11-2021

Generating music with sentiments

Chunhui BAO
Singapore Management University, chbao.2019@phdcs.smu.edu.sg

GENERATING MUSIC WITH SENTIMENTS

BAO CHUNHUI

SINGAPORE MANAGEMENT UNIVERSITY
2021

Generating Music with Sentiments

BAO Chunhui

Submitted to School of Computing and Information Systems
in partial fulfillment of the requirements for the
Degree of Master of Philosophy in Information Systems

Master's Thesis Committee:

SUN Qianru (Supervisor / Chair)
Assistant Professor of Computer Science
Singapore Management University

GAO Wei
Assistant Professor of Computer Science
Singapore Management University

NGO Chong Wah
Professor of Computer Science
Singapore Management University

Singapore Management University
2021

I hereby declare that this Master's thesis is my original work
and it has been written by me in its entirety.
I have duly acknowledged all the sources of information
which have been used in this thesis.

This Master's thesis has also not been submitted for any
degree in any university previously.

Bao Chunhui

BAO Chunhui
20 October 2021

Generating Music with Sentiments

BAO Chunhui

Abstract

In this thesis, I focus on the music generation conditional on human sentiments such as positive and negative. As there are no existing large-scale music datasets annotated with sentiment labels, generating high-quality music conditioned on sentiments is hard. I thus build a new dataset consisting of the triplets of lyric, melody and sentiment, without requiring any manual annotations. I utilize an automated sentiment recognition model (based on the BERT trained on Edmonds Dance dataset) to "label" the music according to the sentiments recognized from its lyrics. I then train the model of generating sentimental music and call the method Sentimental Lyric and Melody Generator (SLMG). Specifically, SLMG is consisted of three modules: 1) an encoder-decoder model trained end-to-end for generating lyric and melody; 2) a music sentiment classifier trained on labelled data; and 3) a modified beam search algorithm that guides the music generation process by incorporating the music sentiment classifier. I conduct subjective and objective evaluations on the generated music and the evaluation results show that SLMG is capable of generating tuneful lyric and melody with specific sentiments.

Index Terms: Conditional Music Generation; Seq2Seq; Beam Search; Transformer

Contents

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor SUN Qianru, whose expertise was invaluable in formulating the research questions and methodology. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank my friends in Singapore Management University. I could not have completed this thesis without the support of my friends, CHEN Zhaozheng and YU Sicheng, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me.

Dedication

This thesis is dedicated to my dad and mom, who taught me to perform all of life's tasks, no matter big or small, to the best of my ability and without complaint.

Chapter 1

Introduction

Summary

A brief introduction of this work is illustrated in this chapter. The research background of deep learning based music generation is introduced in Section 1.1. The problem statement is given in Section 1.2, in which the specific objectives of this work are demonstrated. In Section 1.3, the main contributions for solving the problem are introduced, which to my best knowledge, is the first attempt to automatically generate both lyric and melody with a specific given sentiment using artificial neural networks.

1.1 Research Background

Music is the art of arranging sounds in time, which can be used to express human sentiments. In contemporary pop music, sentiments are mainly conveyed by lyric and melody. Melody is a temporal sequence consisting of musical notes, and lyric is natural language representing music themes.

Melody and lyric provide complementary information in understanding human beings' sentiments in songs.

In recent years, deep learning has made great progress in sequential data generation tasks, such as natural language [1], audio [2], as well as music [3; 4]. Music generation is a human creation activity with a long history to express human sentiments. Musicians composite pleasing sounds according to professional music knowledge, such as harmonious relationships between pitch, duration, velocity, and tempo. Benefit from the advent of large music datasets, such as LMD-full MIDI Dataset [5] and reddit MIDI dataset [6], deep learning models recently have been used to "composite" high-quality music [7; 8; 9]. Mainstream works are focused on how to generate human-like music to be real enough such that listeners can not distinguish whether it is created by human composers or generated by deep learning models [9; 10; 11]. However, they do not care whether generated music represents human sentiments.

It has been said by Hagel that music is the art of mood. However, due to the insufficiency of music datasets annotated with sentiment labels, training a deep learning model to generate music with specific sentiments is difficult. Music generation with sentiments is still unexplored well and a challenging problem in deep learning area.

In this work, I focus on the problem of how to generate lyric and melody with a specific given sentiment. By solving this problem, deep learning models can be used to help human beings composite music with sentiments even if they don't have professional music knowledge.

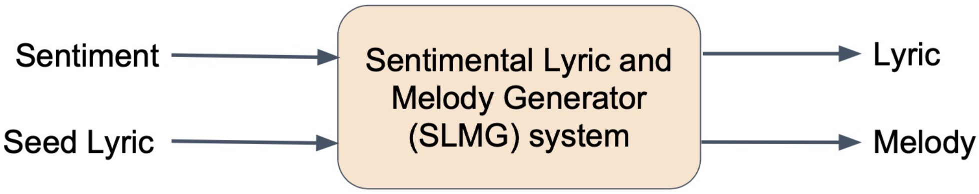


Figure 1.1: A brief introduction of the goal of this work, which takes a specific given sentiment (positive or negative) and a piece of seed lyric as input, output a music segment consists of melody and lyric.

1.2 Problem Statement

As shown in Figure 1.1, the goal of this work is to design a deep learning system which takes a specific given sentiment (positive or negative) and a piece of seed lyric as input, output a music segment consists of melody and lyric. Formally, the problem can be defined as

$$(S, M) = SLMG(e, Seed), \tag{1.1}$$

the required sentiment e and seed lyric $Seed = \{s_1, s_2, ..., s_t\}$ are input to the Sentimental Lyric and Melody Generation (SLMG) system, and a music segment perceived to have the given sentiment is output, which consists of lyric $S = \{s_1, s_2, ..., s_t, ..., s_T\}$ and melody $M = \{m_1, m_2, ..., m_T\}$. I hope that this work can help people who doesn't have professional music knowledge composite music to express their sentiments.

1.3 Main Contributions

The first difficulty in generating music with sentiment is that there is no large-scale dataset available. In 2019, Ferreira et al. [12] built a music dataset annotated by volunteers called VGMIDI, which composed of 95 labelled piano pieces and 728 unlabelled pieces. Then, they proposed a deep generative model which was able to control the polyphonic music generation with a given sentiment. In 2020, Ferreira et al. [13] expanded the VGMIDI dataset from 95 to 200 labelled pieces and presented a system called Bardo Composer, which used a GPT-2 [14] model to generate music with sentiments for role-playing games. More recently, a symbolic music dataset called EMOPIA was constructed by Hung et al. [15], in which there are 1078 music clips from 387 songs with Valence-Arousal emotion labels. Nevertheless, the number of labelled data in both VGMIDI and EMOPIA are too small to generate music with high quality, and they only pay attention to the melody. We all know that not only the melody, but also the lyric of music is an important part for people to express their feelings and sentiments [16; 17]. How to generate lyric and melody with specific sentiments is a difficult problem that has not been researched well.

