and the optimized number of features respectively; c and γ are SVM parameters; ACC indicates the average accuracy.

The RBF-kernel SVM is adopted for classification. Since the dataset is recorded by 11 violinists, we simply take 11-fold cross validation, by using the data of 10 violinists as the training set and the other as the testing set. Then the feature selection is performed in each fold of the cross validation individually. After sending the top-n_{opt} most relevant features into classification, the resulting performance is obtained from optimizing the parameters c and γ of the SVM. In this work, the SVM parameters are set according to the highest average classification accuracy across the 11 folds. In the future, we will consider other data splitting settings, for example using an independent held-out set for parameter tuning.

In our classification experiment, we exclude the case of *None* and consider a 10-class multi-class classification problem, because the calculation of C_M aggregation method requires that the non-expressive version is known *a priori*. The classification accuracy of random guess is 0.152 on average.

As we want to find out the relevant interpretational factors, we only report below the results obtained by the top-n_{opt} relevant features selected by ReliefF.

4. EXPERIMENT RESULTS

4.1 Overview

Table 3 lists the original feature number n, the optimal feature number n_{opt}, the average accuracy (the ratio of true positives and the number of data) computed over the 11 folds, and the corresponding optimal c and γ for each experimental setting. The upper part of the table shows the result of the baseline feature set, where ReliefF selects $n_{\text{opt}} = 107$ out of 317 features and achieves an accuracy of 0.473. From the lower part of the table, when C_M aggregation method is considered, the proposed feature set achieves an accuracy of 0.531 when choosing $n_{\text{opt}} = 36$ out of 69 features, showing a significant improvement from the baseline feature set as validated by a one-tailed t-test ($p<0.05$, d.f.=20). Finally, after fusing the baseline features and all the proposed features, the average accuracy comes to 0.589, using $n_{\text{opt}} = 68$ out of 386 features.

#	Baseline	Proposed ('all')	Fusion
1	24th MFCC-M	4MD-C_M	vibRatio
2	18th MFCC-S	vibRatio	24th MFCC-M
3	26th MFCC-M	D-Max-C_M	ND-M
4	31st MFCC-M	D-M-C_M	18th MFCC-S
5	25th MFCC-S	FPD-C_M	VR-Min-M
6	15th MFCC-S	ND-M	31st MFCC-M
7	21st MFCC-S	D-maxPos-M	26th MFCC-M
8	31st MFCC-S	VR-Min-M	4MD-C_M
9	9th MFCC-S	1MD-C_M	25th MFCC-S
10	entropy-SΔ	2MD-C_M	9th MFCC-S
11	17th MFCC-S	FPD	FPD-C_M
12	24th MFCC-S	2MD-M	24th MFCC-S
13	16th MFCC-S	1MD-M	15th MFCC-S
14	23rd MFCC-M	ND-C_M	23rd MFCC-M
15	22nd MFCC-S	VR-M-M	31st MFCC-S
16	15th MFCC-M	4MD-S	D-maxPos-M
17	30th MFCC-M	4MD-M	16th MFCC-S
18	10th MFCC-S	D-maxPos-S	21st MFCC-S
19	16th MFCC-M	D-maxPos-C_M	10th MFCC-S
20	29th MFCC-S	D-M-M	entropy-SΔ

Table 4: The first 20 ranked features of the feature sets.

4.2 The contrast value C_M

Table 3 also shows how important using the contrast between the expressive and non-expressive version improves the performance. Comparing the left-hand side (without C_M) and the right-hand side (with C_M) of the table, using C_M constantly improves the average accuracy. Salient improvement can be observed for dynamic features ($p<0.05$) and duration features ($p\simeq0.05$), implying that the change of dynamics, note duration, downbeat or phrase might be important interpretation factors when comparing the expressive and non-expressive performance. The improvement is not significant for vibrato ($p>0.5$), possibly because the ratio of strong vibrato (expressive) and "almost no vibrato" (non-expressive) is not a stable feature. Table 3 also shows that using C_M on the proposed ('all') and the fusion feature sets leads to significant improvement for both cases ($p<0.005$). Taking the contrast of feature values between expressive and non-expressive performance seems to be critical in modeling musical expression.

4.3 Feature importance analysis

Table 4 lists the top-20 relevant features for the baseline, proposed ('all') and fusion feature sets. The list is generated by summing the rank of each feature over the results of 11 folds, and by sorting the summarized rank again.

From the leftmost column, we can see that most of the relevant features in the baseline set are MFCCs. Despite its accuracy is inferior to the proposed features, this result shows the generality of MFCCs in audio classification.

