where c, f and t are chord labels, chord functions, and tempo, respectively. c^*, f^* and t^* are the related groundtruth attributes. CCE represents the categorical cross-entropy and MSE represents the mean-squared error. The weight of each error was heuristically determined by observing the reduction in loss during training.

3.3 Training Strategies

When a sufficient amount of emotion-labelled musical data is available for training, the model can be commonly trained with a backpropagation algorithm. However, when the amount of emotion-labeled data is not sufficient, transfer learning strategy enabling to include data without emotion labels can be effective. In such case, encoders are pre-trained using music examples without emotion labels, then the pre-trained encoders (weights are fixed) and randomly initialized decoders are concatenated and retrained only for the subset of tracks with emotion labels. However, groundtruth data may not provide the best examples since there are several possible arrangements following music theory and perception considerations [3]. As in [3], training was stopped to a fixed number of epochs (500) without using validation, when the loss was significantly reduced (learning rate = 1e-3) and subjectively consistent arrangements were generated with the test data. Results on the effectiveness of the transfer learning strategy are reported in Section 5.4.

4. MUSIC EMOTION QUANTIFICATION

In order to input emotional conditions to the networks, perceived or felt emotions associated to music have to be represented numerically. We investigated three ways to map emotions into Russell's arousal-valence (AV) space [17].

4.1 Emotion Average Representation

In the Deezer Mood Detection Detaset [18], the emotional tags from last.fm [1] were mapped to arousal and valence values using [19]'s results. In [19], statistics on participant ratings for emotion words are reported for valence, arousal and dominance on a nine-point scale. In most cases, multiple emotion tags can be associated to music content. To address this, one of the methods used e.g. in [18, 20], consists in summarising multiple emotions tags using the geometrical mean of the tag projections in the emotion space (e.g. AV space). We used such method yielding two-dimensional emotional features from a set of mood tags for songs. These were normalized in the range 0-1 for network input. We refer to this emotional representation as emotion average representation (EAR) in the remainder.

4.2 Emotion Surface Representation

The EAR expresses emotions locally in the AV space. However, as investigated e.g. by [20], there is a possibility that a same song suggests/induces multiple emotions to a same listener, or different emotions to different listeners. Therefore, similar to [21], two-dimensional Gaussian mixture models (GMMs) can be used to represent perceived/felt emotions associated to multiple mood tags associated to songs in the AV space. The average and standard deviation corresponding to the mood tags associated to a song are obtained from the experimental results in [19]. Based on the average and standard deviation, random sampling is performed and 10000 samples are generated for each tag. Like for EAR, the emotional features were normalized in the range 0-1. Clustering is performed using two-dimensional GMMs based on the randomly sampled points. Two Gaussian components were assumed sufficient to represent the emotional feature surface in the AV plane as in [21]. Finally, the average and standard deviation of each Gaussian component were used as the network input representation. We refer to this emotional representation as the emotion surface representation (ESR) in the remainder.

4.3 Emotion Category Representation

EAR and ESR are both continuous. However, music emotions may not be best represented by a dimensional model [22]. Studies on music emotion recognition, such as [23], used discrete representations of emotions through categorical variables. We also tested emotional representations with discrete categories. We distinguish four quadrants in the AV space; Q1: high arousal & high valence [joyful], Q2: low arousal & high valence [relaxing], Q3: low arousal & low valence [sad], Q4: high arousal & low valence [angry]). The AV space quadrant with the highest AV annotations determines the emotional category of the music. We refer to this emotional representation as emotion category representation (ECR) in the remainder.

5. EXPERIMENTAL EVALUATION

In this section, the proposed emotion-driven automatic music arrangement systems are evaluated for differences in network architectures and the effect of transfer learning.

