An intelligent music generation based on Variational Autoencoder

Tao Wang
School of Information Engineering
Zhengzhou University
Henan, China
794574617@qq.com

Junzhe Liu,Cong Jin,Jianguang Li*
School of Information and Communication Engineering
Communication University of China
Beijing, China
Corresponding Author:
jincong0623@cuc.edu.cn

Shihao Ma
School of Information Engineering
Zhengzhou University
Henan, China
923799765@qq.com

***Abstract*— In this paper, GAN and VAE are combined with deep learning network to generate intelligent music based on music theory rules, and to explore intelligent music generation algorithm. Different from the traditional algorithmic composition, it is not necessary to manually add complex rules, but trains the initial music set, evaluates and filters the music collection, and ultimately generates music via the RVAE-GAN neural network. The fitness function calculates a weighted sum of a series of features of a piece of music, such as pitch and rhythm distribution, and can also calculate a series of theoretical rules of music theory from the distance between a particular piece of music. Furthermore, we take the music theory and practical experience into account, put forward the expressions and a rule function of rhythm, which are given in a mathematical way. On this basis, the semi-supervised algorithm is used to form the chord structure model, combined with the feature extraction of music, and the intelligent generation music based on GAN confrontation generation network and VAE network combined with music theory rules is proposed and proposed. And mass production has important theoretical and practical significance.**

Keywords-VAE; GAN; Music theory rules; Rhythm

I. INTRODUCTION

Algorithmic composition, or automated composition, is an attempt to use a formal process to minimize the involvement of a person (or composer) in the use of computers for music creation [1]. In a series of techniques, music is characterized by extracting parameters such as pitch, time value and controlling them separately. From each parameter, select the possible values in turn and organize them into a sound column parameter value which can be changed based on the current sound column or the inversion or inversion of the sound column. From 1906 to 1912, Markov proposed and studied a general schema that can be used to study natural processes using mathematical analysis methods—the Markov chain, which is known to present the next note based on probability [2][3]. Cybernetic Composer, developed by Ames and Domino, is a typical system for successful use of Markov chains and stochastic processes [4]. In the field of music, rhythm is the source of the vitality of music. Waltzes in classical music, for example, are characterized by its triple meter style, in which the dance steps rise and fall with the beat, giving one a magnificent and elegant enjoyment. Or swing music in jazz [5], where the abundance of syncopation and the microtiming deviations (MTDs) [6] created by the performer's innate sense of music, make the music so infectious that one feels the urge to dance with it. All of the above examples impress the listener because of their unique rhythmic style. It can be said that rhythm is an important and inseparable part of music, any style or genre of music can not exist without rhythm [7]. With the research of people, [8][9] to establish a model in a well-defined field or to introduce a clear structure or rule, it is necessary to use a music knowledge base, Ebciogln [10] established a backtracking Backtracking specification language (BSL), which is used to implement CHORAL, a rule-based expert system that can construct a four-part chorus with Bach style for the monophonic main melody, and has certain practical value. Then Mozer constructed CONCERT using recursive neural network technology [11], and used back propagation learning algorithm to train CONCERT to create melody in one sound and one tone. With the rapid development of deep learning in various application fields, people's research on it is getting deeper and deeper. Here we propose a model based on deep learning to build music generation using the rules we define.

Unsupervised learning networks are most suitable for generating data with time-series information, especially for the recursive neural networks (RNN) and long-term and short-term memory networks (LSTM), which perform quite well in this field. However, these networks are underperforming in dealing with the disappearance and explosion gradient problems. In order to solve these problems, scientists are constantly researching new network structures. Currently, variable-point automatic encoders (VAEs) and generative confrontation networks (GANs) are widely used to generate complex structural models. Music is generated intelligently based on the VAE-GAN network in combination with the rules we have specified.

The chapters of this paper are arranged as follows: the second part introduces the theoretical part and the prescribed rules; the third part introduces our experimental part; the fourth part introduces the experimental results and analysis; the fifth part carries on the summary to the paper, and has made some plans to the future development.