In this thesis, I focus on the problem of how to generate lyric and melody with a specific given sentiment. To the best of my knowledge, there is no paired lyric-melody dataset annotated according to sentiments. Based on the dataset created by Yu et al [18], I build a new paired lyric-melody dataset with sentiment labels, which composed of 11528 songs with English lyrics. I cut each song into segments with fixed length, and then a Bert [19] model fine-tuned on Edmonds Dance dataset [20] is used to annotate these segments. The details of the dataset construction are introduced in Chapter 3. Because there are more than 170000 segments, manually annotation is costly. Using deep learning models trained for natural language sentiment classification is an optional method, but the disadvantage is that melody is ignored in the annotation process.

The second difficulty is that how to design the algorithm to generate lyric and melody with a specific given sentiment. Inspired by the great success of deep learning techniques for lyric and melody generation, and variations of beam search algorithms for controlling the generation process [21; 22], I propose the Sentimental Lyric and Melody Generator (SLMG) system, which to my best knowledge, is the first attempt to automatically generate both lyric and melody with a specific given sentiment using artificial neural networks. SLMG consists of the following three parts: 1) Lyric and melody generator: a novel encoder-decoder architecture that can generate lyric and melody by accepting a small piece of initial seed lyric as input. 2) Music sentiment classifier: a classifier for lyric and melody segments. 3) Sentimental beam search (SBS) algorithm: a modified beam search algorithm for controlling the music generation process with a given sentiment.

My contributions are thus the following four-fold:

- A large-scale paired lyric-melody dataset with sentiment labels consisting of 11528 MIDI songs is built.
- Both GRU and Transformer based encoder-decoder networks are trained to generate lyric and melody.
- A modified beam search algorithm SBS is proposed to bias the music generation process to match a particular sentiment.
- The subjective and objective evaluations are combined to verify the effectiveness of the proposed SLMG system.

Chapter 2

Related Works

Summary

The lecture review related to this work is given in this chapter. Firstly, the previous research works of symbolic music composition are introduced in Section 2.1. Then, a series of works for generating music with a given sentiment are shown in Section 2.2, in which a dataset called VGMIDI is built by Ferreira et al. and a dataset called EMOPIA is constructed by Hung et al. Finally, a series of works for generating music with lyrics are introduced in Section 2.3.

2.1 Symbolic Music Composition

With the advent of large music datasets, deep learning models have recently achieved high-quality results in music composition tasks. DeepBach [23] is proposed by Hadjeres et al., which uses a dependency network and a Gibbs-like sampling procedure to generate Bach's four parts chorales,

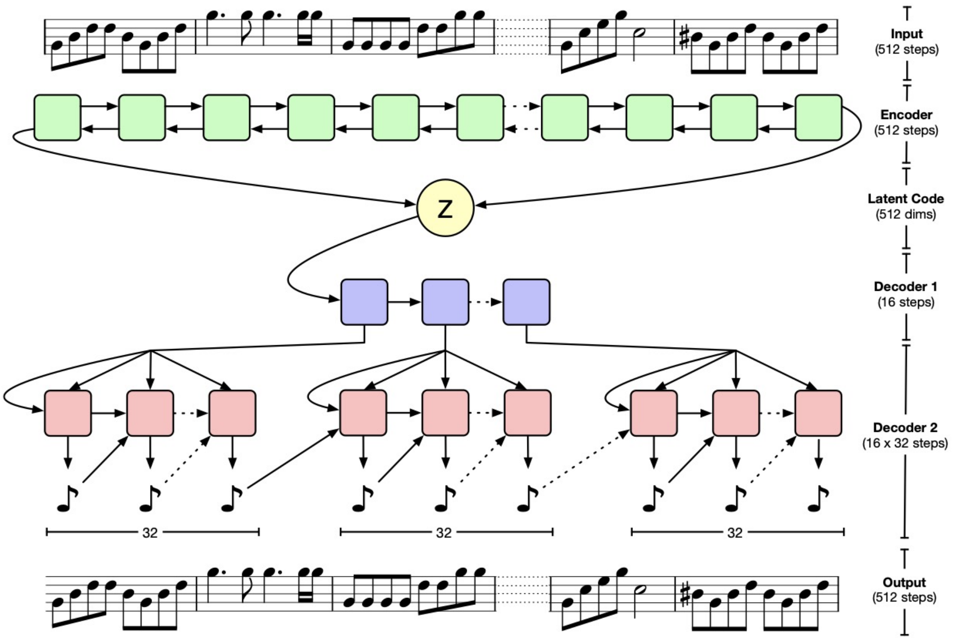


Figure 2.1: The architecture of recurrent hierarchical melody VAE.

including the lists of melody, rhythm and fermatas. Roberts et at. proposed a recurrent variational auto-encoder (VAEs) [24] model to reproduce polyphonic music sequences. The architecture of recurrent hierarchical melody VAE is shown in Figure 2.1, which uses recurrent neural network (RNN) as encoder and decoder for music generation. In addition to RNN, convolutional neural network (CNN) has also been successfully applied to the field of music generation, in which music is represented similar to images as shown in Figure 2.2. As shown in Figure 2.3, MuseGAN [7] is a Generative adversarial network (GAN) [25] based architecture to compose polyphonic music with 5 sound-tracks. RNN based GAN is used in C-RNN-GAN [26] model to generate polyphonic continuous music sequence. However, these models mainly trained for generating human-like music, the sentiment expression of generated music was ignored. In this work, I focus on how to generate music with a specific given sentiment.