5.1 Dataset and training

Network training requires a dataset in which melody, chords, tempo and emotions (perceived or felt) are simultaneously available. As in [3], we rely on the HTPD3 dataset which provides symbolic melodies, chords, and tempo. Time units were set to two beats (half bars in the 4/4 time signature). The chord played for the longest time over two beats was selected as the chord in that time unit. Notes that spanned a segmentation were split. However, tracks in HTPD3 are not labelled with emotion features. We retrieved crowd-sourced mood tags from last.fm [1] and allmusic.com [2] for the song and artists contained in HTPD3. Tags from last.fm [1] and allmusic.com [2] have previously been used in music emotion studies, see e.g. [18]. As some of the last.fm tags are not related to emotions, we filtered these out using the criteria proposed in [24]. This

[1] https://www.last.fm/

[2] https://www.allmusic.com/

method allowed us to tag approximately 4000 tracks available in HTPD3. We refer to this expanded dataset as the HTPD3 Emotion Dataset (HED) available at the link below [3]. There is some uncertainty on whether crowdsourced mood tags relate to perceived or felt emotions. We assume here that these tags relate to perceived emotions, however this can limit the performance of the model in the experiments. Another caveat is that the mood tags relate to audio versions of songs, whereas our model deals with symbolic music. Hence, as in [5], they are considered as "weak labels" for symbolic music. The dataset was divided into 90% training data and 10% test data. As suggested in [25], the use of commercial songs to train AI models for research may be considered fair use.

5.2 Evaluation conditions and music stimuli

We conducted a listening experiment to assess the performance of the proposed model. The study received ethics approval from our institution. Four models (BLSTM without [BL] and with [BT] transfer learning, Transformer without [TR] and with [TT] transfer learning) and the groundtruth (G) were used to create five stimuli for each melody. As we do not investigate the role of key root in this study, the chosen input melodies were converted to either C major or C minor (both training and testing data). Since it is not possible to test all possible EAR and ESR in a continuous way, 15 emotion input presets were prepared for the evaluation. Four types of emotion are represented by EAR, another four by ECR, and seven emotion distributions using the ESR. These input emotions were determined heuristically in order to be able to express as various emotions as possible. The specific configurations are shown on the study website [4]. The chord progressions and tempo generated by the models were converted to MIDI along with the input melody. Audio was rendered from MIDI files using FluidSynth [26] using a SoundFont called SGM-V2.01. As the research [27] has suggested that humans can distinguish the timbre of brass instruments and guitars clearly, the melody part was played by a saxophone and the harmony part by a classical guitar.

5.3 Procedure and Participants

Participants were given instructions on how to complete the study on a dedicated website [5] which can be used to listen to examples of generated accompaniments. The website displayed participants melodies (sampled randomly from the test set) which were represented in the piano roll style. For each of the 15 cases (corresponding to emotional presets not explicitly revealed to participants), participants had to press the "Execute" button to obtain five (four models BT, TR, BT, TT, and groundtruth, G) musical arrangements to rate using three questions:

Q1 The melody and accompaniment were musically coherent. (Likert item: 0 [Disagree] - 6 [Agree]). Participants were instructed that musical coherence here represents a measure of how well the musical accompaniment matches the melody.

Q2 How exciting (arousal) do you perceive the music to be? (Continuous value from 0.0 to 1.0)

Q3 How negative or positive (valence) do you perceive the music to be? (Continuous value from 0.0 to 1.0)

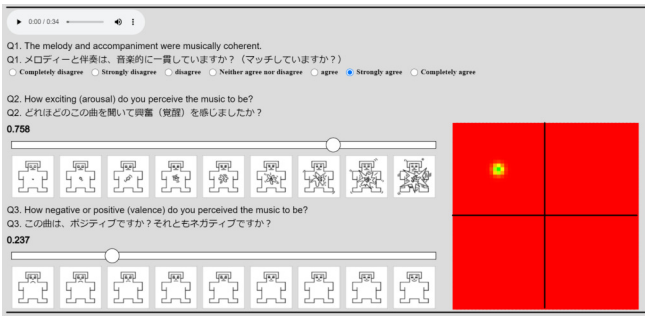

Figure 2. The evaluation interface (five panels had to be completed for each melody; four models plus groundtruth).