978-1-7281-8138-7/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCST50977.2020.00082

II. RELATED THEORIES AND FODELS

1. *VAE network*

A variational auto-encoder (VAE) is a directional model that uses good approximation inference and can be trained purely using a gradient-based approach. The VAE first samples z from the code distribution$P_{\text{model}}(\text{z})$. The sample is then passed through the micro-generator network g(z). Finally, x is sampled from the distribution $P_{\text{model}}(\text{x};\text{g}(\text{z})) = P_{\text{model}}(\text{x}|\text{z})$. The VAE consists of an encoder $q_\lambda(x|z)$ approximating a posteriori $p(z|x)$ and a decoder $p_\theta(x|z)$ of parametric likelihood $p(x|z)$. In practice, the approximate posterior distribution and likelihood distribution ("encoder" and "decoder") are parameterized by the neural networks of parameters λ and θ, respectively. Following the framework of variational reasoning, we minimize the KL divergence by maximizing the lower bound (ELBO) [2], using the KL divergence between the encoder and the posterior $p(z|x)$ to make the posterior it infers to $z\sim q_\lambda(z|x)$ and KL(• ‖ •). Calculating the gradient by ELBO is not feasible due to the sampling operation used to acquire z. So in the common case where $p(z)$ is a diagonal covariance Gaussian, this can be done by replacing $z\sim N(\mu,\sigma)$.

The encoder and decoder network in VAE, specifically the encoder $q_\lambda(z|x)$ is a cyclic neural network (RNN) that processes the input sequence $x = \{x_1, x_2, \ldots, x_T\}$ and produces A series of hidden states h_1, h_2, ..., $h_{(t)}$. The distribution parameter on the latent code z is then set to a function of h(t). The decoder $p_\theta(x|z)$ sets the initial state of the decoder RNN using the sampled potential vector z, and the decoder RNN automatically generates the output sequence $y = \{y(1), y(2), \ldots, y(t)\}$. The model is trained to reconstruct the input sequence (ie, $y_i = x_i, i \in \{1, \ldots, T\}$) and learn an approximate posterior $q_\lambda(z|x)$ close to the previous p(z). For the encoder $q_\lambda(z|x)$, we use a bidirectional RNN network. By analyzing this decoder, a hierarchical RNN decoder is used [12].

2. *GAN network*

Generative Adversarial Networks (GAN) is a deep learning model and one of the most promising methods for unsupervised learning in complex distribution in recent years. The model produces fairly good output through mutual game learning in the framework (at least) two modules: the Generative Model and the Discriminative Model. Samples generated directly by the generator network are as follows:

$$x = g(z;\theta^{(g)}) \tag{1}$$

Its opponent, the discriminator network attempts to distinguish between samples extracted from the training data and samples extracted from the generator. The discriminator emits a probability value given by $d(x;\ \theta(d))$, indicating that X is the probability of a real training sample rather than a forged sample extracted from the model.

The training purpose of the discriminant model D is to maximize its discriminative accuracy. When this data is judged to be from real data, the label 1 is from the 0 when the data is generated. Contrary to this purpose, the training objective of generating the model G is to minimize the discrimination accuracy of the discriminant model D. In the training process, GAN adopts a very direct alternating optimization method, which can be divided into two stages. The first stage is to fix the discriminant model D, and then optimize the generated model G, so that the accuracy of the discriminant model is reduced as much as possible. The other stage is to generate the model G fixedly to improve the accuracy of the discriminant model, which objective function is:

$$\min_G \max_D V(D,G) = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))] \tag{2}$$

This drives the discriminator to attempt to learn to correctly classify the sample as true or falsified, while at the same time, the sample of the generator is indistinguishable from the actual data when it converges, and the discriminator outputs one-half everywhere. There by achieving the effect of learning.

Generating a confrontational network framework makes it possible to train any kind of generator network. Most other frameworks require that the generator network have some specific functional form, such as the output layer is Gaussian. It is important that all other frameworks require a generator network that is spread over non-zero masses. Generating a confrontational network can learn to generate points only on a thin manifold that is close to the data. There is no need to design a model that follows any kind of factorization, and any generator network and any discriminator will be useful. Compared to VAE, it has no lower limit of variation. If the discriminator network is perfectly suited, then this generator network will perfectly restore the training distribution. In other words, the various antagonistic generation networks will be asymptotically consistent, while the VAE has some bias.

