AUTOMATIC VIOLIN SYNTHESIS USING EXPRESSIVE MUSICAL TERM FEATURES

Chih-Hong Yang, Pei-Ching Li, Alvin W. Y. Su

Scream Lab., Department of CSIE,
National Cheng-Kung University,
Tainan Taiwan
P76034193@mail.ncku.edu.tw,
P78021015@mail.ncku.edu.tw,
alvinsu@mail.ncku.edu.tw

Li Su, Yi-Hsuan Yang

MAC Lab., Research Center for Information Technology
Innovation, Academia Sinica,
Taipei Taiwan
lisu@citi.sinica.edu.tw
yang@citi.sinica.edu.tw

ABSTRACT

The control of interpretational properties such as duration, vibrato, and dynamics is important in music performance. Musicians continuously manipulate such properties to achieve different expressive intentions. This paper presents a synthesis system that automatically converts a mechanical, deadpan interpretation to distinct expressions by controlling these expressive factors. Extending from a prior work on expressive musical term analysis, we derive a subset of essential features as the control parameters, such as the relative time position of the energy peak in a note and the mean temporal length of the notes. An algorithm is proposed to manipulate the energy contour (i.e. for dynamics) of a note. The intended expressions of the synthesized sounds are evaluated in terms of the ability of the machine model developed in the prior work. Ten musical expressions such as *Risoluto* and *Maestoso* are considered, and the evaluation is done using held-out music pieces. Our evaluations show that it is easier for the machine to recognize the expressions of the synthetic version, comparing to those of the real recordings of an amateur student. While a listening test is under construction as a next step for further performance validation, this work represents to our best knowledge a first attempt to build and quantitatively evaluate a system for EMT analysis/synthesis.

1. INTRODUCTION

Expression plays an important role in music performance. For the same musical score, different performers would interpret the score with their personal understandings and experiences and instill their feelings and emotions into it, thereby creating large variations in their actual performances. These variations can be observed in interpretational properties like timing, modulation, and amplitude. Therefore, in automatic music synthesis, an important step is to characterize and to control such expressive parameters.

Expressive music performance has been studied in the last few decades [1, 2, 3, 4, 5, 6]. For example, Bresin *et al.* [7] synthesized music of six different emotions by using performance rules such as duration contrast, punctuation, and phrase arch. Maestre *et al.* [8] characterized dynamics and articulation parameters related to the expressivity of saxophone. D'Incà *et al.* [9] considered four sensorial adjectives (*hard*, *soft*, *heavy*, and *light*) and four affective adjectives (*happy*, *sad*, *angry*, and *calm*) based on a set of audio cues. Grachten *et al.* [10] used both predictive and explanatory framework to model three categories of dynamics markings in piano. Erkut *et al.* [11] captured information of guitar performance such as damping regimes and different pluck styles used in a plucked-string synthesis model. More recently, Perez *et al.* [12] combined the modeling of the characteristics of the performer (i.e., pitch, tempo, timbre, and energy), the sound as well as the instrument in order to render natural performances from a musical score.

Surprisingly, among all the elements of expressive synthesis, the *expressive musical terms* (EMT) that describe feelings, emotions, or metaphors in a piece of music have been rarely discussed, even though they have been widely used in Western classical music for hundreds of years. To fill this gap, in a prior work [13] we presented a computational analysis of ten EMTs — including *Tranquillo* (calm), *Grazioso* (graceful), *Scherzando* (playful), *Risoluto* (rigid), *Maestoso* (majestic), *Affettuoso* (affectionate), *Espressivo* (expressive), *Agitato* (agitated), *Con Brio* (bright), and *Cantabile* (like singing) — using a new violin solo dataset called SCREAM-MAC-EMT. [1] The dataset contains ten classical music pieces, with each piece being interpreted in six versions (five EMTs and one mechanical, deadpan version denoted as *None*) by eleven professional violinists, totaling 660 excerpts [13]. With this dataset, we built supervised machine learning models to recognize the EMT of a music excerpt from audio features. We compared the performance of two types of features for the classification task. The first type of features includes a set of audio features characterizing the three interpretational factors, dynamics, duration, and vibrato, whereas the second type of features are standard timbre, rhythm, tonal, and dynamics features such as Mel-frequency cepstral coefficients (MFCC) extracted from the MIRtoolbox [14]. Our evaluation shows that the first feature set, which has clearer music meanings, achieves better classification accuracy than the standard features do, showing the importance of these interpretational features in characterizing EMTs.