For Q2 and Q3, the self-assessment manikin [28] was used to support associations between numerical rating values and corresponding emotions, together with a representation of the rating in the AV space. Figure 2 provides a screenshot of the emotion rating interface. Participants rated 75 songs in total (15 emotion presets x 5 models) each lasting between approximately 15 to 30 seconds. As participants had to rate new arrangements, we assumed that familiarity with the melody, if any, did not affect the ratings (this would have to be assessed in future work). The whole experiment took about 45 minutes to complete.

Participants could not identify the models nor the groundtruth. Different melodies were randomly chosen from the test dataset and used for each emotional preset as, in a pilot testing phase, some participants expressed that listening to the same tune over and over made it difficult to evaluate after a while. For a given emotion preset, the five stimuli (BT, TR, BT, TT, and G) were generated for a same melody, to enable fair comparison between models.

The experiment was completed by 20 participants (7 females, 13 males). All participants were Japanese residents, 19 were Japanese and one was German. Their age ranged between 19 to 59 years (M=30.15, SD=14.05), where M and SD refer to mean and standard deviation, respectively. Six of them had at least one year of formal training in music theory. Eight had more than five years of formal training in their instrument (including voice).

5.4 Results

In statistical analyses, a Type I error of 0.05 was used except when mentioned otherwise.

5.4.1 Perceived musical coherence

Figure 3(a) illustrates the mean and standard error for each model, computed on the perceived musical coher-

[3] http://coconuts-palm-lab.com/EH/HED.zip
[4] http://coconuts-palm-lab.com/EmotionPresets/
[5] http://coconuts-palm-lab.com/EH/

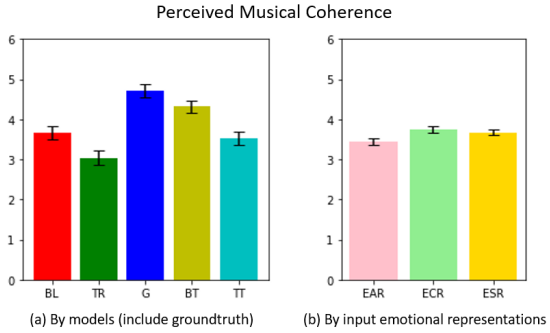

Figure 3. Results of a subjective evaluation experiment, with statistics on the musical coherence across comparisons. The means and standard errors are given.

ence considering all emotion presets. An analysis of variance (ANOVA) shows that the perceived music coherence yielded by the models presented significant differences (df=299, F = 66.750, p<.0001). Tukey's honestly significant difference (HSD) test shows that there is a significant difference between all pairs, except between the BLSTM melody context encoder without transfer learning and the Transformer melody context encoder with transfer learning. The BLSTM model with transfer learning achieves a significantly higher musical coherence compared to the other models. However, compared to the groundtruth, the arrangements generated by the machine learning models were found to be significantly less coherent.

Figure 3(b) displays the mean and standard error of the musical coherence for each emotional expression. ANOVA results showed a significant difference (df=319, F = 3.661, p = 0.025) and Tukey's HSD test showed that there was a significant difference only between EAR and ECR. This suggests that the type of emotional representation impacts perceived musical coherence. In particular, it is suggested that EAR may reduce the coherence of the generated music more than the other representations.

5.4.2 Errors between perceived and target emotions

By observing the error between target and rated emotions, it is possible to assess how well each model expresses the target emotions. The errors for the EAR were obtained using Euclidean distance. The ECR error was defined using the shortest distance between the emotion AV rating and the emotion category AV quadrant. The ESR error was defined by the negative log-likelihood that the rating belongs to the two Gaussians of the GMM component. It should be noted that each absolute emotion error is thus calculated on a different scale.

We also analyzed the relative emotion error, which indicates the degree of reflection of emotion, based on the ratio between the absolute error of the groundtruth and the absolute error of the arrangement generated by the model. If the absolute error for the music generated by the models is smaller than the absolute error for groundtruth, it suggests that the model may have been properly trained. Furthermore, the relative emotion error also makes it easier to compare different emotional representations. The relative error e_r is calculated as follows:

$$e_r = \frac{a_m}{a_g + \epsilon} \tag{2}$$

where a_m is the absolute emotion error for a model and a_g is the absolute emotion error for the groundtruth. ϵ is a small regularising term, avoiding cases where the denominator tends to zero.