III. MULTIMODAL NEURAL NETWORK - RVAE-GAN

1. *Network model*

In view of the above two, combined with the advantages and disadvantages of each network, we propose a new multi-modal neural network model: rule-based neural network (RVAE-GAN). The network model includes an encoder E, a generator G (decoder) and a discriminator D. as the picture shows:

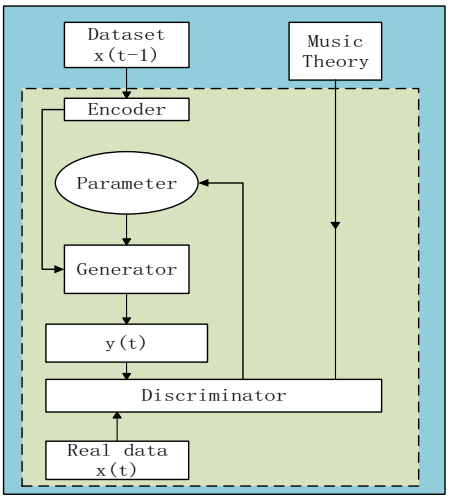

Figure1. System chart

The VAE decoder and GAN generator are combined into one by having them share parameters and training together. The main architecture of the three networks used in this section has been added to the convolutional neural network. The input time related data is treated as a series of individual frames with internal correlation. These frames are extracted using a convolutional neural network while maintaining their interdependencies. For each pair of consecutive frames, the encoder is used to encode the previous frame to its corresponding potential information. The generator then attempts to generate (predict) information for subsequent frames from the potential distribution of the previous frame. This combines the current information with the information from the previous frame to generate the next desired content. Each pair of current real training frames and composite frames is then forwarded to the discriminator as real data and dummy data respectively[13].

2. Rule algorithm

- Melody rules: The reward function is formed by a series of creation rules or the completion of constraints formed by music theory. The melody line is one of the most important means of expression in music. In the general sense, it is impossible for music to leave the melody line and rhythm. The performance and regularity of the melody line are based to a considerable extent on the downward natural relationship of the upward trend of tension and the gradual decline of tension. According to the constraints of the interval in the melody line, we try to avoid large jumps exceeding five degrees, while the intervals in the melody line avoid large jumps exceeding octave while avoiding large jumps in the same direction in three notes. When the rising melody line is composed in turn, after a big jump, it is generally followed by a small interval to form a descending melody line in three notes, we try to avoid using the three full or minus five degrees in the melody line and repeating the same note continuously. If the interval difference between two adjacent notes is greater than octave, it will be recorded as 0, otherwise it will be recorded as 1, which is the average value of each piece of music. That is: if a piece of music has a group of adjacent notes whose interval difference is greater than the octave, the remaining group b is less than or equal to the octave. It's expressed by the following formula:

$$g_1(x) = \frac{a*0+b*1}{a+b} \tag{3}$$

In terms of musicality, music that rises or falls continuously in pitch is not appropriate. Therefore, we use $g_1(x)$ to evaluate the overall outline of a piece of music. The pitch line is the line of the rhythm of time. The pitch line is dominated by progressive or small jumps, supplemented by big jumps. To guide the algorithm design, construct a pitch rhythm rule model (denoted as R_p), as shown in the following equations :

$$\Delta p = |p(i) - p(i-1)| \in S_I \tag{4}$$

$$S_I = \{I / I \geq 0,\ I \leq I_{\max}, I \in Z\} \tag{5}$$

$$I_{\max} = \begin{cases} 5, \ if\ p(i) \in S_P, p(i-1) \in S_P \\ 12, if\ p(i) \in S_P, p(i-1) \notin S_P \end{cases} \tag{6}$$

In equation (6), $P(i)$ is the pitch of the i-th note, SI is the sample set of the intervals of adjacent notes, and each sample value represents the number of semitones of the two tones (the intervals are expressed by the number of semitones) In equation (8), S_p is a set of pitch samples in the unified body of perception. The sensible unity is something that belongs to the same structure in the human sense. If the two sounds are not in a unified body, even if they are adjacent, there will be no sense of unity, or the experience of unity is weak. The music information expressed by R_p is "only use progressive or small jumps within the perceptual unity, and large jumps can be used between the perceptual units".