Extending from this prior work, we investigate in this paper the use of a small set of such interpretational features for synthesizing music of different EMTs. Specifically, we aim to manipulate features of vibrato, dynamics, and duration to synthesize expressive sounds from a mechanical interpretation. The way of manipulation is learned from a training set SCREAM-MAC-EMT. To quantitatively evaluate the performance of the proposed expressive synthesis method, we make use of the classification model developed in our prior work [13] again to see if the intended EMT of the synthesizer can be correctly recognized. Specifically, we recruit a professional violinist and an amateur student to record new data in accordance with the collection method of the SCREAM-MAC-EMT dataset. That is, both of them perform the sixty classical excerpts (i.e. six different versions of the ten pieces) individually. Then, we compare the performance of the real and synthetic versions by means of the same EMT analysis and classification procedure. In other words, the objective evaluation of the ten EMTs

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

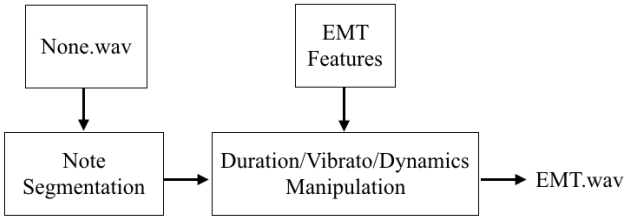

Figure 1: *Flowchart of the EMT synthesis system.*

recognition is applied to these outside data through the classification models constructed from the preliminary work [13]. Therefore, we can observe not only the differences between the two people who have distinct musical trainings and skills from their violin performances, but also the result of synthesis based on the two unique sources. The synthesized sound samples can be found online. [2] This paper is organized as follows. Section 2 describes the expressive synthesis method, together with the EMT features, and the setting of classification. In Section 3, the experimental results are presented. Finally, we conclude in Section 4.

2. METHOD

Figure 1 shows the EMT synthesis diagram, whose goal is to convert a mechanical interpretation of a music piece into an expressive one. We refer to the mechanical interpretation as the "*None*" signal. As the manipulation process is usually done for each note, at the first stage note segmentation is applied to the *None* audio file. Then the manipulations of duration, vibrato, and dynamics of each segmented notes are performed according to the pre-learned parameters of the target EMT (see Sections 2.1–2.3 for details). Lastly, when concatenating all the manipulated notes back into a complete music piece, we adopt fade-in and fade-out operations to eliminate the crackled sounds.

To synthesize expressive sounds, the parameter values of the ten EMTs are calculated by averaging over the corresponding music pieces because each EMT is interpreted in five different excerpts. Moreover, as we have the performance from eleven musicians for each music piece and each EMT, the parameters will be averaged again across the violinists. The *EMT feature set* listed in the Table 1 is used in the proposed synthesis system. It includes seven relevant features, namely `vibRatio`, `ND-`C_M, `4MD-`C_M, `FPD-`C_M, `D-M-`C_M, `D-Max-`C_M, and `D-maxPos-M`, as well as two fundamental features of vibrato, `VR-M-M` and `VE-M-M`. The first seven features are found to be more important than other possible interpretational features for classifying EMTs [13]. The last two features are found less useful in classifying EMTs, but they are still needed to manipulate added vibrato to a note.

The manipulations of duration and vibrato are implemented by means of the phase vocoder, which is a mature technique of time stretching and pitch shifting [15, 16, 17]. Given an audio input, the short-time Fourier transform (STFT) converts the signal from time domain into a time-frequency representation. The *time stretching* (expansion/compression) is achieved by modifying the hop size and then by performing the inverse STFT with the overlap-add method. The *pitch shifting* is accomplished by resampling the time stretched signal back to the original length. More details of the synthesis method, together with the meaning of the features listed in the EMT feature set, and the setting of EMT classification, are introduced in the following sections.

In what follows, we assume that all the audio excerpts are sampled at 44.1 kHz.

[2] `http://screamlab-ncku-2008.blogspot.tw/2016/03/music-files-of-expressive-musical-term-experiment.html`

Table 1: *The EMT feature set used in the synthesis system. The terms 'M' and '*C_M*' denote mean and contrast of mean, respectively. Please refer to [13] for details.*

Features	Abbreviation	Description
Vibrato	`vibRatio`	percentage of vibrato notes in a music piece
	`VR-M-M`	mean vibrato rate
	`VE-M-M`	mean vibrato extent
Dynamics	`D-M-`C_M	mean energy
	`D-Max-`C_M	maximal energy
	`D-maxPos-M`	relative time position of the energy peak in a note
Duration	`ND-`C_M	mean length of every single note
	`4MD-`C_M	mean length of a four-measure segment
	`FPD-`C_M	mean length of a full music piece