The upper part of Figure 4 illustrates the mean and standard error of the absolute emotion error (a_m) for each model and emotional representation. The bottom part of Figure 4 shows relative emotion error (e_r) boxplots. The medians and quartiles are more appropriate to interpret the results than the averages (green triangles) which are strongly influenced by outliers. The blue dotted line is a threshold representing the absolute emotion error yielded by the groundtruth. If the relative emotion error is smaller than the threshold, it suggests that the model has been able to generate arrangements that are closer to the target emotion than the original groundtruth arrangement.

The hypothesis that all means are equal in the absolute error of the EAR is rejected based on the ANOVA (df=79, F=4.388, p=0.004). The results of Tukey's HSD test showed that there was a significant difference in absolute error between the BLSTM with transfer learning and the Transformer without transfer learning (p = 0.003). Moreover, based on the relative emotion errors of the EAR, it was found that the median and quartiles are smaller for BLSTM with transfer learning compared to without transfer learning. In particular, up to the third quartile, the errors were below the threshold, indicating that 75% of the arrangements generated by the model with the transfer-learned BLSTM are closer to the target emotion than the original one. The results show that the BLSTM with transfer learning can reduce the emotion error significantly more than the Transformer one.

ANOVA results suggest that the mean of the absolute error for ECR is not significantly different (df=79, F=1.757, p=0.155). Unlike for EAR, transfer learning does not seem to have had any particular impact for ECR.

According to ANOVA results, the mean of the absolute error for ESR is not significantly different (df=139, F=2.539, p=0.0557). The results of Tukey's HSD test also showed that there were only significant differences between the BLSTM with transfer learning and the Transformer without transfer learning (p = 0.042). When observing the relative error of the ESR, the smallest median, quartiles and mean were obtained with the BLSTM with transfer learning. Thus, when ESR was used, the emotional error was the smallest when BLSTM with transfer learning was used as for the EAR.

In addition, Figure 5 summarises the tempo of the generated arrangements for all the test data. The plotted points represent the actual tempo and the box plot shows the statistics. In all models, the tempo varied significantly when using ESR. This shows that ESR is the emotional representation that generates the most diverse tempi. The

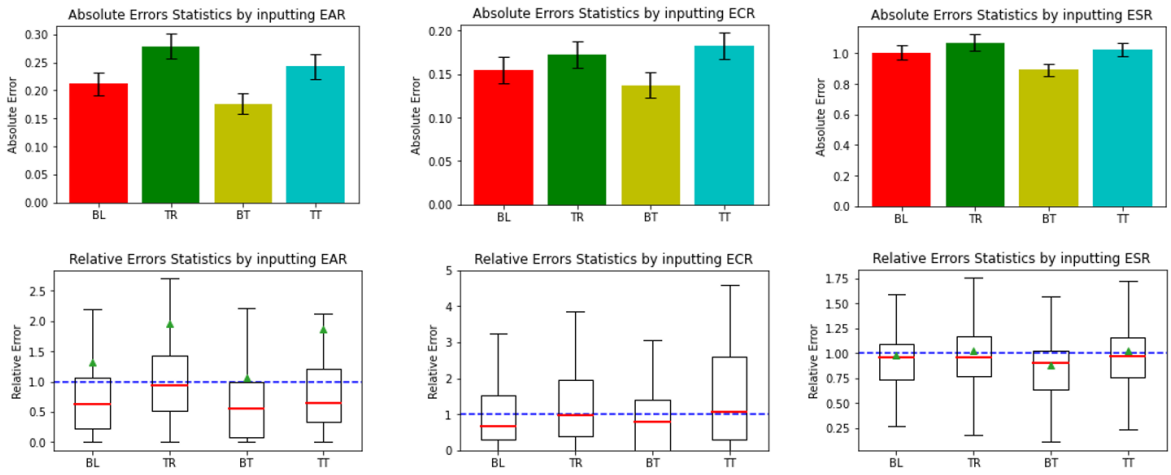

Figure 4. Results of the subjective evaluation experiment, with statistics on the absolute emotion error and relative emotion error across models.