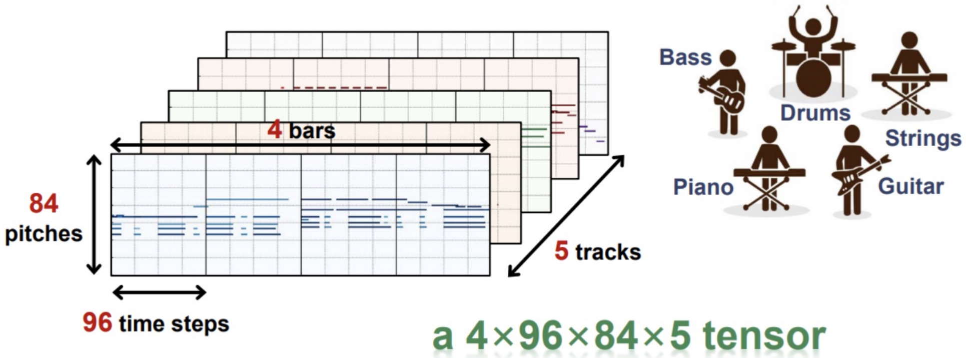
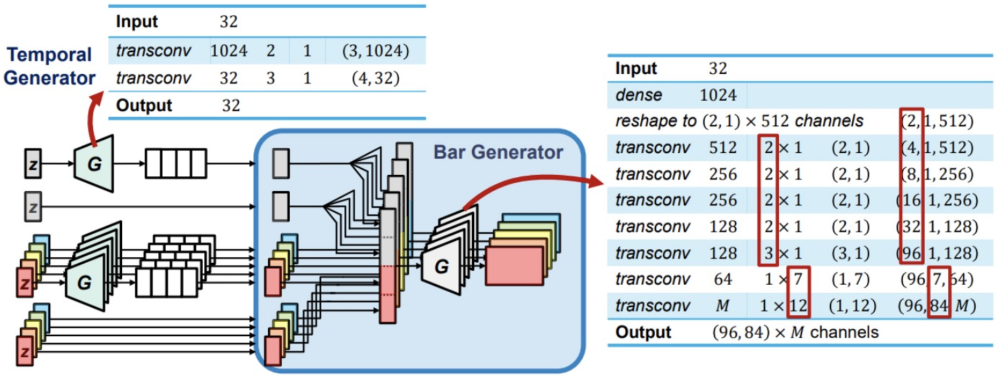


Figure 2.2: Data representation for MuseGAN.



Input	32			
transconv	1024	2	1	(3, 1024)
transconv	32	3	1	(4, 32)
Output	32			

Input	32			
dense	1024			
reshape to (2, 1) × 512 *channels*				(2, 1, 512)
transconv	512	2 × 1	(2, 1)	(4, 1, 512)
transconv	256	2 × 1	(2, 1)	(8, 1, 256)
transconv	256	2 × 1	(2, 1)	(16 1, 256)
transconv	128	2 × 1	(2, 1)	(32 1, 128)
transconv	128	3 × 1	(3, 1)	(96 1, 128)
transconv	64	1 × 7	(1, 7)	(96, 7, 64)
transconv	M	1 × 12	(1, 12)	(96, 84 M)
Output	(96, 84) × M channels			

Figure 2.3: Neural architecture and parameter settings for MuseGAN.

2.2 Generate Music with a Given Sentiment

Music is a way for humans to express their sentiments. However, it is too expensive to manually annotate sentiment labels for music datasets, which causes great difficulties for music generation tasks conditioned on sentiments. In 2019, Ferreira et al. [12] proposed a deep generative model based on mLSTM, which was the first work to explore deep learning models for symbolic music sentiment analysis. They also built a new music dataset with manually sentiment labels called VGMIDI, which consists of 95 labelled piano pieces and 728 unlabelled pieces. In 2020, a GPT-2 model was used by Ferreira et al. [13] to generate music with a spe-

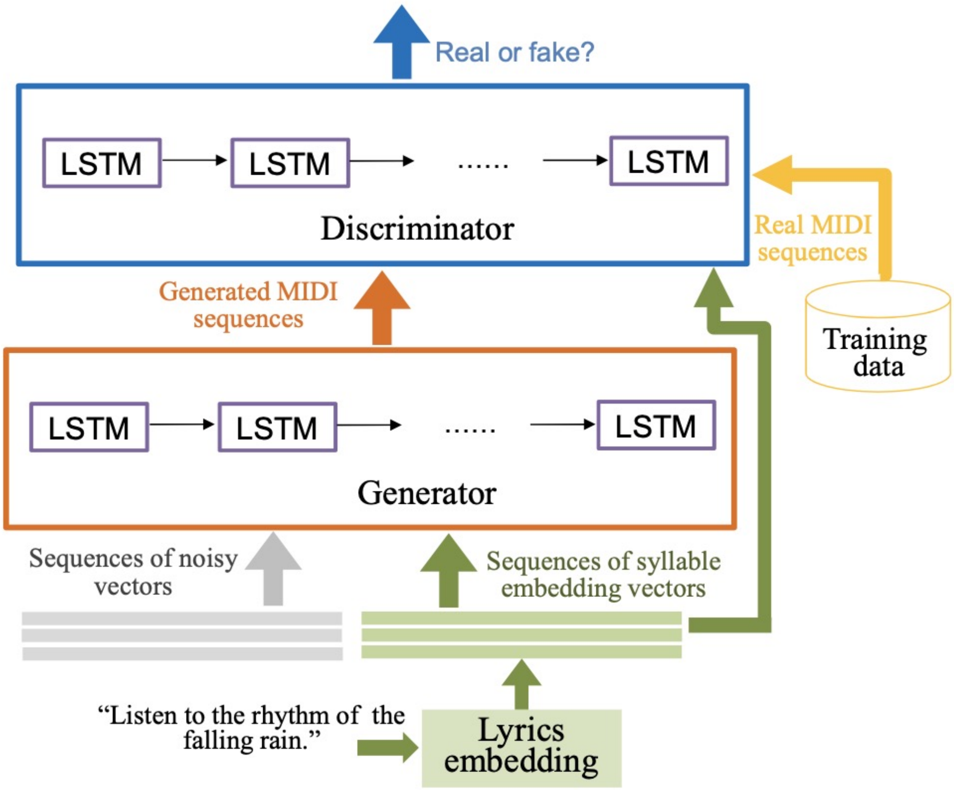


Figure 2.4: Conditional LSTM-GAN for melody generation from lyrics.

cific sentiment and the VGMIDI dataset was extended to 200 labelled data. In [27], a model called CVAE-GAN was proposed for sentiment-conditioned symbolic music generation, which synthesized Conditional-VAE and Conditional-GAN [28]. More recently, Hung et al. built an emotion-labeled symbolic music dataset called EMOPIA [15], which consists of 1078 music clips from 387 songs. They also verified that the proposed dataset can be used for generating music conditioned on emotions. Nevertheless, existing music datasets with sentiment labels are both small in size. Therefore, I create a new large-scale paired lyric-melody dataset with sentiment labels for generating harmonious music that can evoke sentiments.