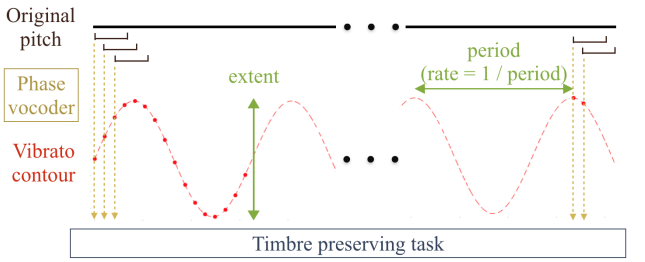

Figure 2: *Illustration of adding vibrato to a non-expressive note. The note is divided into a sequence of fragments whose pitches will be individually shifted by means of the phase vocoder. The timbre preservation is applied to each fragment. The vibrato contour is sampled at sixteen times per cycle to avoid artifacts.*

2.1. Vibrato Features

Vibrato is an essential factor in violin performance and its analysis/synthesis has been studied for decades [18, 19]. Vibrato is defined as a frequency modulation of F0 (fundamental frequency) and is typically characterized by the rate and extent [20]. The vibrato *rate* means the number of periodic oscillations per second while the vibrato *extent* specifies the amount of frequency deviation. In the EMT feature set, `VR-M-M`, `VE-M-M` and `vibRatio` are related to vibrato. The first two are defined as the mean value of the vibrato rate and extent, and the last one means the ratio of the number of vibrato notes over total notes in a music piece. The detailed criteria of determining whether a note is vibrato could be found in [13]. Vibrato is a common interpretation in violin performance, but the `VR-M-M` and `VE-M-M` are found to have weak discrimination power in classifying EMTs [13], possibly due to their subtle difference. The mean values of the two features among the

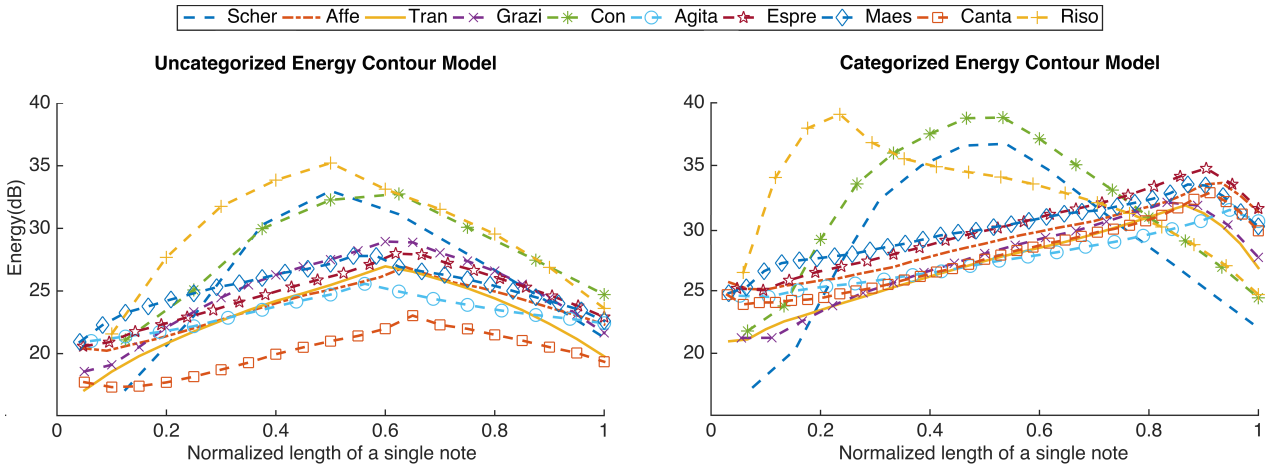

Figure 3: *The energy contours are modeled by means of the EMT parameter (UEC; left) and the categorized method (CEC; right).*

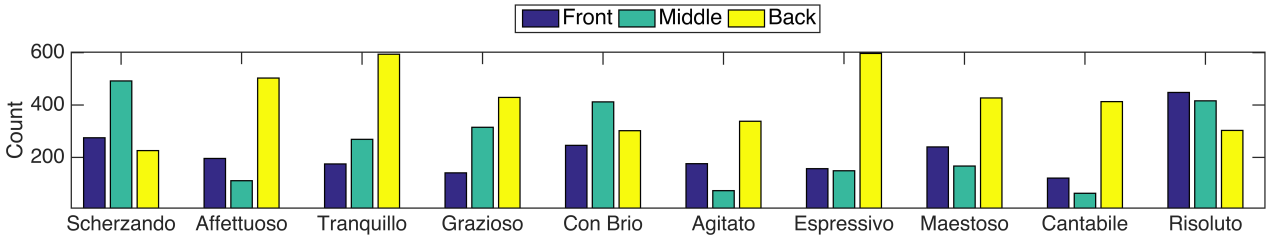

Figure 4: *The number of notes in the three categories of the ten EMTs*

ten expressive musical terms are between 6.3–6.8 Hz (STD=0.15) and 0.27–0.38 semitones (STD=0.03) separately. In contrast, the `vibRatio` has strong discrimination power and its mean values are between 52–74% (STD=6.82) among the ten expressions. Assuming that there are no vibrato at all in the *None* signal, to implement `vibRatio` we need to determine how many vibrato notes there should be, and which notes should be manipulated to have vibrato. Firstly, the amount of vibrato notes is easy to calculate and is expressed by:

$$\#\,\text{Vibrato notes} = \#\,\text{notes in a violin piece} \times \text{vibRatio}\,, \tag{1}$$

where the value of vibRatio is set differently for different EMTs and it's set according to its average value in the training set (i.e. SCREAM-MAC-EMT) per EMT. Secondly, according to our observation, a note with longer duration will more likely have vibrato. Hence, we sort all the notes in descending order of duration and add vibrato to the top longest ones (the exact number of notes is determined by equation (1)).

Moreover, we remark that the continuity of the pitch contour is important to obtain a naturally synthesized vibrato. Therefore, we use a sequence of short fragments to model the modulation of the original frequency of a non-expressive note. Specifically, we shift the pitch of each fragment through the phase vocoder to fit a vibrato contour. For the purpose of avoiding weird artifacts, we sample at sixteen times per cycle of a vibrato contour so that the sampling period is approximately 2.4 milliseconds (410 samples). Accordingly, the first step of the vibrato manipulation process shown in Figure 2 is that partitioning a non-expressive note into a sequence of fragments of 2,048 samples with an 80% overlap (1,638 samples). Next, given a fragment and its corresponding pitch on the particular vibrato contour generated by the `VR-M-M` and `VE-M-M`, the pitch shifting is carried out with a Hanning window of 256 samples as well as a hop size of 64 samples. Then, the timbre preserving method proposed by Röbel and Rodet [21] is adopted. According to this method, both the spectral envelope, measured by a true envelope estimator, and the pre-warping factor of the original fragment are calculated before the pitch shifting takes place. The timbre preservation is therefore realized by means of the multiplication of the pre-warping factor and the pitch-shifted fragment. Finally, we overlap and add the fragments to achieve the synthesized vibrato note.

2.2. Dynamic Features

One of the most prominent characteristics to distinguish expression is dynamics. According to Gabrielsson and Juslin [22], the dynamics and the temporal envelopes of individual notes are different for distinct expressions. The EMT feature set has three dynamic features, `D-M-C`$_M$, `D-Max-C`$_M$ and `D-maxPos-M`, which indicate the mean energy, the maximal energy, and the relative time position of the maximal energy peak in a note (denoted as *maxPos*), separately. To utilize these features for synthesis, we need to know the dynamic envelops of the ten EMTs ahead. However, the specific envelopes are still unknown within these features so we need to model the energy contour, which characterizes the instantaneous energy as a function of time. To make the energy contour as close to the real acoustic envelope as possible, we consider the data of the three consultants, who helped us in the cre-

ation of SCREAM-MAC-EMT [13], and the dynamic level function, which is calculated by summing the spectrum over the frequency bins and expressing in dB scale with frames of 1,024 samples at increments of 256 samples [13]. According to the ways of deciding the values of *maxPos* for each EMT, we implement two types of energy contour model: one is directly using the parameter, `D-maxPos-M`, in the EMT feature set (denoted as *UEC*), while the other is based on a categorized method (denoted as *CEC*).

The UEC modeling of dynamics is implemented as follows:

STEP 1 Calculating the dynamic levels of all the notes among the five excerpts corresponding to a particular EMT across the three consultants (15 excerpts in total).

STEP 2 Resampling all the dynamic levels so that the values of *maxPos* are equal to the `D-maxPos-M` parameter.

STEP 3 Averaging the whole dynamic levels.

The *UEC* model of the ten EMTs is shown in the left side of the Figure 3. We see that *Scherzando*, *Con Brio*, and *Risoluto* have relatively large variation of the energy contours, while the remaining ones have relatively flat ones. Besides, we observe that the values of *maxPos* for all EMTs lie in the interval of 40–70%. However, this phenomenon is unfortunately not consistent with our observation, as the maximal energy would not always lie in the middle of a note. Some notes have strong attacks and others have maximal energy in the back even within a music piece with a particular EMT. The `D-maxPos-M` falling into the middle portion is probably due to the fact that we have taken average on all the notes in the dataset. This motivates us to take the following alternative model.

The CEC modeling of dynamics classifies the notes into three categories among the 15 excerpts of each EMT:

$$\text{note} \in \begin{cases} \text{Front}, & \text{if } \textit{maxPos} < 0.33 \\ \text{Middle}, & \text{if } 0.33 \leq \textit{maxPos} < 0.66 \\ \text{Back}, & \text{otherwise.} \end{cases} \tag{2}$$

After doing this, we count the number of notes for each category. Certainly, as seen in Figure 4, the dominant one of each EMT is different. We simply select the relative category of each EMT, i.e., discarding the remaining ones, to construct a new energy contour model. For example, as most of notes are classified into the back category in the *Tranquillo* case, we take such notes to modeling its own energy contour.