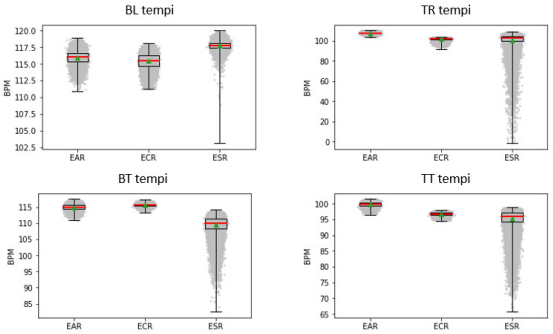

Figure 5. Statistics of the tempi generated for each model from the all melodies of the test data. Each point represents a generated tempo.

use of Gaussian variance, as in ESR, may support more complex emotional expressions.

5.5 Discussion

The highest perceived musical coherence and the lowest absolute and relative emotion errors were obtained when BLSTM with transfer learning was used for the melody context encoder. The results also show that transfer learning, a semi-supervised learning strategy, seems to be effective at generating emotional arrangements. A comparison of the relative errors of the BLSTM with transfer learning showed that the third quartile was not below the threshold only for the ECR, indicating that both EAR and ESR are superior representations for the emotion conditioning.

The generated arrangement could be computed quickly even with Intel(R) Core(TM) i7-9700K CPU (less than approximately 2.5 seconds per 16 bars of melody for any compared models). This could be useful for real-time music generation scenarios. Due to the simplicity of the model, the arrangement may be generated more quickly than state-of-the-art techniques such as that in [12] (approximately 50 seconds per 16 bars of melody for a lead sheet on the same CPU). However, the method in [12] generates the entire leadsheet, so the computation time cannot be directly compared to the ones reported here.

Several limitations should be highlighted. Observing the perceived musical coherence, there was no model in this experiment that could match the groundtruth. The reasons may be overfitting and a lack of data in the dataset. In most cases, the performance of the model was improved by transfer learning, so it is expected that more data and appropriate validation will help to build better models. Emotion errors remain relatively large and it is difficult to prove that the model consistently expresses the desired emotions. As it is unclear whether emotion labels in the training dataset represents perceived or felt emotions, this may contribute to inference errors. More knowledge about the listener's state and context would be needed to gauge more comprehensively emotion perception [9]. Results cannot be generalised since participants were from one provenance (Japan) and the sample size was small (20). More participants of different nationalities would be required to assess the generalisability of the proposed models.

6. CONCLUSION

We devised techniques for automatic harmonic and tempo arrangement of melodies controlled by emotional features, suitable for near real time applications. A network architecture to generate music expressing target emotions by predicting chord progressions and tempo for an input melody was proposed. In addition, three methods to quantify musical emotions were compared. To evaluate the results, we conducted an online listening experiment. For the melodic context encoder, the BLSTM model with transfer learning produced the most coherent arrangement and the one that best reflected the targeted emotions. The proposed method finds applications in assisting tools to create new music arrangements based on emotional directions and affective video games.

7. REFERENCES

[1] Z. Wang, D. Wang, Y. Zhang, and G. Xia, "Learning Interpretable Representation for Controllable Polyphonic Music Generation," in *Proc. of the 21st Int. Society for Music Information Retrieval Conference*, Montréal, Canada, 2020. [Online]. Available: https://github.com/

[2] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv Preprint arXiv:2005.00341*, 2020.

[3] Y.-C. Yeh, W.-Y. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H.-M. Liu, H.-W. Dong, Y. Chen, T. Leong, and Y.-H. Yang, "Automatic melody harmonization with triad chords: A comparative study," *Journal of New Music Research*, vol. 50, no. 1, pp. 37–51, 2021.

[4] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach: A steerable model for bach chorales generation," in *International Conference on Machine Learning*. Proceedings of Machine Learning Research (PMLR), 2017, pp. 1362–1371.

