Deep Learning for Expressive Music Generation

José Maria Simões
CISUC Department of Informatics Engineering of University of Coimbra
Coimbra, Portugal
josecs@student.dei.uc.pt

Penousal Machado
CISUC Department of Informatics Engineering of University of Coimbra
Coimbra, Portugal
machado@dei.uc.pt

Ana Cláudia Rodrigues
CISUC Department of Informatics Engineering of University of Coimbra
Coimbra, Portugal
anatr@dei.uc.pt

ABSTRACT

In the last decade, Deep Learning (DL) algorithms have been increasing its popularity in several fields such as computer vision, speech recognition, natural language processing and many others. DL models, however, are not limited to scientific domains as they have recently been applied to content generation in diverse art forms - both in the generation of novel contents and as co-creative tools. Artificial music generation is one of the fields where DL architectures have been applied. They have been mostly used to create new compositions exhibiting promising results when compared to human compositions. Despite this, the majority of these artificial pieces lack some expression when compared to music compositions performed by humans. In this document, we propose a system capable of artificially generating expressive music compositions. Our main goal is to improve the quality of the musical compositions generated by the artificial system by exploring perceptually relevant musical elements such as note velocity and duration. To assess this hypothesis, we perform user tests. Results suggest that expressive elements such as duration and velocity are key aspects in a music composition expression, making the ones who include these preferable to non-expressive ones.

KEYWORDS

Artificial Neural Network (ANN), Deep Learning (DL), Recurrent Neural Networks (RNN), Expressive Music Composition, Co-Creative Tools

ACM Reference format:
José Maria Simões, Penousal Machado, and Ana Cláudia Rodrigues. 2019. Deep Learning for Expressive Music Generation. In *Proceedings of ARTECH 2019, 9th International Conference on Digital and Interactive Arts (ARTECH 2019)*, October 23–25, 2019, Braga, Portugal. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3359852.3359898

ACM ISBN 978-1-4503-7250-3/19/10. . .
https://doi.org/10.1145/3359852.3359898

1 Introduction

The efficiency of machine learning algorithms is remarkable when considering its applications. From content filtering to recommendations these algorithms have proved their adaptability and effectiveness when applied to user-based services [14]. Besides commercially oriented applications, these artificial learning methods showed promising results in different areas, especially in content recognition (such as faces or objects) and text transcription from speech. All this is possible using DL - a branch of ML - which allows computational models composed by multiple processing layers to learn representations of data with several abstraction levels [14]. Over the past recent years, we have been witnessing the re-emergence of DL algorithms due to technological advances but also because of the current amount of data that is available to the average user, which did not exist before [8]. These recent advances have shown that the plasticity of DL algorithms are not limited to *simple* classification problems. In fact, a new field related to artificial arts [8] holds the application of these techniques to content generation such as image (generating paintings or style transfer), texto
(composing poems, for example) and music. Regarding the latter, different approaches have been studied to assess the ability of ANN to learn musical structures and generate music composition. Yet, only a few have taken into consideration music aspects besides note sequences.

Human and computer performances can be compared through files containing musical elements. For instance, a live human performance can be recorded and converted to a MIDI file, in which are encoded notes, durations and velocities (i.e. intensities). When opening the file in a software like GarageBand or MuseScore, we can hear the piece exactly as it was performed. This is where the computer performances meet human ones. As such, our computational model is capable of learning how to generate expressive sequences (i.e. duration and velocity) that are then converted into a readable file. Although we are aware that many other musical expressive elements may exist, we decided to work with these two for the purpose of this experiment.

Studies performed by Filippo Carnovalini and Antonio Rodà pointed out the importance of expressiveness in a musical composition [10]. The authors argue: *"the aspect of expressive performance is often overlooked by researchers in algorithmic composition"*, meaning that in most cases, this type of algorithms only learn how to generate the pitch and the quantized duration of notes, regardless of the dynamics [10]. Additionally, research

on music psychology has suggested that expression and dynamics in a piece are important aspects of music that avoid a music composition from sounding 'mechanical' when performed by a computer [9].

Designing an ANN to produce extremely good melodies has been proved to be a difficult task itself, as it is hardly comparable to human compositions. Google Brain's Team Project Magenta developed what might be one of the few models in which expressive factors have been taken into consideration [1] [4].
Although most of the Project Magenta's models are open source, it would be beneficial to have total control over the architecture and model itself, especially when handling expressive elements as it was important to assess the influence, they might have on music compositions. For such, we conceived a model not only capable of generating pleasant melodies but also capable of applying expressive variations to the notes.

In this article, we detail a model based on Recurrent Neural Networks (RNN) - particularly Long Short-Term Memory (LSTM) - that generates a series of artificial expressive compositions. The generated compositions are tested along with human performances in order to study the influence that expressiveness has on musical compositions.
Our contribution to the field comprehends a framework to improve a music piece by applying expressive elements on top of the melody. With this, we'll be able to better understand not only their importance but also how they might be used to improve a musical composition. Additionally, our model can be seen as an auxiliary tool to creatively aid human in expressive music composing.