Accordingly, the *CEC* model is realized as follows:

STEP 1 Computing the amount of notes in the front/middle/back category of the ten EMTs using the equation (2).

STEP 2 Selecting the relative majority category of each EMT and taking the notes belonging to this particular group for the dynamic level calculation.

STEP 3 Repeating the three steps of the *UEC* model.

The right-hand side of the Figure 3 shows the estimation of energy contours for each EMT based on the *CEC* model. We notice that *Risoluto* has a strong attack, and *Scherzando* as well as *Con Brio* still has a maximal energy in the middle. Besides, the others have slowly increasing curves which reach the highest energy in the end of a note. The performance of these two models will be evaluated in the classification experiment. Ultimately, we carried out the dynamic manipulation by means of applying a particular energy contour to each note, together with the multiplication of the mean/maximal energy of every note and the parameter, `D-M-`$\mathrm{C_M}$/`D-Max-`$\mathrm{C_M}$.

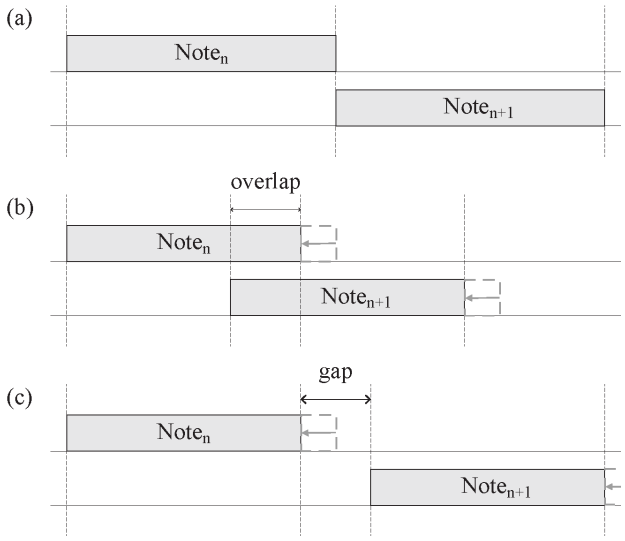

Figure 5: *Illustration of time shortening: (a) The original duration of two consecutive notes, (b) The shortened Note$_n$ followed by an overlapping Note$_{n+1}$, (c) A silent gap between the shortened Note$_n$ and Note$_{n+1}$ if ND–C_M < FPD–C_M.*

2.3. Duration Features

The deviation of timing is also an important expressive factor used by performers [8]. In the EMT feature set, we use `4MD-`$\mathrm{C_M}$, `FPD-`$\mathrm{C_M}$ and `ND-`$\mathrm{C_M}$, defined as the mean length of a four-measure segment, of a full music piece, and of every single note, respectively. Firstly, we stretch a non-expressive note through the phase vocoder and the time-scaling factor is according to the parameter, `ND-`$\mathrm{C_M}$ for each EMT. The length of synthesized note is described as follows:

$$\mathrm{ND_{Synthesis}} = \mathrm{ND_{None}} \times \mathrm{ND\text{–}C_M}\,. \tag{3}$$

Next, we take the `FPD-`$\mathrm{C_M}$ into account and calculate the reasonable onset position of stretched note by the following equation:

$$\mathrm{Onset_{Synthesis}} = \mathrm{Onset_{None}} \times \mathrm{FPD\text{–}C_M}\,. \tag{4}$$

In general, there is an overlap between two consecutive notes in the time shrinking case. However, an abrupt and silent gap may occur in some expressions such as *Tranquillo* if `ND-`$\mathrm{C_M}$ < `FPD-`$\mathrm{C_M}$. This is illustrated in Figure 5. In such a case, the synthesized tone can not keep the temporal continuity of sound. To address this issue, the `FPD-`$\mathrm{C_M}$ will be set equal to the `ND-`$\mathrm{C_M}$ in such condition. Moreover, we stretch every four-measure segment according to the value of `4MD-`$\mathrm{C_M}$ for each EMT. A Hann sliding window of 1,024 samples and a fine hop size of 100 samples are adopted in the phase vocoder module.