[5] S. Sulun, M. E. Davies, and P. Viana, "Symbolic music generation conditioned on continuous-valued emotions," *IEEE Access*, 2022.

[6] C. L. Krumhansl, "An exploratory study of musical emotions and psychophysiology." *Canadian Journal of Experimental Psychology/Revue*, vol. 51, no. 4, p. 336, 1997.

[7] S. Swaminathan and E. G. Schellenberg, "Current emotion research in music psychology," *Emotion Review*, vol. 7, no. 2, pp. 189–197, 2015.

[8] H. H. Tan and D. Herremans, "Music fadernets: Controllable music generation based on high-level features via low-level feature modelling," *arXiv Preprint arXiv:2007.15474*, 2020.

[9] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[10] A. J. Blood, R. J. Zatorre, P. Bermudez, and A. C. Evans, "Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions," *Nature Neuroscience*, vol. 2, no. 4, pp. 382–387, 1999.

[11] E. Coutinho and A. Cangelosi, "Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements." *Emotion*, vol. 11, no. 4, p. 921, 2011.

[12] D. Makris, K. R. Agres, and D. Herremans, "Generating lead sheets with affect: A novel conditional seq2seq framework," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[13] Y. Frachi, T. Takahashi, F. Wang, and M. Barthet, "Design of emotion-driven game interaction using biosignals," in *Proc. HCII*, 2022.

[14] R. Guo, I. Simpson, T. Magnusson, C. Kiefer, and D. Herremans, "A variational autoencoder for music generation controlled by tonal tension," *arXiv Preprint arXiv:2010.06230*, 2020.

[15] H. Lim, S. Rhyu, and K. Lee, "Chord generation from symbolic melody using blstm networks," *arXiv Preprint arXiv:1712.01011*, 2017.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[17] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[18] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *arXiv Preprint arXiv:1809.07276*, 2018.

[19] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[20] M. Barthet, D. Marston, C. Baume, G. Fazekas, and M. Sandler, "Design and evaluation of semantic mood models for music recommendation," in *Proc. International Society for Music Information Retrieval Conference*, 2013.

[21] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *International Society for Music Information Retrieval (ISMIR) 2010*, vol. 86, 2010, pp. 937–952.

[22] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content-to context-based models," in *International Symposium on Computer Music Modeling and Retrieval*. Springer, 2012, pp. 228–252.

[23] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen, "Exploring relationships between audio features and emotion in music," in *ESCOM 2009: 7th Triennial Conference of European Society for The Cognitive Sciences of Music*, 2009.

[24] M. Buffa, E. Cabrio, M. Fell, F. Gandon, A. Giboin, R. Hennequin, F. Michel, J. Pauwels, G. Pellerin, M. Tikat *et al.*, "The wasabi dataset: Cultural, lyrics and audio analysis metadata about 2 million popular commercially released songs," in *European Semantic Web Conference*. Springer, 2021, pp. 515–531.

[25] OpenAI, "Uspto comment regarding request for comments on intellectual property protection for artificial intelligence innovation," 2019, available: https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

[26] J. Newmarch, "Fluidsynth," in *Linux Sound Programming*. Springer, 2017, pp. 351–353.

[27] S. McAdams, "Musical timbre perception," *The Psychology of Music*, pp. 35–67, 2013.

[28] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behaviour Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

USING ACTIVATION FUNCTIONS FOR IMPROVING MEASURE-LEVEL AUDIO SYNCHRONIZATION

Yigitcan Özer[1] **Matěj Ištvánek**[2] **Vlora Arifi-Müller**[1] **Meinard Müller**[1]
[1] International Audio Laboratories Erlangen, Germany
[2] Brno University of Technology, Brno, Czech Republic
yigitcan.oezer@audiolabs-erlangen.de

ABSTRACT

Audio synchronization aims at aligning multiple recordings of the same piece of music. Traditional synchronization approaches are often based on dynamic time warping using chroma features as an input representation. Previous work has shown how one can integrate onset cues into this pipeline for improving the alignment's temporal accuracy. Furthermore, recent work based on deep neural networks has led to significant improvements for learning onset, beat, and downbeat activation functions. However, for music with soft onsets and abrupt tempo changes, these functions may be unreliable, leading to unstable results. As the main contribution of this paper, we introduce a combined approach that integrates activation functions into the synchronization pipeline. We show that this approach improves the temporal accuracy thanks to the activation cues while inheriting the robustness of the traditional synchronization approach. Conducting experiments based on string quartet recordings, we evaluate our combined approach where we transfer measure annotations from a reference recording to a target recording.