- Harmony rules: Try to use the consonant interval (that is, the interval with a small common multiple of the vibration frequency ratio, such as pure one, four, five, octave, and three or six degrees), with or without the use of the uncoordinated interval (i.e. the ratio of the vibration frequencies). A large multiplier, such as two or seven degrees. But too much use of the consonant interval will produce a "single" and "not rich" feeling in the auditory, both in the sound process and in the melody interval. Applicable, because even if the physical stimulation of the sound stops, the auditory impression remains in the brain. And this consideration also has a certain contradiction with the pitch rhythm rule R_p . In order to balance the beauty and performance of the music and R_p, here set a "uniform rule" Rule, as shown in the following equations :

$$I_b \in \{0,2,3,4\} \tag{7}$$

$$I_{bb} \in \{0,2,3,4,5,7\} \tag{8}$$

$$I_{\to P} \in \{7,9,10,12\} \tag{9}$$

$$N_{i=1} \in \{1,2\} \tag{10}$$

Among them, I_b, I_{bb} and $I_{\to P}$ represent the interval sample values within the bar, between the bars, and the climax, which is the number of times of the minor second interval (the number of semitones is 1) in a passage, and the number of uses is limited to not exceed 2, because the interval has a strong tension.

- Rhythm rules: In the quantitative analysis of the time series identified by the model, the following rules are obtained by combining the relevant definitions of music theory (where r denotes the reference value of the rhythm variable, and p denotes the pitch variable.

  - A rest: a rhythm variable with a corresponding duration expressed as a pitch variable of 0, as in:

$$r_0 = r(p = 0). \tag{11}$$

  - A weak rise: expressed as dividing the rhythm variable of the corresponding duration into two, with the first half as a rest, as in:

$$r_1 = \frac{r_0}{2} + \frac{r_1}{2}. \tag{12}$$

  - A dotted note: the extension of the duration is 1.5 times of the reference value, as in:

$$r_2 = 1.5r. \tag{13}$$

- A tie: for a tie that connects n notes, it is expressed as an extension of the duration to n times the reference value, as in:

$$r_3 = nr. \tag{14}$$

- A triplet and more: for an n-linking note, it is expressed as a continuous n rhythm variable whose duration is 1/n, as in:

$$r_4 = \sum_{i=1}^{n} \frac{r}{n}. \tag{15}$$

IV. Experiment Procedure

1. Experimental data

The data used in this paper is the classical piano MIDI dataset, which contains music files of different formats for all classical music. These songs are transmitted on the digital piano through the sequencer on the MIDI dock, and then converted to audio format. And we used the Pretty-MIDI [14] toolkit to process the individual MIDI files in the dataset, marked the time intervals and note types to obtain rhythm sequences. Each MIDI file is divided into separate sections, and we remove the songs that are not in the C key or do not use the four shots. For each measure, we set the width (time resolution) to 96 to model common time patterns, such as triplets, sixteen notes, and (minimum notes) the 32th note. We set the height to 84 to cover the pitch from C1 to C8. Therefore, the size of each data tensor is 96 (timestep) * 84 (note) * 5 (track). The value of each element of the matrix is the velocity (volume) of the notes at a certain time step. The sequence of n bars is represented by $X = \{x_0, x_1, \dots, x_{(t-1)}, x_t, \dots, x_n\}$, where $x_{(t-1)}$ and x_t are two consecutive pieces of information.

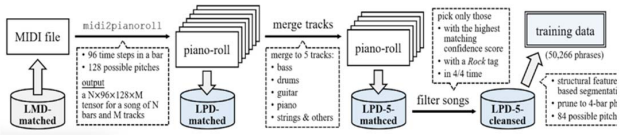

Figure2. Data preprocessing

2. Experiment procedure

The VAE decoder and the GAN generator are merged into one by sharing parameters and training together. The primary architecture for the three networks used in this section is CNN. The input time related data is treated as a series of individual frames with internal spatial correlation. These frames are generated separately using CNN while maintaining the dependencies between them. For each pair of consecutive frames, E is used to encode the previous frame to its corresponding potential representation. Then, G attempts to generate (predict) subsequent frames from the potential distribution of the previous frame. This combines history with information from previous frames to generate the next desired content. Each pair of current real training frames and synthesized frames is then forwarded to D as real data and dummy data respectively.