2 Background

In this section, we briefly address topics regarding research on artificial music generation.

2.1 Music Elements

The main component of a music composition is the sequence of notes that make the piece. Notes vary according to their pitch. The pitch of a note is how loud it sounds (i.e. regarding its frequency). There are seven notes represented by letters: A, B, C, D, E, F and G (*la, si, do, re, mi, fa* and *sol* in Romance languages), naming all the natural notes within one octave (including Sharps and Flats) [20]. A sequence of notes results in a melody. A melody can be either monodic - or monophony - (only one note is played at the same time) or polyphonic (more than one note can be played at the same time) [8]. Among the expressive elements in study we have duration and velocity. Duration specifies the amount of time a certain note must be heard (i.e. last) [20] whereas velocity relates to the note intensity (i.e. how hard the note was/must be played). In the present work, we explore notes' pitch besides notes' duration and velocity.
Although these elements are described or written on a musical sheet, factors like duration and velocity are subject to variations concerning the performer himself: one must not rely on the fact that a performer plays a note according to the specified duration in a perfect manner (i.e. with no delay). The same idea applies to velocity, where different performers naturally apply different intensities throughout a certain piece.
Keeping this information in mind, it is interesting to observe that some factors affecting how a musical composition is perceived are actually not part of the composition itself (at least the written part of it), but rather subject to various interpretations of the performer.

2.2 LSTM

RNNs (a class of ANN) are dynamic systems - meaning that their internal state changes over time on the course of multiple iterations. This dependency on previous states introduces a notion of time to the model making it capable of handling sequences with greater accuracy [16]. In practical terms, the network processes an input sequence (one element at a time), keeping in its hidden units a 'state vector' containing information about the history of all the past elements of that same sequence [14]. This ability to memorize events is a key factor when applying RNNs to tasks such as predicting the next word in a sequence [14] [17]. Although this seems an almost perfect scenario to handle sequences and state dependencies, RNNs had been found to be difficult to learn to store information on long-term [14] [16]. The main problem causing this harness is the nominated vanishing or exploding gradient problem [16] [5].

One way to overcome this is by augmenting the network including in it a state of explicit memory [14]. Originally proposed by Hochreiter and Schmidhube in 1997 ([6] and [13]) LSTM networks solve this specific problem [14]. LSTMs are a special kind of RNNs capable of learning long-term dependencies, explicitly designed to avoid this issue [19]. This is possible as these artificial networks use a special hidden unit (a memory cell) that works as an accumulator. In specific, this means that it copies its own real-valued state and accumulates the external signal but also that it learns to decide when to erase (or forget) the content of that memory [14]. The effectiveness of LSTMs on handling sequences when compared to usual RNNs was one of the main reasons that led us to follow this architecture as an initial approach, as music composition is highly correlated with sequences. In the following section we discuss some related work on artificial music generation using LSTM-based approaches.

2.3 Related Work

Regarding content generation, some research and experiments presented multiple architectures when addressing artificial music compositions problems [8]. We focus on LSTM-based models to generate musical compositions due to the good performance of LSTM when dealing with long sequence predictions. In 2008, Douglas Eck et al. [12] introduced a *"music-specific sequence learner"* using an LSTM architecture. As explained by the authors, this specific type of architecture was chosen based on the need of dealing with long-term sequences once *"LSTM's architecture is designed to allow errors to flow backwards in time without degradation"* - a consequence of the absence of the already mentioned vanishing gradient problem. The network was trained using a MIDI dataset in order to generate music. The results were reported as positive in the sense that the model could learn musical structures and long-timescale dependencies in a time series. As this research was conducted mainly to get a better

understanding of how LSTM units would perform in this field, this specific model does not concern performed music [12].
A few years later, Aran Nayebi et. al (2015) compared the performance of LSTM and Gated Recurrent Units (GRU) - an architecture similar to LSTM - regarding musical generation [18]. With this specific experiment, the authors aimed to produce compositions that sounded unique and musically coherent and also analyze through comparison the performance of both architectures. In addition, this test was conducted using raw audio waveforms as an input (i.e. training data). As the authors reflect, the pieces generated by the LSTM architecture *"were significantly more musically plausible than those of the GRU"*.

Another experiment strengthens the LSTM's capability of learning sequences of musical events - this time with training data represented as text [11]. On their research paper "Text-based LSTM networks for Automatic Music Composition", Keunwoo Choi et. al (2016) present a model consisting of a text-based LSTM designed to learn relational patterns in text documents representing chord progressions and drum tracks - considering the scope of our work, we have analyzed the chord progression results. An interesting point of this experiment is the assumption taken by the authors, who assume that *"there is no constraint on the form of the text representation of music"* in order to observe if the network is able to learn musical patterns and structures based on such a *"weak assumption"* [11]. As a global view of the gathered results, the authors conclude that both architectures provided well-structured results, detailing that the networks learned *"the local structures of chords and bars after a sufficient number of iterations"*.