1. INTRODUCTION

In music information retrieval (MIR), synchronization techniques are essential for several applications including score following [1], content-based retrieval [2], automatic accompaniment [3], or performance analysis [4, 5]. Beside these applications, music synchronization has a great potential to simplify data augmentation, data annotation, and model evaluation. For example, one can use music synchronization to obtain additional training data for deep learning methods semi-automatically by transferring annotations from one recording to another recording. Furthermore, using music synchronization, one can transfer measure positions between audio recordings for navigation purposes, structural segmentation, and cross-version analysis [6, 7].

While traditional synchronization approaches typically rely on alignment algorithms such as dynamic time warping (DTW) and conventional chroma features used as the input representation [8, 9], the integration of additional onset-related information has proven to enhance the synchronization accuracy [10–12]. Inspired by the combined approach from [12], where *decaying locally adaptive chroma onset* (DLNCO) features are integrated into the synchronization pipeline, we incorporate in this paper onset, beat, and downbeat activation functions to obtain a better temporal accuracy while retaining the robustness of the original chroma-based synchronization approach (see Figure 1 for an illustration of the overall approach). The addition of activation functions results in a grid-like structure in the cost matrix, which guides the alignment through activation cues that point to note onsets or other musical events.

While the integration of DLNCO and spectral flux (SF) have led to substantial improvement of synchronization results [12,13], the detection of soft onsets constitutes a challenging problem due to their long attack phase with a slow rise in energy. To adapt the onset detection task to music recordings which comprise soft onsets and temporal–spectral modulations such as vibrato (e.g., string music), Böck and Widmer [14] introduced the superflux (SF*) feature. Furthermore, deep learning (DL) methods such as bidirectional long short-term memory (BLSTM) networks [15] and convolutional neural networks (CNN) [16] have led to significant improvements compared to conventional onset detectors.

As the main contribution of this paper, we show how one can integrate conventional and DL-based activation functions into the synchronization pipeline. Different from the approach in [12], we do not apply any hard peak picking but directly use onset-related activation cues. Furthermore, we go beyond onsets by integrating activation functions that indicate onset, beat, and downbeat positions. In particular for music with noisy and unreliable onset cues, we show that beat and downbeat cues are more reliable and better suited for improving the synchronization accuracy. For extracting beat and downbeat activation functions, we build on recent work by Böck et al. [17, 18], using recurrent neural network (RNN) models for extracting beat and downbeat activation functions.

To better understand our improved synchronization pipeline, we compare several synchronization approaches