According to the above music theory constraints, the network parameters and variables can be set more reasonably. The output layer of E is a fully connected layer with 256 hidden units, and the first and second 128 units are respectively considered as the average μ and covariance. Used to represent the potential Z_t of dimension 128. Projecting the potential zt and a normal distribution Z_{pt} (128 dimensions) to G, outputting the synthesized information xt, applying another convolution to map the number of output channels (1 in this paper) before the Tanh layer of G. An additional convolution is applied before the Sigmoid layer of D to represent the output through a one-dimensional feature map. The network takes the two-dimensional matrices x_t and $x_{(-t)}$ as inputs and predicts whether they are real or generated MIDI information.

All models were trained with minimum batch random gradient descent (SGD) with a minimum batch size of 64. The momentum of the Adam optimizer was 0.5, the learning rate for E and G was 0.0005, and D was 0.0001.

V. Experimental Analysis

Combining the theories of melody rule model, harmony rule model and rhythm rule model, different styles of music have been successfully realized, such as pop, classical, jazz, melody + chord, etc.

As shown in the figure, the result map of the generated music tracks of different music:

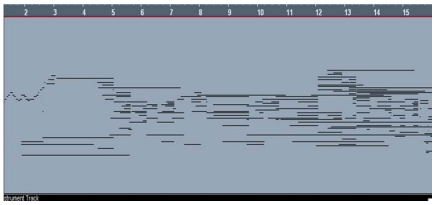

Figure3. Pop style music generation

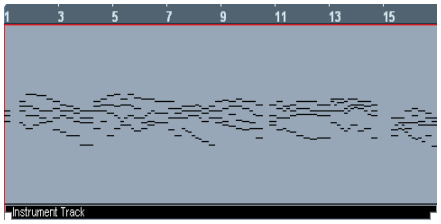

Figure4. Jazz style music generation

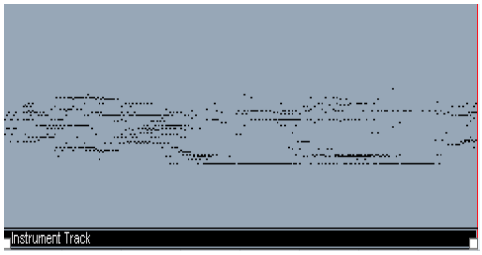

Figure5. Classical Bach Style Music Generation

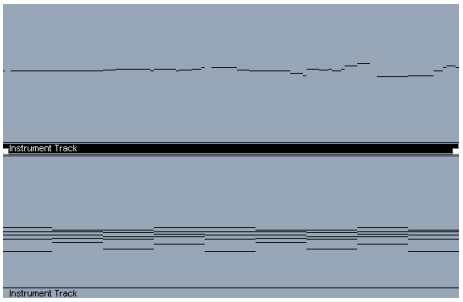

Figure6. Music melody + chord generation

1. Objective Evaluation

In the objective evaluation, we chose the music samples generated by magenta [15] to compare with the music samples that use our model to generate their rhythm sequences. We randomly generated 300 music samples from both models and import them into a digital audio work-station, then introduce a beat and downbeat tracking model [16]. This model is composed of two parts: one is to calculate the occurrence probability of rhythm and downbeat by recurrent neural network RNN, the other is to apply dynamic Bayesian network DBN to the output of RNN to make the binary decision of the occurrence of rhythm and downbeat.

Using the model to track the results, three estimates were calculated:

- The standard deviation of the beat length, which is denoted as DBEL.
- The standard deviation of the bar length (i.e. the interval between successive downbeats), which is denoted as DBAL.
- The average probability of a beat, which is denoted as APB.

The closer the evaluation data is to the actual value, the higher the quality of the rhythm sequence generated by the model is and the more practical significance is.

From DBAL, the result of our samples is closer to the real data than magenta samples, which indicates that the consistency of music is better. From APB, the result of our samples still shows a small gap, thus we can see the salience of the rhythm.

2. Subjective Evaluaion

We posted the resulting rhythm audios on a test site and invited a number of professionals and music lovers to subjectively assess the salience, consistency, audibility, and fluency of the rhythm. We asked participants to listen to five sets of rhythm audios generated by the two models and rate each item on a scale of 10 based on their experience. Average the results of the scores collected, and the results are shown in table 2.

According to the result of score, the musical rhythm generated by the model in this paper has some advantages.