In 2017 Feynman Liang et. al. presented the "BachBot", an end-to-end automatic composition system designed not only to compose but also to complete musical compositions in the style of *Johann Bach* (chorales) using LSTM. To test BachBot's generated pieces success in a measurable way the authors developed a publicly accessible musical discrimination test provided online. The results gathered after the tests phase showed that the participants distinguished BachBot creations from *Bach* ones only 51% of the times, which the authors consider being a suggestion that *"BachBot successfully composes and completes music that cannot be distinguished from Bach significantly above the chance level"* [15].
As we mentioned before, Google Brains' Project Magenta [1] has also used RNNs to generate musical compositions. Performance RNN model is also based on a LSTM architecture designed to generate music with expressive timing and dynamics [4]. As detailed by the authors, this model was thought considering the essential role dynamics and expressiveness play in music [4]. The Performance RNN was trained with MIDI files of live piano recordings, having a vocabulary consisting of some MIDI events (such as duration and velocity). The results are presented as excerpts around 30 seconds each and described by the authors as lacking overall coherence but still *"quite expressive"*. This type of network and model are - to our knowledge - the closest to our intentions in the sense that it takes into consideration the expressive factor of a musical composition.

3 Model

In this section, we detail the architecture used to build our model, as well as all its functional process and generated results. Although the previously mentioned Magenta's models are open source, we had the need to create our own model. This allowed us to have total control over the implemented architectures and to perform different settings as the model improved. In this section, we detail the model implementation. Figure 1 illustrates an overview of the model.

3.1 Architecture

The model is composed of three independent LSTM-based networks, each one assigned to train sequential relationships between specific elements - Notes Network, Velocities Network and Durations Network. Each network holds the same structure: three LSTM layers with 100 neurons each, three Dropout layers [7] set to 0.3 and two Dense (fully connected) layers - the first with 256 neurons and the last one with the number of the existent vocabulary, using *'softmax'* as the activation function and a categorical cross-entropy loss function. The sequence below illustrates the layer-architecture used.

Input → LSTM → Dropout → LSTM → Dropout → LSTM → Dense → Dropout → Dense → Output

Regardingthe training sessions, we have found Adam optimizer to provide better learning performances when compared to Root Mean Square Propagation (RMSProp) and Stochastic Gradient Descent (SGD), particularly when training Velocities Network and Durations Network.

3.2 Process

The model receives MIDI files as an input. These files are then parsed using Music21 library [3], resulting in three different musical elements vectors (notes, velocities and durations). Each network is trained regarding its own assigned element to learn its relational patterns. After such, each network generates its own vector that is combined with the others to generate the artificial composed piece (i.e. a vector with all notes and respective velocities and durations). This final vector is then converted to a MIDI file resulting in the final audible piece.

The decision of building a model comindependennetworks was taken based on time limitations and computational constraints. We can illustrate this by looking at the information given to the model as training data. Each note extracted from a MIDI file has several attributes, among which we find velocity and duration. This means that before separating the data into three different vectors, a note is, in fact, a block (or an object) with both expressive elements appended. With that, for instance, a 'G' note with a duration of 1.0 (equivalent to a quarter note) is seen as different from a 'G' note with a duration of an eighth note. This would hinder the learning process, making it harder to find sequential patterns among notes, velocities our durations. If we were to keep a single network to learn from these 'full' blocks we would then need a significantly more powerful network, a much large dataset and consequently more training time. We have decided to address this issue by creating three independent networks, each in charge of learning from each set of elements.

The representation methodology is the same for each vector element. We encoded each element by addressing an integer to each unique value, forming a new vector of integers representing the sequences. The latter is then converted into a binary class matrix to match the network's parameters.