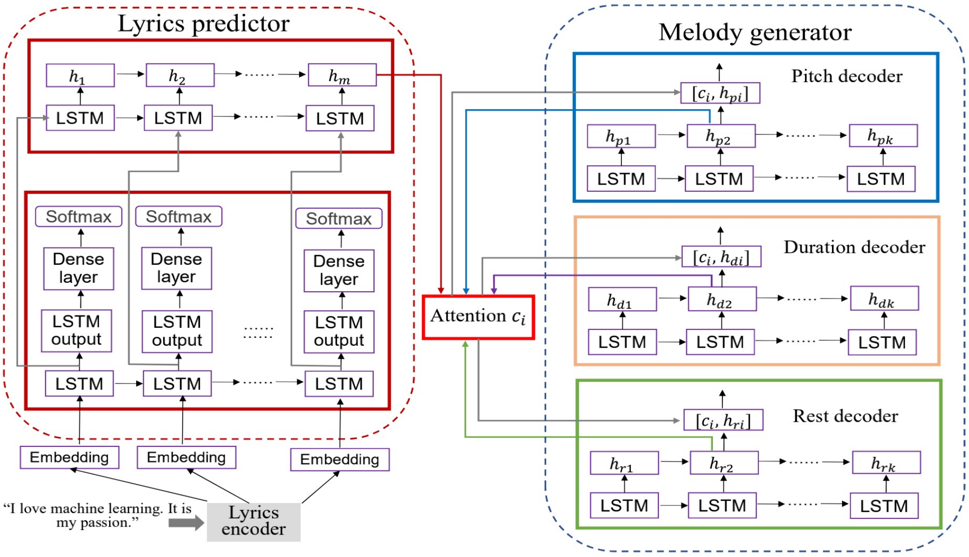


Figure 2.5: Graphical representations of AutoNLMC's neural network architecture.

2.3 Generate Music with Lyrics

In recent years, with the advent of music datasets with lyrics, deep learning was also researched for mining musical knowledge between lyrics and melodies. Songwriter proposed by Bao et al. [29] focused on lyric-conditional music generation. They first divide the input lyric into sentences, then use the seq2seq-based model to generate melody from the input lyric, and finally merge these segments into a complete melody. As shown in Figure 2.4, Yu et al. [18; 30] utilized conditional LSTM-GAN to generate melody from given input lyric, in which the generator and discriminator were LSTM networks with lyric as condition. AutoNLMC [31] proposed by Madhumani et al. can create songs with both lyrics and melodies automatically. It was an encoder-decoder LSTM network where the encoder was designed to generate lyric and three decoders are trained to generate pitch, duration and rest of melody respectively, whose architecture is shown in

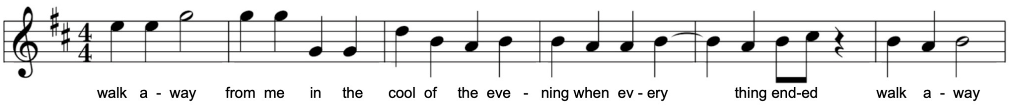
Figure 2.5. Jukebox [9] trained on raw audio data can also generate music with lyrics. In this work, I propose a novel encoder-decoder architecture for lyric and melody generation. The melody is represented to a sequence of tokens and only one decoder is trained to generate melody, which can be easily controlled to match a particular sentiment.

Chapter 3

Dataset Construction

Summary

There is no large-scale music dataset with sentiment labels publicly available for sentiment-conditioned music generation. Therefore, it is valuable to build a large-scale music dataset, the detailed method for building the new dataset used to generate lyric and melody with sentiments will be introduced in this chapter. There are many different ways to represent music for deep learning, inspired by Yu et al. [18; 30] and Madhumani et al [31], each note of the music is represented as a four-dimensional tuple $n = (n_{syllable}, n_{pitch}, n_{dur}, n_{rest})$, the detailed form of music representation of this work is introduced in Section 3.1. The basic information of the paired lyric-melody English songs dataset is introduced in Section 3.2, which are collected from LMD-full MIDI Dataset [5] and reddit MIDI dataset [6]. The method that I used to annotate music is introduced in Section 3.3. The detailed analysis of the annotated dataset is given in Section 3.4.



Pitch	80	80	83	83	83	71	71	78	75	73	75	75	73	73	75	73	75	77	75	73	75
Duration	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	0.5	0.5	1.0	1.0	2.0
Rest	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Figure 3.1: An example of paired lyric-melody music data. Each note of the music is represented as a four-dimensional tuple $n = (n_{syllable}, n_{pitch}, n_{dur}, n_{rest})$.

3.1 Data Representation

Inspired by Yu et al. [18; 32] and Madhumani et al. [31]. I represent music as a sequence of syllable-note pairs. As shown in Figure 3.1, lyric as natural language sentences are made up of words. English words are made up of one or more syllables, for example, “do” is made up of one syllable “do” and “doing” is made up of two syllables “do” and “ing”. Melody can be defined as a sequence of musical notes. Each note of the melody is represented as a three-dimensional tuple $n = (n_{pitch}, n_{dur}, n_{rest})$:

- n_{pitch}: in music, the pitch is what decide of how the note should be played, it can take any integer from 0 to 127.
- n_{dur}: the length of time that a note is played, which depends on the note type. The standard unit is one beat, if the duration of a note is one beat, denote its duration as 1.0.
- n_{rest}: the duration of the rest before the note. 0.0 means no rest before the note.

Therefore, music segments with length N can be defined as $M = \{m_1, m_2, ..., m_N\}$, where each m_i is a (syllable, pitch, duration, rest) four-dimensional tuple. For simplicity, I do not consider the velocity and tempo of the music. And suppose that the lyrics and melodies can be paired as one-syllable-to-one-note.

3.2 Data Collection

The dataset used in this work initially created in [18], which comes from two large-scale MIDI music datasets: LMD-full MIDI dataset [5] and reddit MIDI dataset [6]. MIDI is the abbreviation of musical instrument digital interface, which is an industry standard that describes the interoperability protocol between various electronic instruments, software and devices. The MIDI file records all the information of the music and saves it on the computer. The LMD-full dataset contains a total of 176581 different MIDI files, but most of them do not contain lyrics. In this work I only use the files with sufficient English lyrics, so only 7497 MIDI files could be used. Similarly, the reddit MIDI dataset contains 130000 different MIDI files but only 4031 with enough English lyrics could be used. Altogether there are 11528 MIDI files in the dataset. Paired lyric-melody sequences are obtained by parsing the MIDI files as follows:

- Open the file, find out the beginning of the lyric and its corresponding note.
- If a note has a corresponding English syllable, its pitch, duration and rest are stored.
- If a syllable corresponds to multiple notes, only the information of the first note is recorded.