2.4. Classification

To evaluate the performance of the synthesis result, we take advantage of the classification models constructed from the prior work [13] for the machine recognition of the ten EMTs. Specifically, the radial-basis function (RBF) kernel Support Vector Machine (SVM) implemented by LIBSVM [23] is adopted for classification. In the training process, we use SCREAM-MAC-EMT

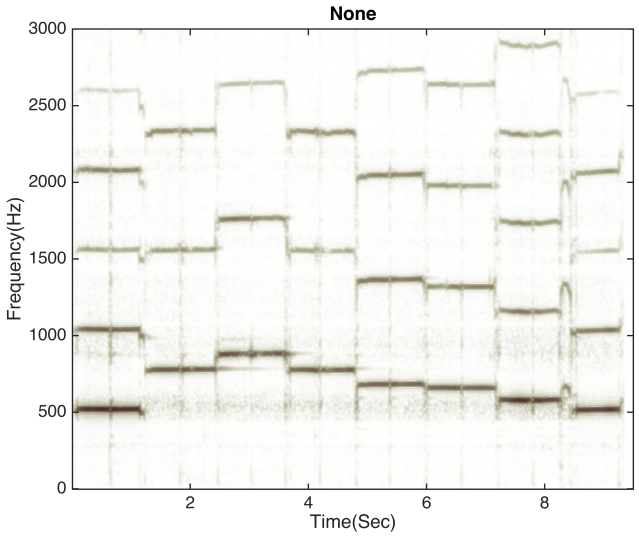

Figure 6: *The first phrase of Mozart's Variationen with a mechanical interpretation performed by an amateur student.*

and take 11-fold cross validation, that is, leave-one-violinist-out in each fold. Besides, the feature selection process is performed by using the ReliefF routine of the MATLAB statistics toolbox [24]. In order to obtain optimized SVM models, the parameters c and γ of the SVM and the top-n' most relevant features are taken based on the highest average accuracy across the 11 folds. In the testing process, each of the outside data, two real recordings collected in this study and six synthetic versions (see Section 3.1 for details), need to be normalized prior to classification. Then the data are fed into the eleven SVM models in conjunction with corresponding relevant features produced in each fold, and the accuracy is computed by averaging over the eleven results. According to [13], the values of c, γ for the SVM classifier and the optimal feature number $n_{\rm opt}$ are set to 1, 2^{-6}, and 36 separately.

3. RESULTS

3.1. Synthesis Results

In this paper, we consider three different sources of the ten non-expressive music pieces, that is, MIDI, amateur student, and professional violinist, in order to observe the differences between the people who have distinct violin trainings and skills. The last two data are recorded in accordance with the collection method of the SCREAM-MAC-EMT. To compare the real recordings with the synthetic versions, both of them perform not only the mechanical interpretation of the ten classical music pieces but also the EMT versions according to the settings of the dataset. In other words, the two people record all the sixty excerpts one by one in a real-world environment. Similarly, to evaluate the proposed synthesis method, each non-expressive excerpt is synthesized in five distinct expressive versions. Moreover, based on the two energy contour models, all the three sources have two types of synthesized sounds. In sum, we have two original and six synthetic data, and each data has sixty excerpts. The following figures, restricting spectrograms from 0 to 3 kHz, illustrate the variations in vibrato, dynamics, and duration. Figure 6 shows an example of mechanical interpretation performed by the amateur, while the three particular expressive versions are demonstrated in Figure 7 which contains original and corresponding synthesized versions in the upper and lower rows respectively. Comparing to the original, we notice that *Scherzando* has a little faster tempo but both *Risoluto* and *Maestoso* have slower one. Besides, all the three synthetic results have more powerful dynamics and stronger vibrato.

Table 2: *The average accuracy compared between the original and synthetic versions which utilize the uncategorized and categorized energy contour models (UEC and CEC, respectively), across three distinct sources.*

Data	MIDI	Amateur	Expert
Original	—	0.293	0.605
UEC	0.578	0.656	0.595
CEC	0.482	0.687	0.615

Table 3: *F-scores of the ten EMTs compared with the original as well as the synthetic version based on the categorized energy contour (CEC) model.*

EMT	MIDI	Amateur		Expert	
	CEC	Original	CEC	Original	CEC
Scherzando	0.878	0.407	0.857	0.653	0.738
Affettuoso	0.317	0.270	0.503	0.436	0.561
Tranquillo	0.923	0.711	0.835	0.838	0.866
Grazioso	0.252	0.250	0.542	0.468	0.516
Con Brio	0.381	0.047	0.770	0.442	0.606
Agitato	0.524	0.397	0.981	0.764	0.922
Espressivo	0.472	NaN	0.231	0.588	NaN
Maestoso	0.483	0.345	0.519	0.815	0.557
Cantabile	0.132	NaN	0.667	0.610	0.459
Risoluto	0.500	0.286	0.855	0.549	0.660