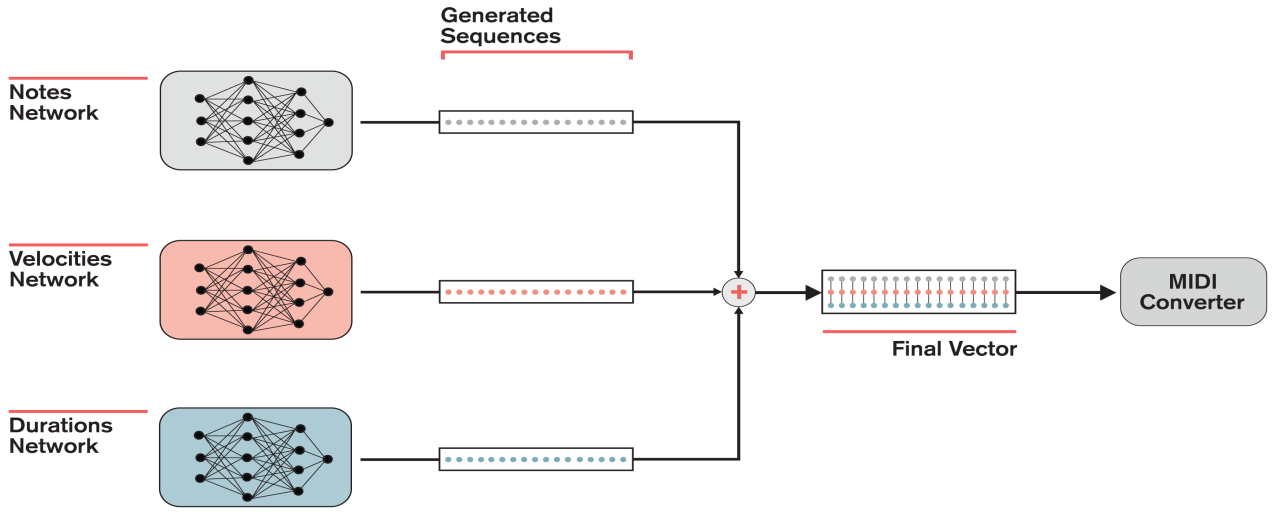

Figure 1: Overview of the model

The Notes Network was trained using a dataset composed of 36 *Johann Bach* MIDI files (a total of 60382 notes). Once that Velocities Network and Durations Network were both in charge of learning expressive elements, we used a dataset composed of various MIDI files recorded from live performances - also of *Bach* compositions - provided by Yamaha International E-Piano Competition [2]. We selected the number of live performances files to contain a close number of elements (having 62611). The choice of using a dataset performed by humans was based on two major points: First, although note durations are written and specified on a musical sheet, we can not rely on the fact that a performer is capable of following those exact timings with no delay or latency - which, in fact, could be easily done by a computer and therefore deviating from a human 'feel'. Secondly, most MIDI files used in Notes Network had a default velocity value assigned to each note, where live performances files had registered every single intensity throughout all compositions.
Overall, the full model took about 2 months of training until all networks showed significant improvements. (1 month for Notes Network and two weeks for Velocities and Durations Networks) - due to computational constraints, the training sessions were performed separately.

3.3 Artificial Expressive Music Generation

In order to evaluate the model's capability to produce expressive music, several compositions were generated.
We began by listening to 30 artificially generated pieces and then identified the consonant melodies of the compositions that we considered to have an interesting sequence of notes or variations. Afterwards, we selected the ones resembling expressiveness, that is the ones we considered being the most expressive. The decisions made on expressiveness were based on the full compositions, that is, we did not focus on a particular expressive element at this stage (durations or velocities) but rather at the impact that both could have on the music composition (and performance). This selection process resulted in a final set of 8 musical compositions (i.e. potential compositions to be used in User-Testing). We analyzed some of these pieces regarding their expressive patterns in comparison to non-expressive ones. Figure 2 and Figure 3 illustrate the graphical representations of expressive elements behaviour in a generated composition by the model and in a MIDI file with velocity and duration set to a default value.