TABLE I OBJECTIVE EVALUAION DATA

Music Samples	DBEL	DBAL	APB
Magenta Samples	0.0635	0.1072	0.1638
Our Samples	0.0628	0.1847	0.2389
Real Data	0.0589	0.2177	0.2043

TABLE II SUBJECTIVE EVALUAION SCORE

Music Samples	Salience	Consistency	Audibility	Fluency
Magenta Samples	6.73	7.32	6.89	7.26
Our Samples	7.59	7.47	7.91	7.44

VI. SUMMARY

This paper introduces an RVAE-GAN model for generating sequence data. The model consists of three networks (encoder, generator and discriminator) in which the local correlation of the data in each frame is spatially learned using convolutional neural networks and constraints based on music theory knowledge. Each frame is sampled by using a potential distribution obtained by the encoder mapping the previous frame. The consistency between the frames in the sequence of the final generated results is relatively consistent. Compared with the prior art, in this paper, the experiment of piano music automatic generation has achieved good results in rhythm and melody, and the network structure and constraints have been improved to make the music automatic generation better.

ACKNOWLEDGMENT

This paper is supported by “the Fundamental Research Funds for the Central Universities”. This document is the results of the research project funded by the National Natural Science Foundation of China (Grant No. 61631016 and 61901421), National Key R\&D Program of China (Grant No. 2018YFB1403903) and supported by the Fundamental Research Funds for the Central Universities (Grant No. CUC200B017, 2019E002 and CUC19ZD003).

REFERENCES

[1] Alpan A.Techniques for alogorithmic composition of music.1995.http://alum.Hampshire.edu/~adaF92/algocomp/algoeomp95 .html

[2] Basset BA， Neto JJ. A stochastic musical composer based on adaptive algorithms.1999.http://gsd.ime.usp.br/sbcm/1999/papers/Bruno_Basseto.pdf

[3] Bartetzki A. CMask, a stochastic event Generator for Csound. 1997.http://gnom.kgw.tu-berlin.de/abart/CMaskMan/CMask-Manual.html

[4] Ames C, Domino M. Cybernetic composer: an overview. In: Balaban M, Ebcioglu K, Laske O, eds. Understanding Music with AI. Cambridge: AAAI Press, 1992. 186-205.

[5] Datseris, G., Ziereis, A., Albrecht, T. et al. Microtiming Deviations and Swing Feel in Jazz. Sci Rep 9, 19824 ,2019.

[6] Tao Shen.: Methods of rhythm control in algorithmic composition [D]. Wuhan Conservatory of Music,2010.

[7] Jiewen Zhu. The use of rhythm in music creation [J]. Drama and Film Journal, 2009, 000(002):97.

[8] Yi L, Goldsmith J. Automatic Generation of Four-part Harmony[C]// UAI Bayesian Modeling Applications Workshop. DBLP, 2009.

[9] M. I. BELLGARD, C. P. TSANG. Harmonizing Music the Boltzmann Way[M]// Musical networks. MIT Press, 1999:281-297.

[10] YK. Au expert system for harmonizing chorales in the style of J.S. Bach. In:Balaban M, Ebcioglu K, Laske O, eds. Understanding Music with A1.Cambridge:AAAI Press,1992. 294-334.

[11] GMozer MC. Neural network composition by prediction: Exploring, the benefits of psychophysical constraints and multiscale processing.Cognitive Science,1994,6:247-280.

[12] Adam Roberts;Jesse Engel;Colin Raffel;Curtis Hawthorne;Douglas Eck. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. Sound. 2018.

[13] Mohammad Akbari, Jie Liang.Semi-Recurrent CNN-Based VAE-GAN for Sequential Data Generation. IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.2321-2325.

[14] Raffel, C., Ellis, D.P.: Intuitive analysis, creation and manip-ulation of midi data with pretty midi. In: 15th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers, pp. 84–93. Taipei, Taiwan ,2014.

[15] Magenta: Make Music and Art Using Machine Learning. https://magenta.tensorflow.org/. 2016.

[16] Sebastian Bo¨ck, Florian Krebs, and Ger- hard Widmer. Joint beat and downbeat tracking with recur- rent neural networks. In International Society for Music Information Retrieval, pag-es 255–261, 2016.