After parsing, there are 1971257 notes in total and the average length of music segments is 171 notes. The pitch distribution of these selected songs is shown in Figure 3.2a, from which we can see that the pitch distribution approximately obeys a normal distribution with a mean of 66.58 and a standard deviation of 9.96. Similarly, the duration distribution is shown in Figure 3.2b, we can observe that most of them fall in the interval [0.5, 2.0], and the mode is 1.0. Rest distribution is shown in Figure 3.2c, we can observe that most of the rests are zero. For the lyrics, there are 20934 unique syllables and 20268 unique words in total.

3.3 Data Annotation

For the above large-scale dataset, manually labelling sentiments expressed in music by humen is expensive. Therefore, in this work I exploit the deep learning models to automatically annotate the paired lyric-melody dataset. There are many datasets that can be used to train the annotator, such as large-scale social media or dialog datasets with emotion labels [33], relatively small-scale lyric datasets for lyric sentiment classification [17; 20; 34] and small-scale sentiment-labelled music datasets without lyric [12; 15]. In this section, I explore the reliable method to train the annotator.

Understanding emotions expressed in natural language has been widely researched in resent years. The largest human annotated dataset for text emotion classification is GoEmotions [33], which consists of 58000 carefully selected Reddit comments and labelled for 27 emotion categories or neutral. Table 3.1 shows illustrative samples of GoEmotions dataset, each sample text has one or more corresponding labels.

For music emotion analysis, I firstly used fine-grained emotions classifiers to annotate the dataset, which annotates lyrics to Ekman's 6-emotion model [35; 36]: anger, disgust, fear, joy, sadness and surprise. However, on the fine-grained annotated dataset, the number of data of different labels is very unbalanced. I selected some data segments to manually label them, most of the labelling results disagree with the annotation results. And I also cannot train a good classifier on this dataset with high-accuracy. Since both humans and deep learning models cannot successfully grasp the difference between different labels, I had to simplify the emotion model to binary sentiment groups: positive or negative. According to binary group method proposed by the authors of GoEmotions [33], the labels are divided into 4 categories as shown in follows:

- **positive:** admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
- **negative:** anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
- **ambiguous:** confusion, curiosity, realization, surprise
- **neutral**

The advantage of GoEmotions dataset is its large scale, but the disadvantage is that there's a domain gap between Reddit comments and song lyrics.

There's some relatively small-scale lyric datasets manually labelled according to human emotions. Recently, Edmonds et al. constructed Edmonds Dance dataset [20], which consists of lyrics retrieved from 524 English songs. As shown in Table 3.2, there's 8 emotion categories in the

Sample Text	Label(s)
You know the answer man, you are programmed to capture those codes they send you, don't avoid them!	annoyance, surprise
I've never been this sad in my life!	remorse
I don't necessarily hate them, but then again, I dislike it when people breed while knowing how harsh life is.	anger, disappointment
You're right. Sorry for the poor reply.	relief
Absolutely. I'd love it. No matter how much I like the guy, if he just goes for it that's not cool.	embarrassment, joy

Table 3.1: Examples from GoEmotions dataset.

Sample Text	Label(s)
Just one day in the life. So I can understand. Fighting just to survive. But you taught me I can. We are the lucky ones. We are...	joy, trust, surprise
Hypnotized, this love out of me. Without your air I can't even breathe. Lead my way...	joy, trust
You ruined my life. What you said in your message that night. left me broken and bruised but now i know that you were wrong...	sadness, disgust, anger

Table 3.2: Examples from Edmonds Dance dataset.

Edmonds Dance Dataset and each song has one or more corresponding labels. Same as GoEmotions, the 8 categories are grouped into positive, negative or ambiguous:

- **positive:** anticipation, joy, trust
- **negative:** anger, disgust, fear, sadness
- **ambiguous:** surprise

In order to have a common model for sentiment classification, I train Bert-base [19] models on GoEmotions and Edmonds Dance Dataset. Bert stands for Bidirectional Encoder Representations from Transformers [37], which has been pre-trained on Wikipedia and BooksCorpus and given

Train dataset	Acc	Precision	Recall	F1 score
GoEmotions	52.44	45.50	55.26	49.93
Edmonds Dance	77.90	81.82	80.67	81.23
Both	**79.02**	**82.85**	**81.88**	**82.31**

Table 3.3: Classification results (%) of Bert models trained on GoEmotions dataset and Edmonds Dance dataset, tested on Edmonds Dance dataset. "Both" means first trained on GoEmotions dataset and then fine-tuned on Edmonds Dance dataset.

state-of-the-art results on a wide variety of natural language processing tasks. When train the Bert model, the learning rate is set to 5e-5 with gradually decay. The model fine-tuned for 10 epochs with the warm-up proportion as 0.1 and batch size as 16. Because there's no domain gap between Edmonds Dance Dataset and my dataset, I randomly select 1/10 data from the Edmonds Dance Dataset as test data. The experimental results are shown in Table 3.3, to my surprise, the Bert model trained on GoEmotions dataset has relatively worse performance for lyric emotion classification. It means that the emotion classifiers trained on large-scale out-of-domain data do not generalize well to song lyrics. However, the Bert model directly trained on Edmonds Dance Dataset achieves better performance, despite the in-domain dataset is magnitude smaller than out-of-domain dataset. In addition, pre-training the Bert model on GoEmotions dataset and then fine-tuning the model on Edmonds Dance Dataset can slightly improve the classification accuracy of song lyrics.

In addition to lyrics, is there any way that can utilize the melodies for annotation? In order to answer this question, I train deep learning models on the EMOPIA dataset [15] and evaluate if they can be used on my dataset. The EMOPIA dataset consists of 1078 clips from 387 piano solo performances. They are labelled corresponding to the Russell's

Length	**Annotations**		**Total**
	High-valence	**Low-valence**	
20	25743	77797	103540
50	7069	36833	43902
100	2699	20163	22862

Table 3.4: Annotation results of the Bi-LSTM trained on EMOPIA, all segments are labelled to High-valence or Low-valence.