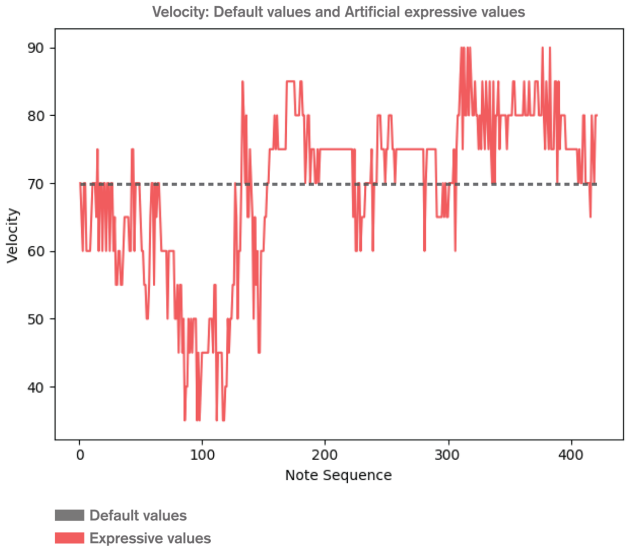

Figure 2: Velocity variation throughout the piece

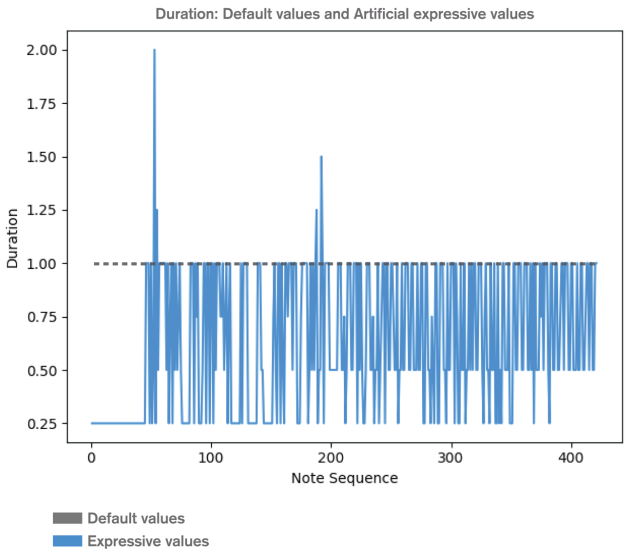

Figure 3: Duration variation throughout the piece

The expressive graphs depict interesting variance when confronted with the monotonous non-expressive ones. We observe that the expressive velocity graph shows different intensity moments in the piece just as a narrative: it gets softer after the first notes and ends in a 'striking' way. On the durations graph, we observe more abrupt changes making the graph look denser, with occasional longer notes.

Regarding note sequences, most of the compositions are not fully variant as some repetitive patterns may emerge during a small period of time, yet most of the melodies heard among the selected generations were considered pleasant, with some interesting expressive moments which we consider to be improving the dynamics of the pieces.

In order to get a better understanding of the actual impact that the introduction of expressive elements has on the pieces, we conducted user-testing sessions to address this issue, as we discuss in the next section.

Some of the generated expressive compositions can be heard by clicking the following link: *https://soundcloud.com/dlemg/sets/deep-learning-for-dynamic/s-cx3Y9*.

4 User-Testing

In this section, we describe the setup and procedures of the conducted user-tests. Participants were asked to rate excerpts according to their preference. We must state that we did not set any kind of threshold to define the success / level of expression of the output sample.

Instead, we decided to analyze the results based on their global tendencies and use those values as indicators of suggesting proof. Essentially, we have used these tests to examine how the generated compositions would be rated when compared to pieces composed and performed by humans. In addition, these terms of comparison let us understand if there is any sign of influence by the studied expressive elements - whether in human compositions or artificially generated.

3.1 Setup

To design the user-test scheme, it was important to narrow down the main purposes of this phase. Below we list the points to be addressed:

- Compare artificial compositions with human compositions;
- Study which expressive element has more influence on human/artificial compositions;
- Study how much full expressiveness influences human/artificial compositions;
- Test if artificial expressiveness can be perceived in the same way human expressiveness does in human compositions.

Having defined these terms, we designed two different sets: one consisting of the expressive generated compositions and the other composed of live *Bach* performances. The first set was built with the compositions studied in the last section. As for the second one, we opted to use live performances of *Bach* pieces - extracted from the same database used before (International E-Piano Competition [2]). As we intended to test several parameters, we decided to use only excerpts of the chosen compositions. Considering the extension of the user-tests designed, maintaining the original length of all compositions would mean an extensive test once some of the gathered live performances could last for 2 or more minutes - as for the generated ones. Thus, we believe, it could possibly hinder the analytical process as it would require quite long user-test sessions - which might be tiresome - and a great amount of concentration as it is easier to remember an excerpt than a two-minute piece for comparison purposes. Each of the selected compositions we extracted an excerpt with at least 15 seconds and no longer than 30 - naturally depending on the composition itself. To ensure impartiality, the artificial excerpts used in the tests were analyzed by an algorithm that compared the note sequences with training set ones. This algorithm was designed to find similarities between the given sequences. None of the excerpts used in the tests showed relevant similarities to the dataset - the most similar one had only a sequence of ten notes that was found on the dataset (equivalent to a second).

To attend all the aforementioned points, we divided the test into ten different parameters (or sections). These parameters were developed based on the different combinations of cases to evaluate. First, we pointed out all three elements of the music composition under examination: Notes, Velocities and Durations. Next, several parameters combinations were determined to include all the aspects to study. Figure 4 illustrates the parameters relational table.

Notes	Velocities	Durations
H	O	O
	H	H
	C	C
C	O	O
	C	C

H Human Composition/Performance
C Artificial Composition/Performance
O Element absence (Average)

Figure 4: Parameters relational table

These parameters were defined assigning the letter H (Human) and C (Computer) to identify the origin of the element - that is whether the melody of the excerpt (i.e. notes) was composed by humans - in this case, *Bach* - or artificially generated by our model - computer. The same logic is applied to both expressive elements, assigning an H to identify human live performances and C for artificial expressiveness. Apart from the letters defining computer (C) or human (H) compositions/performances, a capital letter O was introduced to represent the 'absence' of a certain expressive element regarding the parameters. To completely or partially exclude expressiveness from the excerpts we essentially have cancelled any variance within note velocities and/or durations. This was achieved by simply calculating the average value of each expressive element considering the full piece to which excerpt belongs, assigning that value to all elements (as a default value).