3.2. Classified Results

The objective evaluation of the machine recognition of the ten EMTs is applied to these outside data via the classification models built up from the preliminary work. Hence, the average accuracy predicted by the eleven SVM models among original and synthetic data across the three sources is listed in the Table 2. Additionally, the performances of the two energy contour models are also displayed. Firstly, the MIDI achieves higher classified accuracy when using the *UEC* model. However, all the other synthetic versions have better performance than MIDI. Secondly, the amateur attains an accuracy less than 30% based on the original data but more than 60% among the synthetic ones. There are highly significant differences on both the synthetic data from the original one as validated by a one-tailed t-test (p<0.00001, d.f.=20). In particular, the *CEC* version, using the categorized method to model the dynamic envelopes, achieves the highest accuracy of 0.687, showing a slight improvement from the *UEC* version (p<0.05). Finally, the original data of the expert attain a great performance and the average accuracy comes to 0.605. Besides, both the synthetic versions have nearly the same classified results as the original. In addition, the average F-scores of the ten EMTs comparing between the original and the *CEC* synthetic version are listed in the Table 3. *Espressivo* and *Cantabile* have unrepresentable values, NaN, in the synthetic version of expert and the original version of

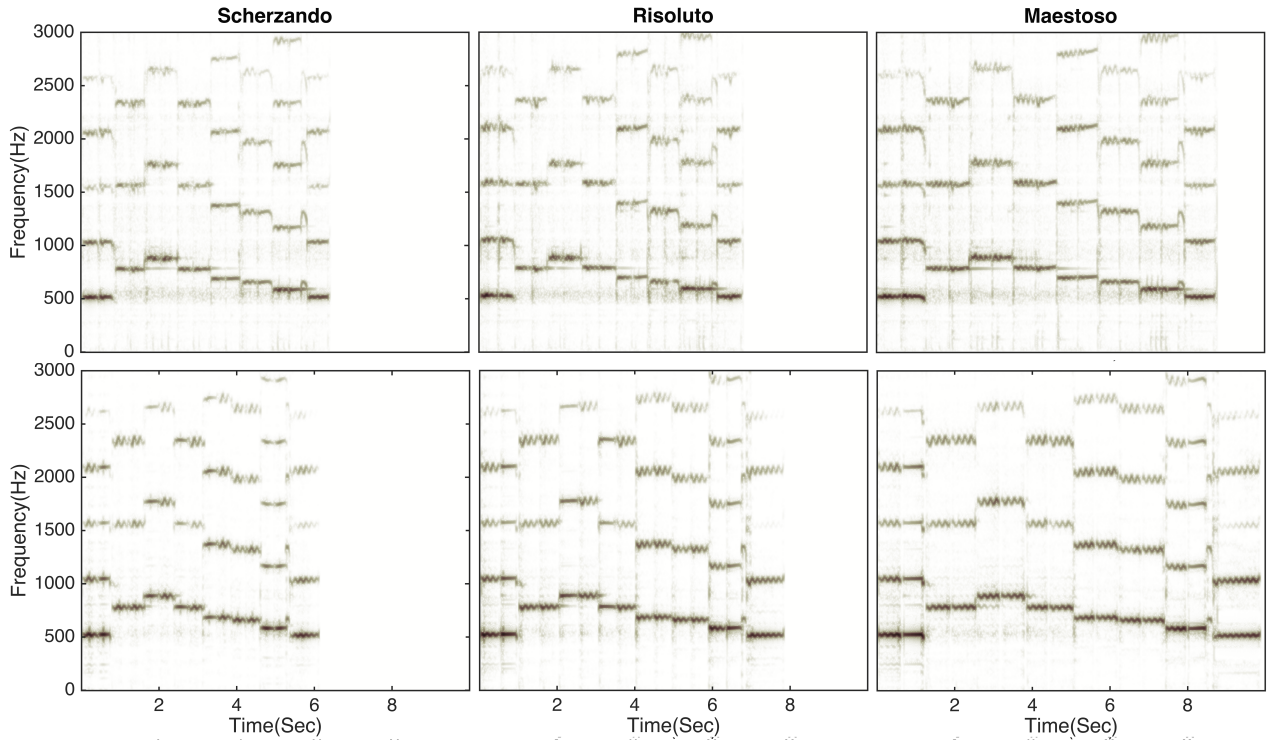

Figure 7: *The first phrase of Mozart's Variationen with three particular EMTs. The upper row shows the original recordings of an amateur while the lower one displays the corresponding synthetic versions based on the specific non-expressive version (see Figure 6).*

amateur respectively because their true positives are zero. Apart from this exception, we find that all the ten EMTs attain higher F-scores in the synthetic version compared with the original of the amateur. Moreover, *Scherzando*, *Tranquillo* and *Agitato* are easily recognized among the five data probably because the first two have lighter dynamics than other EMTs and the last one has faster tempo in most cases.