2-dimensional model [38], which represents music emotion using a valence-arousal pair. Arousal indicates emotion intensity and valence indicates the positive or negative sentiment. Thus, the clips with high valence label can be considered as positive data and the clips with low valence label are negative data. I train a bidirectional LSTM with self-attention to classify the music clips according to their valence, and achieves 83.3% test accuracy on EMOPIA dataset. Then, this model are used to classify the melodies of my dataset, the results are shown in Table 3.4, we can see that the classification results are catastrophically unbalanced, even though the training data in EMOPIA dataset is balanced. I also manually verify randomly selected data of the classification results, the unanimous ratio is less than 50%. Therefore, I think the deep learning model trained on EMOPIA cannot be used to annatate my dataset because of the following reasons: 1) There's a domain gap between piano solo performances and pop songs' melodies in my dataset. 2) EMOPIA is a small-scale dataset. 3) The unanimous ratio of automatic labelling and manual labelling is less than 50%.

Length	**Annotations**			**Total**
	positive	**negative**	**unlabelled**	
20	48659	17019	37862	103540
50	18557	9341	16004	43902
100	8968	5712	8182	22862

Table 3.5: Annotation results for my dataset, all segments are labelled to positive, negative or unlabelled.

Sample Text	**Label**
When I look into your eyes your love is there for me And the more I go inside the more there is to see	positive
I believe in angels Something good in everything I see I believe in angels When I know the time is right for me	positive
Please forgive me I stop loving you deny me this pain going through Please forgive me I need you	negative
Please forgive me I know not what I do Please forgive me I stop loving you deny me this pain	negative
Quit playing games with my heart With my heart my heart I should have known from the start	unlabelled

Table 3.6: Examples from annotated dataset.

3.4 Data Analysis

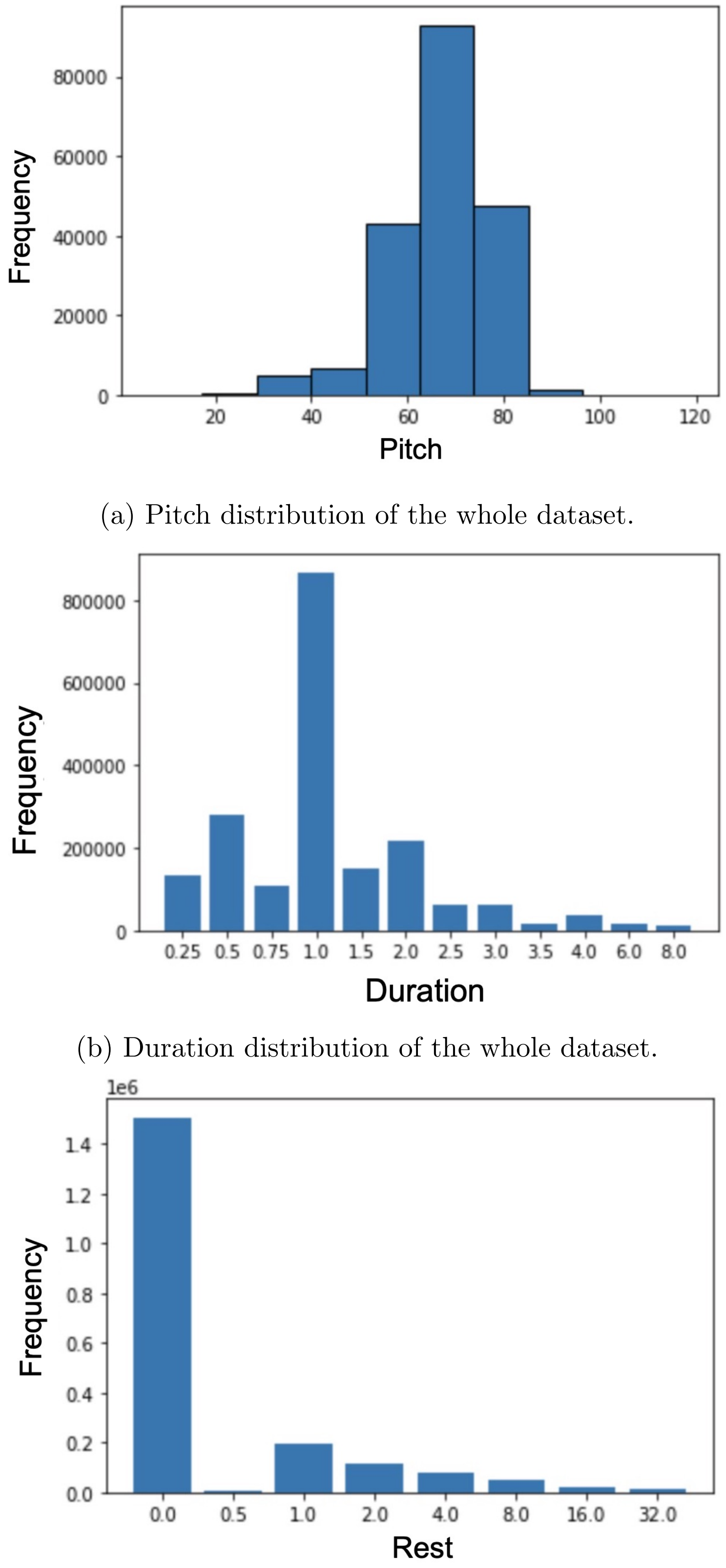
The 11528 MIDI files are cut into small music segments with fixed length N (20, 50 or 100), and gets 103540, 43902, 22862 segments respectively for N equals to 20, 50, or 100. Then, the Bert model trained on GoEmotions dataset and then fine-tuned on Edmonds Dance Dataset is used to annotate these music segments. Specifically, if the music segment is classified as [positive] or [negative] and the confidence is greater than 95%, mark it as positive or negative; if the music segment is classified as [positive, ambiguous] or [negative, ambiguous], the confidence of positive or negative

Items	WD	Positive	Negative
Mean value of pitch	66.58	66.34	66.63
Standard deviation of pitch	9.96	9.98	10.11
Number of pitch value	98	85	79
Maximum pitch value	108	102	105
Minimum pitch value	3	7	21
Mode of duration	1.0	1.0	1.0
Number of duration value	19	19	18
Maximum duration value	32.5	32.5	32.0
Minimum duration value	0.25	0.25	0.25
Percentage of 1.0	45.17	45.68	48.83
Mode of rest	0.0	0.0	0.0
Number of rest value	8	8	8
Maximum rest value	32.0	32.0	32.0
Percentage of 0.0	80.25	83.55	86.66
Percentage of major keys	**57.73**	**61.59**	**55.83**
Percentage of minor keys	**42.27**	**38.41**	**44.17**

Table 3.7: Detailed quantitative comparison of melody distributions, include the whole dataset (WD), positive and negative.

is greater than ambiguous and surpass 95%, mark it as positive or negative; if the music segment is classified as [ambiguous], [positive, negative] or [positive, negative, ambiguous], this music segment is unlabelled. if the music segment is classified as [positive], [negative], [positive, ambiguous] or [negative, ambiguous], but the confidence of positive or negative is less than 95% or ambiguous, this music segment is unlabelled.