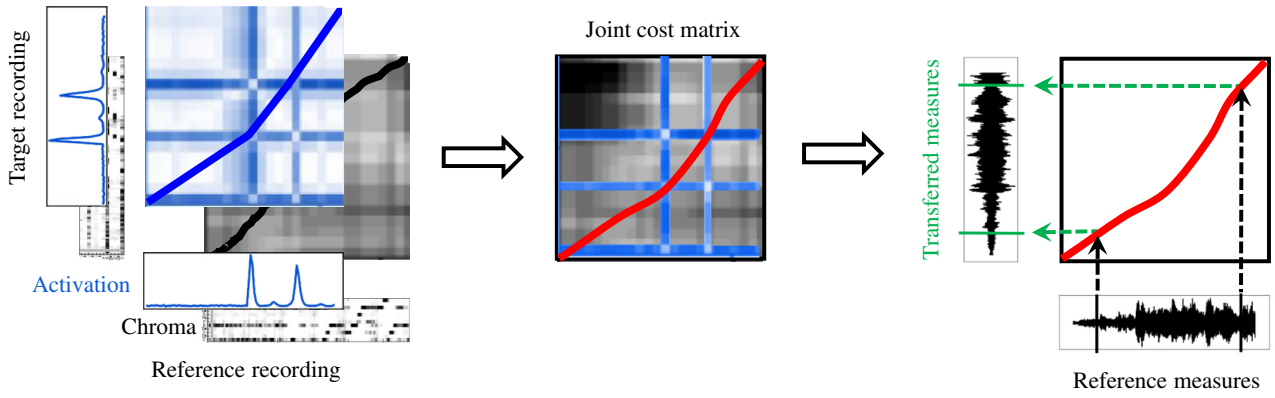

Figure 1: Overview of the combined approach integrating activation functions into a conventional chroma-based synchronization pipeline. The cost matrix computed with activation cues (blue) yields a grid-like structure to guide the alignment of musical events, whereas the chroma-based cost matrix (black) accounts for the robustness of the overall synchronization. The resulting warping path (red) is used for transferring measure positions.

where we transfer measure annotations from a reference recording to a target recording, similar to [19]. In particular, we conduct systematic experiments based on three versions of the String Quartet No. 12 in F major, Op. 96, composed by Antonín Dvořák. As string music generally comprises vibrato, tremolo, rubato, and abrupt tempo changes, which increase the musical complexity, the synchronization of string quartets is a challenging scenario. We show that integrating DL-based activation functions significantly improves the temporal accuracy while retaining the robustness of chroma features.

The remainder of the paper is organized as follows. In Section 2, we introduce our combined synchronization approach, explore conventional and DL-based activation cues, and show how to integrate activation functions into the synchronization pipeline. In Section 3, we present our dataset, the measure transfer between string quartet recordings as our application scenario, and report on our systematic experiments and empirical results. Finally, we conclude in Section 4 with prospects on future work.

2. COMBINED SYNCHRONIZATION APPROACH

In this section, we show how activation functions can be integrated into the synchronization pipeline to enhance the temporal accuracy of traditional chroma-based synchronization approaches. Here, we regard the activation function as a function that yields a value between 0 and 1. Each entry in the function indicates the likelihood of a certain musical event, e.g., onsets, beats, or downbeats, for each frame (time position). In the ideal case, the value of the activation function is one when an event occurs and zero otherwise. Note that we do not apply any peak picking in our approach, but only use the activation functions as temporal cues. This is opposed to onset detection or beat tracking where one needs to apply a temporal decoding method to obtain an explicit representation of onset and beat positions from the activation functions.

In the following, we first explore conventional onset-based activation functions in Section 2.1. Then, in Section 2.2, we investigate DL-based onset, beat and downbeat detectors. Finally, in Section 2.3, we explain how we integrate activation functions into the synchronization pipeline.

2.1 Conventional Onset-Based Activation Functions

2.1.1 DLNCO

DLNCO features are 12-dimensional pitch-based onset features, which combine the robustness of the chroma features with the accuracy of one-dimensional onset features. To compute DLNCO features, we first apply a pitch-wise audio decomposition. Then, we derive pitch-wise onset cues by considering points of energy increase (see Figure 2c for an illustration). DLNCO features are particularly suited for the music with clear note attacks such as piano music. For further details about the computation of DLNCO features, we refer to [12].

2.1.2 SF

Spectral flux (SF) captures the changes in the spectral content of an audio signal, and is widely used for onset detection [9, 20]. For the computation of SF, we apply a first-order differentiator on the log-compressed magnitude spectrogram of a music recording. Half-wave rectification follows the differentiation to keep only the positive differences between subsequent frames. As a final step, we subtract a local average function to enhance the peak structure (see Figure 2d).

2.1.3 SF⋆

Superflux (SF⋆) is a modified version of SF for detecting soft onsets [14]. These features are suitable for music recordings with vibrato, such as strings quartets. Similar to the SF algorithm, SF⋆ also relies on the detection of positive changes in the energy over time. However, it includes