From the middle column, we see that the top-20 proposed features include 11 duration features, 6 dynamic ones, and 3 vibrato ones. Over half of them are duration features. However, we note that the second feature is about vibrato (vibRatio) and the next two are both dynamic features (D-Max-C_M and D-M-C_M). It is not trivial to conclude that which factor is the most relevant. Dynamics, duration and vibrato all have contribution on music inter-

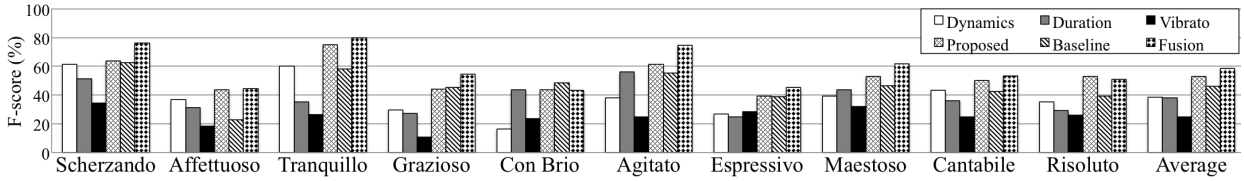

Figure 4: Average accuracy (in F-scores) of expression classification using individual feature sets.

		predicted class										F-score
		Sc	*Af*	*Tr*	*Gr*	*Co*	*Ag*	*Es*	*Ma*	*Ca*	*Ri*	
actual class	*Scherzando*	**45**	0	2	1	3	2	0	0	0	2	0.763
	Affettuoso	1	**22**	3	4	3	3	4	4	8	3	0.444
	Tranquillo	1	1	**46**	3	0	0	0	1	3	0	0.800
	Grazioso	2	3	2	**31**	4	0	6	3	4	0	0.544
	Con Brio	8	4	0	0	**25**	1	2	0	3	12	0.431
	Agitato	0	1	0	1	2	**43**	4	4	0	0	0.748
	Espressivo	1	0	1	9	2	5	**23**	6	5	3	0.451
	Maestoso	0	2	3	4	1	5	4	**34**	1	1	0.618
	Cantabile	1	9	3	5	2	1	4	1	**29**	0	0.532
	Risoluto	4	2	0	1	19	0	0	2	1	**26**	0.510

Table 5: Confusion matrix of musical term classification using the fusion feature set.

pretation in various perspective. What this list provides is the signal-level details useful in synthesizing expressive music, or in education software for teaching expressive performance, something that cannot be achieved using the baseline features such as the MFCCs.

Finally, from the rightmost column we find that the top-20 fusion features contain a blending of the baseline and the proposed features. This shows that the two feature sets are indeed complementary, and it is advisable to exploit both of them if classification accuracy is of major concern.

4.4 Class-wise performance

Table 5 illustrates the confusion matrix of the fusion feature set summarized over the 11-fold outputs. In this confusion matrix, rows correspond to the actual class and columns correspond to the predicted class. The column on the right of the confusion matrix lists the average F-score, which is the harmonic mean of precision and recall.

We see that *Scherzando*, *Tranquillo* and *Agitato* attain relatively high F-scores because the first two have lighter dynamics than other expressions and the last one has shorter duration in most cases, all are fairly easy to be recognized. Interestingly, the low F-scores and the high confusion between the two pairs *Affettuoso*/*Grazioso* and *Cantabile*/*Espressivo* clearly reveal their semantic similarity. The most serious confusion occurs between *Risoluto* and *Con Brio*, perhaps due to their similar tempo and dynamics; the slightly difference of vibrato extent between them is not discriminated in our system, unfortunately.

Figure 4 shows the class-wise F-scores obtained by different feature sets. From the first three feature sets, we can find that using dynamic features outperforms other two for six expressions. Using duration features attains the best results for *Agitato*, *Con Brio* and *Maestoso*; the first two tend to use relatively fast tempo and the last one is prone to use a little slow and stable tempo. Lastly, we see that vibrato features perform slightly better than dynamic and duration features for *Espressivo*, possibly because *Espressivo* is similar to *Cantabile* in dynamic features and is similar to *Grazioso* in duration features.

Comparing the baseline to the proposed ('all') feature sets, the baseline feature set performs better only for *Grazioso* and *Con Brio*. The fusion set generally improves F-scores for all expressions except for *Con Brio*. For all settings, the four expressions, *Affettuoso*, *Grazioso*, *Espressivo* and *Cantabile*, are not easy to be distinguished from each other due to their similar meaning.