Table 3.5 shows the annotation results for my dataset, from which we can see that about 64% are labelled, and the number of positive segments is larger than the number of negative segments. Examples form the annotated dataset are shown in Table 3.6. Detailed quantitative comparison of melody distributions is shown in Table 3.7, it shows that the pitch, duration and rest distributions of positive and negative samples are pretty similar to the whole dataset. I also measure the major-minor tonality of the music segments by using Krumhansl-Kessler algorithm [39]. We can see that the major-minor tonality distributions of positive data and negative data are a little bit different, which indicates that when people create sentimental-positive music, they prefer to use major keys, but when they create sentimental-negative music, more minor keys are used.



(a) Pitch distribution of the whole dataset.

(b) Duration distribution of the whole dataset.

(c) Rest distribution of the whole dataset.

Figure 3.2: Melody distribution of the collected dataset. (a), (b) and (c) show the distribution of pitch, duration and rest of the whole dataset respectively.

Chapter 4

Methodology

Summary

In this work, the sentimental Lyric and Melody Generator (SLMG) system is designed to generate lyrics and melody with required specific sentiment given a piece of seed lyric. A general overview is shown in Algorithm 1 and Figure 4.2. It receives the labelled and unlabelled music segments, a required sentiment and a piece of seed lyric as input. Firstly, skip-gram models are trained on the whole dataset, which aim at mapping each English word, syllable and music note to a vector representation [40], the details are introduced in Section 4.1. Then, an encoder-decoder model is trained end-to-end as the lyrics and melody generator, in which the encoder is lyrics generator and the decoder is melody generator, its structure is illustrated in Section 4.2. Next step, a music sentiment classifier is trained on the labelled data, which is demonstrated in Section 4.3. Finally, an sentimental beam search (SBS) algorithm is proposed in Section 4.4, it takes the required sentiment, lyrics and melody generator, music sentiment

Algorithm 1 Sentimental Lyric and Melody Generator (SLMG)

```
Require: labelled and unlabelled dataset $X_l$ and $X_u$, required sentiment $e$, piece of seed lyric $m$
Initialize word embedding $E_w$
Initialize syllable embedding $E_s$
Initialize music note embedding $E_m$
for $x \in X_l \cup X_u$ do
    Update $E_w$, $E_s$ and $E_m$.
end for
Initialize lyric and melody generator $G$
for $x \in X_l \cup X_u$ do
    Update $G$
end for
Initialize music sentiment classifier $C$
for $x \in X_l$ do
    Update $C$
end for
$y \leftarrow \texttt{SBS}(G, C, m, e)$
return $y$, $E_w$, $E_s$, $G$, $C$
```

classifier and a piece of seed lyric as input and output a music segment.

4.1 Skip-gram Embedding Models

In this work, the English sentences are divided into syllable-level. Different from fastText that decompose words into sub-words based on morphology [41], the syllables of each word are divided according to its pronunciation. Therefore, in order to represent semantic information of each syllable, I train two skip-gram models to obtain the vector representation of the words and syllables respectively.

Given a piece of input text, the skip-gram model architecture tries to predict the surrounding words of the center word. Take Figure 4.1 as an example, suppose the center word is w_t and tokens context window $c = 2$, the conditional probability of the surrounding words can be expressed as

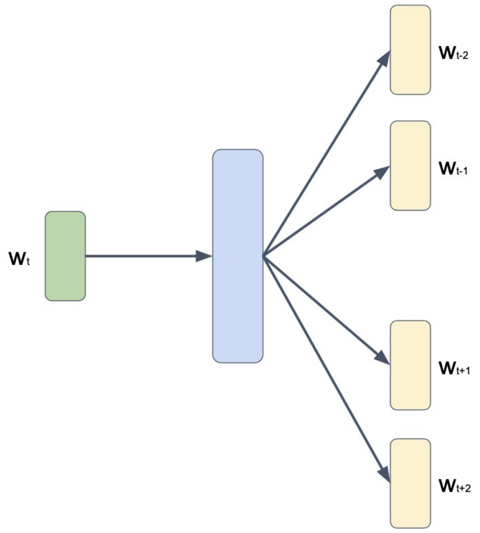


Figure 4.1: The skip-gram model architecture which predicts surrounding words given the center word. w_t is the center word, input it to the model, after passing through the projection layer, the model is learned to predicts its surrounding words w_{t-2}, w_{t-1}, w_{t+1} and w_{t+2}.

$$P(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}|w_t). \tag{4.1}$$

Assume that the appearances of surrounding words are independent of each other, equation 4.1 can be written as

$$P(w_{t-2}|w_t)P(w_{t-1}|w_t)P(w_{t+1}|w_t)P(w_{t+2}|w_t). \tag{4.2}$$

Assume the length of a given sequence of words is T, such as $w_1, w_2, ..., w_T$, the objective of the skip-gram model can be formally defined to maximize the log probability of the surrounding words given by

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq i\leq c, i\neq 0}\log p\left(w_{t+i} \mid w_t\right). \tag{4.3}$$

In the skip-gram model, each word has two vector representations, which are the learnable parameters. For the word w_t, when it is the central word, its vector representation is $\boldsymbol{v}_{w_t} \in \mathbb{R}^d$, and when it is the surrounding word, the vector is expressed as $\boldsymbol{u}_{w_t} \in \mathbb{R}^d$. The conditional probability of w_{t+1} given w_t can be obtained by performing a softmax operation on the inner product of the vectors:

$$P\left(w_{t+1} \mid w_t\right)=\frac{\exp \left(\boldsymbol{u}_{w_{t+1}}^{\top} \boldsymbol{v}_{w_t}\right)}{\sum_{w_i \in \mathcal{V}} \exp \left(\boldsymbol{u}_{w_i}^{\top} \boldsymbol{v}_{w_t}\right)}, \tag{4.4}$$

where $\mathcal{V}$ is the vocabulary of the dataset. From equation 4.4 we can see that if the vocabulary is large, the computational complexity will be pretty high. Hence, negative sampling is defined by

$$\begin{aligned} \log P\left(w_{t+1} \mid w_t\right)= & \log \sigma\left(\boldsymbol{u}_{w_{t+1}}^T \boldsymbol{v}_{w_t}\right) \\ & +\sum_{j=1}^{j=k} \mathbb{E}_{w_i \sim p(w)}\left[\log \sigma\left(\boldsymbol{u}_{w_i}^T \boldsymbol{v}_{w_t}\right)\right], \end{aligned} \tag{4.5}$$

where σ represents sigmoid function, then for calculating the conditional probability, only k negative words should be computed instead of the whole vocabulary and the k negative samples drawn from the smoothed noise distribution $p(w)$ [42] given by

$$p(w)=\frac{f(w)^\alpha}{\sum_{w_i \in \mathcal{V}} f\left(w_i\right)^\alpha}, \tag{4.6}$$

where $f(w)$ is frequency of word w, and α is distribution smoothing parameter.