To ease the reading process and analysis, we list below each parameter acronym and relative meaning (note that by 'human composition' we mean *Bach* composition):

- HOO - Human composition, no expressiveness.
- COO - Computer-generated composition, no expressiveness.
- HCO - Human composition, computer-generated velocities and no durations (mean).
- HHO - Human composition, live performance velocities and no durations (mean).
- CCO - Computer-generated composition and velocities, no durations (mean).
- HHH - Human composition, live performance (no modifications).
- CCC - Computer-generated composition and expressiveness.
- HOC - Human composition, no velocities (mean) and computer-generated durations.
- HOH - Human composition, no velocities (mean) and live performance durations.
- COC - Computer-generated composition, no velocities (mean) and computer-generated durations.

The following link provides examples of expressive and non-expressive excerpts (*https://soundcloud.com/dlemg/sets/user-test-excerpts-examples/s-n2U3R*).

4.2 Participants Demographics

In this subsection, we provide some demographics concerning the participants that performed this user-testing phase. For the present study, we conducted 30 user-testing sessions (i.e. 30 participants). The selection of participants held no strict rule apart from the music education level - we were willing to test our goals with a sample in which the majority of the participants had no musical education. This was based on the fact that a person with a background in music theory and practice might be more 'sensible' regarding expressiveness detection. From all 30 participants, 60% were male and 40% were female of ages between 16 and 51 years old - resulting in the average age of 25. Regarding former musical education, we separated different degrees into four main levels for classification purposes: Low (less than one year of formal music education), Medium (from one to three years), High (three years or more) or None. In this categorization, we did not count mandatory music lessons (i.e. school lessons) as a formal musical education. Of all the 30 participants three had a Low level of musical education, two participants had a Medium level and only one had more than three years of formal musical education. This means that 80% of the participants had no musical education of any kind. Despite the musical background or musical education level, all of the participants were asked to rate the excerpts according to their tastes and opinions only, with no technical terms being mentioned.

4.3 Procedures

The tests were designed in the form of a questionnaire. Each questionnaire was composed of 20 excerpts (two per parameter) with the purpose of ranking them. By having two excerpts per parameter instead of only one we intended to obtain a clearer indication of the participants' ratings. We believe that this approach would increase the reliability of the ratings. Each question (i.e. excerpt) was rated within a range of 1 (*"I don't think this is a good excerpt in any way")* to 5 (*"I really think this is a good excerpt")* according to each participant's judgment. The questionnaire was programmed in order to introduce some randomness to it, meaning that each time a new questionnaire was generated the excerpts were randomly chosen - also avoiding the presence of the same file twice in the same test. Each question was numbered only, meaning that the participants did not have any kind of information on the excerpt beside the question number. As we considered important that all tests were performed under the same circumstances, we have chosen to conduct them personally instead of making an online-form. Although an online-form would possibly ease the process - and eventually gather more participants - we considered that assuring the same conditions in each session was more valuable than the number of participants. With this, the user-tests were conducted under a

controlled environment, meaning this that in each session a group member was present. Also, all participants were given the same environmental conditions and devices. This meant assuring that the tests were conducted in a quiet atmosphere and that through all tests the exact same sound device was used (portable speaker). Each test session had no time limit, meaning that each participant had no time restriction to perform the tests - also to listen to the excerpts the desired amount of times and in no particular order. Each user-test lasted for about 13 minutes (average estimation). No context regarding the aim of the tests was given to the participants. The participants were only asked to evaluate each excerpt according to what they considered being a good or a bad piece and to not to take into account the beginnings and the ends of the excerpt as they were part of a greater composition and some extracted from middle sections.

At the beginning of each session, the participants were asked to sign a consent form in which we assure their anonymity in order to use the collected data.

4.4 Results

To analyse the performed user tests, each set of answers was combined to evaluate the score of each parameter. Each excerpt could be rated from 1 to 5 and once there were two excerpts per parameter, the maximum score would be 10 in each questionnaire - resulting in a final maximum score of 300 considering all questionnaires. For clearance purposes, the final scores were converted to percentages. Figure 5 presents the ratings of all the parameters, ranked from the highest score to the lowest one.

The table (Figure 5) shows that the human excerpts with no expressiveness (HOO) got a total score of 51,66%. By introducing velocities (HHO), we observe an increase of 12,34%. In the case of durations (HOH), when these assume some variation the score increases by 23,34% - suggesting that durations have more impact on the composition, at least on these excerpts. When combining both expressive elements, the full expressive excerpts **(**HHH) assume an increase of 27,67\% when compared to the non-expressive parameter, reaching a score of 79,33%.

In general, we observed that the participants did favour the excerpts containing partial or full expressiveness. Concerning the excerpts generated by our model, Figure **5** shows that the non-expressive artificial compositions (COO) gathered a total score of 63,33%. When velocity expression is added (CCO), it provides an increase of 7%. Looking at the third parameter (COC), we observe that in this case the introduction of different durations caused an increase of 5,67%. It is curious to observe that in these artificially generated compositions the intensity seems to have more influence than note duration variation - although by a small difference - the opposite of what was observed on the last histogram.

Comparing the final parameter (CCC) to the non-expressive one (COO), we note that full expressiveness has promoted an improvement of 8,67%. In general - and similarly to what we observed in human compositions -, this results also suggest that expressiveness has improved the excerpts.