5. CONCLUSION AND FUTURE WORK

In this study, we have presented a method for analyzing the interpretational factors of expressive musical terms implemented on a new dataset comprising of rich expressive interpretations of violin solos. The proposed features, motivated from the basic understanding of dynamics, duration, vibrato, and the information of score, give better performance than the standard feature set in classifying expressive musical terms. Particularly, the contrast of feature values between expressive and non-expressive performance is found critical in modeling musical expression. The importance of the features is also reported. This provides insights into the design of new expression-based features, which may include features for the possible glissando between two adjacent notes, or the variation of the note/measure duration proportion with respect to its measure/excerpt. For future work, we will consider to expand the dataset, to experiment with other features and machine learning techniques, and to devise a mechanism that does not require a non-expressive reference to compute the contrast values.

6. ACKNOWLEDGMENTS

The authors would like to thank the following three professional violinists for consulting: Chia-Ling Lin (Doctor of Musical Arts, City University of New York; concertmaster of Counterpoint Ensemble), Liang-Chun Chou (Master of Music, Manhattan School of Music; concertmaster of Tainan Symphony Orchestra), and Hsin-Yi Su (Master of Music, New England Conservatory of Music; major violin performance of Tainan Symphony Orchestra). We are also grateful to the 11 professional violinists for their contribution to the development of the new dataset. The paper is partially funded by the Ministry of Science and Technology of Taiwan, under contracts MOST 103-2221-E-006-140-MY3 and 102-2221-E-001-004-MY3, and the Academia Sinica Career Development Award.

7. REFERENCES

[1] J. Abeßer, H. Lukashevich, and G. Schuller. Feature-based extraction of plucking and expression styles of the electric bass guitar. In *ICASSP*, pages 2290–2293, 2010.

[2] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad. Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. *Music Perception*, 28(3):265–278, 2011.

[3] R. Bresin and G. U. Battel. Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the andante movement of mozart's sonata in g major (k 545). *Journal of New Music Research*, 29(3):211–224, 2000.

[4] A. Camurri, G. Volpe, G. De Poli, and M. Leman. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1):43–53, 2005.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[6] J. Charles. *Playing Technique and Violin Timbre: Detecting Bad Playing*. PhD thesis, Dublin Instititute of Technology, 2010.

[7] R. L. De Mantaras. Playing with cases: Rendering expressive music with case-based reasoning. *AI Magazine*, 33(4):22, 2012.

[8] A. Friberg, E. Schoonderwaldt, and P. N. Juslin. CUEX: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals. *Acta acustica united with acustica*, 93(3):411–420, 2007.

[9] R. Gang, G. Bocko, J. Lundberg, S. Roessner, D. Headlam, and M. F. Bocko. A real-time signal processing framework of musical expressive feature extraction using MATLAB. In *ISMIR*, pages 115–120, 2011.

[10] M. Grachten and G. Widmer. Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research*, 41(4):311–322, 2012.

[11] S. Hainsworth and M. Macleod. Time frequency reassignment: A review and analysis. Technical report, Cambridge University Engineering Department, 2003.

[12] N. Kroher and E. Gómez. Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors. In *ICMC-SMC*, 2014.

[13] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *DAFx*, 2007.

[14] M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre. The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43(3):303–317, 2014.

[15] M. Molina-Solana, J. L. Arcos, and E. Gómez. Using expressive trends for identifying violin performers. In *ISMIR*, pages 495–500, 2008.

[16] K. Okumura, S. Sako, and T. Kitamura. Stochastic modeling of a musical performance with expressive representations from the musical score. In *ISMIR*, pages 531–536, 2011.

[17] E. Prame. Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America*, (102):616–621, 1997.

[18] R. Ramirez, E. Maestre, and X. Serra. A rule-based evolutionary approach to music performance modeling. *Evolutionary Computation, IEEE Transactions on*, 16(1):96–107, 2012.

[19] C. Raphael. Representation and synthesis of melodic expression. In *IJCAI*, pages 1474–1480, 2009.

[20] A. Rodà, S. Canazza, and G. De Poli. Clustering affective qualities of classical music: beyond the valence-arousal plane. *Affective Computing, IEEE Transactions on*, 5(4):364–376, 2014.

[21] S. Vieillard, M. Roy, and I. Peretz. Expressiveness in musical emotions. *Psychological research*, 76(5):641–653, 2012.

[22] M. R. Šikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53:23–69, 2003.

[23] T.-M. Wang, P.-Y. Tsai, and A. W. Y. Su. Note-based alignment using score-driven non-negative matrix factorisation for audio recordings. *IET Signal Processing*, 8:1–9, February 2014.

[24] L. Yang, E. Chew, and K. Z. Rajab. Vibrato performance style: A case study comparing erhu and violin. In *CMMR*, 2013.

[25] Y.-H. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Trans. Audio, Speech & Lang. Processing*, 19(4):762–774, 2011.