3.3. Discussion

According to the experimental results, the synthetic data produced by means of the proposed system attain high performances. Specifically speaking, almost all the synthetic versions achieve more than 50% accuracy in the EMT classification task. Particularly, the *CEC* synthetic version of the amateur has significant difference than the original, implying that the virtual simulated violinist is closer to the model of eleven violinists than the amateur. However, the average results are based on the 11 SVM models and corresponding relevant features, which are derived from the 11-fold cross validation from the prior work [13]. We adopt this criterion in order to not only carry on the work but also evaluate the performance of synthesis system via a objective method. In the real application, we will use all the training data to generate a unified model.

Although the synthetic versions obtain great accuracy in classifying the ten EMTs, we could not judge that they have the same expressiveness as the original, or even better than that in the amateur case, by means of the machine recognition. Especially, we only use the nine average features in the synthesis system so both the subtle deviation and the diverse interpretation in violin performance could not be modeled. Hence, the human recognition of expressions is necessary. This work represents an important part of our EMT analysis/synthesis project. A listening test is under construction for it is interesting to know how subjects perform in this regard when original and synthesized sounds are presented. It is expected that such results can be beneficial to this study.

4. CONCLUSION AND FUTURE WORK

In this study, we have presented an automatic system for expressive synthesis from a mechanical sound by means of a small set of interpretational factors derived from the preliminary analysis results. The synthetic data coming from three distinct sources with the dynamic, vibrato, and duration manipulations achieve more than 50% accuracy in the expressive musical term classification task. The performance of two energy contour models is also reported. Specifically, the synthetic versions based on the non-expressive excerpts of an amateur student are closer to the classified models than the original, providing insights into the application of computer-aided music education such as performance calibration in pitch, tempo, and articulation. For future work, we will consider to adopt other features for generating more expressive versions and to conduct a listening test for subjective evaluation.

5. ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Science and Technology of Taiwan for its financial support of this work, under contract MOST 103-2221-E-006-140-MY3.

6. REFERENCES

[1] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad, "Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance," *Music Perception*, vol. 28, no. 3, pp. 265–278, 2011.

[2] A. Friberg, "Digital audio emotions-an overview of computer analysis and synthesis of emotional expression in music," in *Proc. of the 11th International Conference on Digital Audio Effects*, 2008.

[3] G. De Poli, A. Rodà, and A. Vidolin, "Note-by-note analysis of the influence of expressive intentions and musical structure in violin performance," *Journal of New Music Research*, vol. 27, no. 3, pp. 293–321, 1998.

[4] G. Widmer and W. Goebl, "Computational models of expressive music performance: The state of the art," *Journal of New Music Research*, vol. 33, no. 3, pp. 203–216, 2004.

[5] 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*, vol. 29, no. 3, pp. 211–224, 2000.

[6] R. Ramirez, E. Maestre, and X. Serra, "A rule-based evolutionary approach to music performance modeling," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 1, pp. 96–107, 2012.

[7] R. Bresin and A. Friberg, "Synthesis and decoding of emotionally expressive music performance," in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1999, vol. 4.

[8] E. Maestre and E. Gómez, "Automatic characterization of dynamics and articulation of expressive monophonic recordings," in *Proc. of the 118th Audio Engineering Society Convention*, 2005.

[9] G. D'Incà and L. Mion, "Expressive audio synthesis: From performances to sounds," in *Proc. of the 12th International Conference on Auditory Display*, 2006.

[10] M. Grachten and G. Widmer, "Linear basis models for prediction and analysis of musical expression," *Journal of New Music Research*, vol. 41, no. 4, pp. 311–322, 2012.

[11] C. Erkut, V. Välimäki, M. Karjalainen, and M. Laurson, "Extraction of physical and expressive parameters for model-based sound synthesis of the classical guitar," in *Proc. of the 108th Audio Engineering Society Convention*, 2000.

[12] Alfonso Perez and Rafael Ramirez, "Towards realistic and natural synthesis of musical performances: Performer, instrument and sound modeling," in *Proc. of the Third Vienna Talk on Music Acoustics*, 2015.

[13] 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," in *Proc. of the 16th International Society for Music Information Retrieval Conference*, 2015.

[14] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proc. of the 10th International Conference on Digital Audio Effects*, 2007.

[15] J.L. Flanagan and R.M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[16] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[17] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th International Conference on Digital Audio Effects*, 2003.

[18] M. Mellody and G.H. Wakefield, "The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 598–611, 2000.

[19] L. Yang, K. Rajab, and E. Chew, "Vibrato performance style: A case study comparing erhu and violin," in *Proc. of the 10th International Conference on Computer Music Multidisciplinary Research*, 2013.

[20] J. Sundberg, "Acoustic and psychoacoustic aspects of vocal vibrato," *Vibrato*, pp. 35–62, 1995.

[21] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. of the 8th International Conference on Digital Audio Effects*, 2005.

[22] A. Gabrielsson and P.N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of Music*, vol. 24, no. 1, pp. 68–91, 1996.

[23] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27, 2011.

[24] M. R. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.