Finally, we analyze the experimental setup of combining human compositions with artificial expressive elements (HCO and HOC). Regarding velocity, we observe that the artificially generated sequences provided good results in comparison to the live performed version (HHO). However, it is important to note that in a previous analysis, durations have shown to be more influential than note intensity in human excerpts. The same is not verified when it comes to artificial durations applied to human compositions (HOC), where the participants usually preferred the original durations applied.

Observing all the given scores to each parameter, there's a suggestion that different combinations have a determined influence over the excerpts - and also that adding expressive elements (alone or simultaneously) tend to increase the value concerning the participant's preferences.

Chosen Combinations	Ratings
HHH	79,33%
HOH	75%
CCC	72%
CCO	70,33%
HCO	70%
COC	69%
HHO	64%
COO	63,33%
HOC	59,33%
HOO	51,66%

Full-expressive excerpts
Non-expressive excerpts

Figure 5: All parameters ratings (ranked)

5 Discussion and Future work

The main goal of our work was not to create a model to improve musical compositions, but instead to study the influence of expressiveness in music (human and artificial) so that the perception of a of an artificial music composition was closer to the one we get in a live human performance.

By examining the ratings obtained on the user-testing phase we noticed that the gathered results did provide some good insights regarding our study on artificial music expressiveness. The ratings exhibited a tendency to increase the preference as expressive elements were added to a musical composition - whether these elements are added individually or collectively. This was verified not only in human performances and compositions - where the full expressive pieces provided an improvement of 27,67% over the participants' preferences - but also in the excerpts generated by

our model with a global increase of 8,67%. These results **come to reinforce** findings and suggest that expressiveness is an important aspect of a music composition as it has the capability to enhance the musical output at a perceptual level, allowing participants to distinguish the multiple expressions that music can take in a natural an effortless way. Comparing the artificial non-expressive compositions with expressive ones we observe that in some cases - especially when the generated piece holds a sequence of stacking notes - expressiveness plays a major role in making an artificial composition to sound more natural, closer to a human one.

Another key aspect in the evaluation concerns the generation of melodies. We consider that to properly study expressiveness the model must be capable of generating sufficiently good note sequences (i.e. good enough so it won't compromise the ratings due to inconsistency our severe dissonance). The validation of this point was done by comparing the artificial excerpts to human ones (i.e. *Johann Bach's*), where the absence of expressiveness (HOO *vs* COO) indicates that the network has learned to generate coherent melodies. Although the non-expressive artificial melodies got a higher rating when compared to non-expressive human ones, the same did not apply when all expressiveness was introduced. It is important to note that this does not necessarily mean that the chosen artificial melodies are better than the human ones. In fact, it is likely to happen due to the exact opposite: *Bach* pieces can be more complex in its structure than the compositions generated by our model. The human chosen excerpts contained some variations regarding note positions, which created some moments of silence between notes or passages where a lot of notes were played in a short amount of time. Our model was not trained to learn this note 'allocations' (i.e. offset), meaning that in each generated composition the notes were positioned according to a generic value, making all notes equally distanced. The variation in *Bach* pieces regarding this factor may have contributed to a cruder sounding in non-expressive compositions than the artificial excerpts where all notes were metrically and equally sequenced.

Nonetheless, there are some points that we consider important to address in future work. First, the quality of the full generated compositions is an aspect that needs improvements. As we presented, the pieces generated by our model hold quite interesting melodies and sequences, but when considering the full composition, we observed some lack of coherence and some sporadic monotony. Being a topic that has already been pointed out by several authors, we agree that it is probably one of the main issues to overcome regarding artificial music. In addition, we estimate that another important factor to improve these musical pieces is to train the network (or model) to learn how to begin and end a composition. The absence of this notion - as seen in our generated compositions - makes the composition less natural to our ears by suddenly ending without any kind of change of 'announcement', which may sound more mechanical and less aware of the full structure - and thus far from human capacities. Nevertheless, it would be interesting in future work to train the model to learn how to position notes with some degree of variance and not equally distanced and test how this factor influences the piece.

Another feature that could be interesting to study is the correlation between the networks their learning and generation process instead of having a model with independent networks like ours. This could be a good way to experiment whether artificial networks find compositional patterns by being dependent - for instance, generating durations based on the pre-generated note sequences.

Regarding our conducted work, we hereby address some points as well. On the one hand and considering that all the architectures used haven't reached a clear state of stabilization, having more training time would possibly produce better results. Although we consider having good results when looking at the available computational resources and timeframe, having a larger dataset or more powerful architecture would theoretically improve the compositions. On the other hand, it is also important to note that regardless of the good gathered results as we have no proof that the model used is the best one to address this problem. It could be valuable to test our process with different models by exploring other architectures. Lastly, we must keep in mind the size of the user-testing sample. As we described, we preferred having local control of the procedures. An online questionnaire would probably increase the number of participants, but that way we would have no control over the environment. Having that said, it could be beneficial to test the different parameters on a larger scale.

In sum, we believe that the present work has positively contributed to the field of artificial music generation by studying how a factor like expressiveness - that is usually despised - may perceptually influence music compositions, providing some good indicators of how much a music piece is affected by it. In addition, our model could also be used as a co-creative tool by exploring different expressive patterns applied to human compositions. Once all networks that compose the model are independent, it provides the freedom to experiment with specific elements. This enables the user/composer to test different expressive approaches in the same melodic composition and even mix expressiveness from baroque (as an example) music performances with note sequences from different genres.

5 Conclusion

In the last decade, different experiments and approaches have been presented in the field of artificial music composition. Although some results show promising melodies, the majority does not take into consideration perceptually relevant musical aspects such as expressive variations.

In this paper, we presented a model capable of generating artificial expressive music compositions in order to study the influence of expressiveness on human and artificial compositions. Based on recent experiments regarding artificial music generation, we built an LSTM-based model composed of independent networks - each one in charge of a different sequence of elements. We evaluated the generated pieces' expressiveness, conducting user-testing sessions where we assessed the impact of different musical elements on human and artificial excerpts.

Based on the experiments conducted, our results suggest that aspects such as expressiveness shall be taken into consideration when generating artificial music that seeks an expression close to the human performance.

ACKNOWLEDGMENTS

This project was funded by the Portuguese Foundation for Science and Technology (FCT) under the grant SFRH/BD/139775/2018.

We would like to thank all the participants that performed the user-test sessions.

References

[1] [n. d.]. Google Brain Magenta. https://magenta.tensor ow.org. [Online; accessed 04-Dec-2018].

[2] [n. d.]. International e-Piano Competition. http://www.piano-e-competition. com/. [Online; accessed 24-May-2019].

[3] [n. d.]. Music21 Library. http://web.mit.edu/music21/. [Online; accessed 18-Nov-2018].

[4][n.d.].PerformanceRNN:GeneratingMusicwithExpressiveTimingandDynam- ics. Magenta Model. https://magenta.tensor ow.org/performance-rnn. [Online; accessed 04-Dec-2018].

[5] Eniola Alese. (2018). The curious case of the vanishing & exploding gra- dient. https://medium.com/learn-love-ai/the-curious-case-of-the-vanishing-exploding-gradient-bf58ec6822eb. [Online; accessed 20-Dec-2018].

[6] FilippoMariaBianchi,EnricoMaiorino,MichaelCKamp meyer,AntonelloRizzi, and Robert Jenssen. 2017. An overview and comparative analysis of recurrent neural networks for short term load forecasting. arXiv preprint arXiv:1705.04378 (2017).

[7] Theodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2015. Where to apply dropout in recurrent neural networks for handwriting recognition?. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 681–685.

[8] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. 2017. Deep learning techniques for music generation-a survey. arXiv preprint arXiv:1709.01620 (2017).

[9] Sergio Canazza, Giovanni De Poli, and Antonio Rodà. 2015. CaRo 2.0: an interactive system for expressive music rendering. Advances in Human-Computer Interaction 2015 (2015), 2.

[10] Filippo Carnovalini and Antonio Rodà. 2019. A multilayered approach to automatic music generation and expressive performance. In 2019 International Workshop on Multilayer Music Representation and Processing (MMRP). IEEE, 41– 48.

[11] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Text-based LSTM networks for automatic music composition. arXiv preprint arXiv:1604.05358 (2016).

[12]DouglasEckmandJasminLapalme.2008.Learningmusicalstructuredirectlyfrom sequences of music. University of Montreal, Department of Computer Science, CP 6128 (2008).

[13]SeppHochreiterandJürgenSchmidhuber.1997.Longshort-termmemory.Neural computation 9, 8 (1997), 1735–1780.

[14] Yann LeCun, Yoshua Bengio, and Geo rey Hinton. 2015. Deep learning. nature 521, 7553 (2015), 436.

[15] Feynman T Liang, Mark Gotham, Matthew Johnson, and Jamie Shotton. 2017. Automatic Stylistic Composition of Bach Chorales with Deep LSTM.. In ISMIR. 449–456.

[16] Zachary C Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019 (2015).

[17] TomasMikolov,MartinKara át,LukásBurget,JanCernocký,and SanjeevKhudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH. [18] Aran Nayebi and Matt Vitelli. 2015. GRUV: Algorithmic music generation using recurrent neural networks. Course CS224D: Deep Learning for Natural Language Processing (Stanford) (2015).

[19] Christopher Olah. (2017). Understanding LSTM Networks. http://colah.github.io/ posts/2015-08-Understanding-LSTMs/. [Online; accessed 21-Dec-2018].

[20] Catherine Schmidt-Jones and Russell Jones. 2007. Understanding Basic Music Theory. Connexions, Rice University, Houson, Texas.