When train the skip-gram models, I keep most of the hyper-parameters set in [18]: tokens context window $c=7$, negative sampling distribution parameter $\alpha=0.75$, the dimension of embedding vectors $v=10,50,100,128$

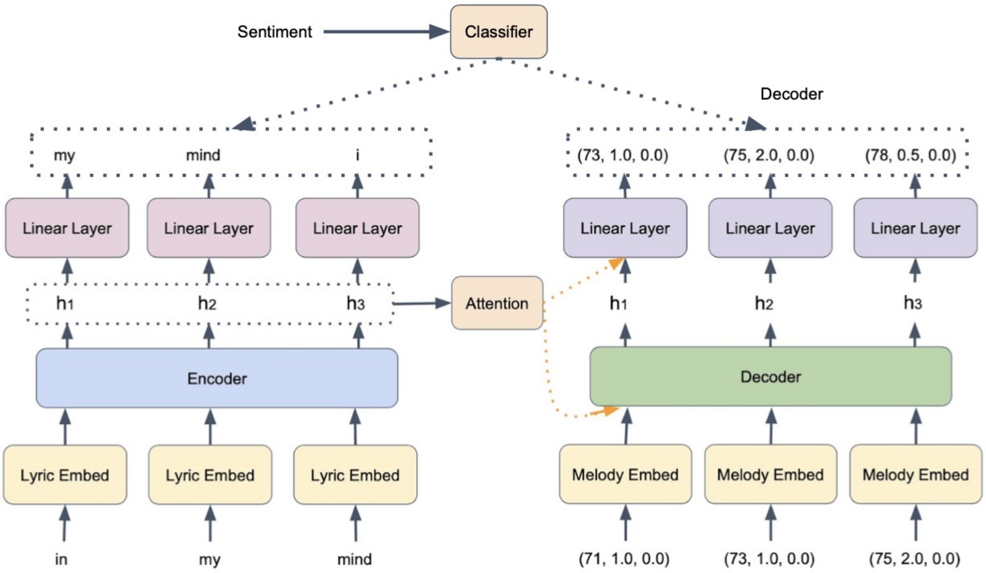


Figure 4.2: The architecture of the proposed SLMG system. The encoder is designed as lyric generator and lyric encoder, which takes a sequence of English syllables as input, uses the pre-trained skip-gram models as embedding layer, and output a context vector as well as predictions of next tokens of the input lyric. The function of the decoder is to generate melody as shown on the right part. It uses attention mechanism to look at the lyric during melody generation process. After training, the generation process is controlled by a classifier using the SBS algorithm.

and the learning rate is set to 0.03 with a gradually decay. After training two skip-gram models on my dataset, I obtain word-level and syllable-level embedding models, denoted as $E_w(\cdot)$ and $E_s(\cdot)$ respectively. For a syllable s comes from word w, it can be represented as the concatenation of $E_w(w)$ and $E_s(s)$, denoted as $E_w(w)||E_s(s)$.

4.2 Lyric-melody Generator

Although the generative models such as GAN and VAE can be modified to generate the data in specific categories. Conditional-GAN can generate certain types of things by inputting the condition vector to both the generator and discriminator. Conditional-VAE is an extension of VAE which can be used for conditional generation.

However, if I use Conditional-GAN or Conditional-VAE here as the generator. Every data needs a clear label. So I have to use "unlabelled" as a label input to the generator. But "unlabelled" is not a label, "unlabelled" here means that the sentiment expressed in these music segments is not obviously positive or negative as illustrated in Section 3.4, but does not mean that these data are similar.

In this work, due to the characteristics of the dataset (part of the data is unlabeled, both lyric and melody are given), I choose to use a seq2seq model for music generation as shown in Figure 4.2, so that the unlabelled data can be used in the training process. And then, the generation process can be controlled by the proposed SBS algorithm for sentiment-conditioned generation.

The architecture of the proposed lyric-melody generator is shown in Figure 4.2, which is a sequential encoder-decoder model trained end-to-end to compose lyrics and melodies.

The encoder is designed as lyric generator and lyric encoder. It takes a sequence of English syllables as input, denoted as $S = \{s_1, s_2, ..., s_T\}$. The lyric embedding layers are skip-gram models [40] trained on the whole lyrics dataset as illustrated in Section 4.1, I keep most of the hyper-parameter settings in [18] for training the skip-gram models. After training, I obtain word-level and syllable-level embedding models, denoted as $E_w(\cdot)$ and $E_s(\cdot)$ respectively. Then, the output vectors of lyric embedding layers are input into the encoder.

The encoder takes the whole syllable sequence S as its input and output a sequence of hidden states as the representation of the input lyric, $H = \{h_1, h_2, ..., h_T\}$. These output hidden states are used for lyric generation. Input H into a fully connected layer, for every unit, the lyric generator is modeled to predict the next syllable token conditioned on all the previous syllables in the input sequence. Thus, the goal of encoder is learning a probability distribution such as

$$p(S) = \prod_{t=1}^{T} p\left(s_t \mid s_1, s_2, \ldots, s_{t-1}\right). \tag{4.7}$$

Therefore, the loss function of encoder is defined as

$$L_{lyric} = -\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \log p_\theta\left(s_t \mid s_1, s_2, \ldots, s_{t-1}\right). \tag{4.8}$$

The decoder takes the corresponding melody sequence as input, $M = \{m_1, m_2, ..., m_T\}$, where each m_i is a (pitch, duration, rest) three-dimensional tuple. Firstly, each m_i is converted to a word form representation, for example, $m = (70, 1.0, 0.0)$ are denoted as 'p_70 d_1.0 r_0.0', then the melody notes can be input into embedding layer as normal words. The output of decoder is also a sequence of hidden states, $\tilde{H} = \{\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_T\}$, which is the representation of the input melody. In addition, attention mechanism [37; 43] is used to insure that lyric is taken into consideration during the melody generation process. The same as encoder, these output hidden states are input to a fully connected layer, for every unit, the melody generator is learned to predict the next melody note conditioned on previous melody notes and the corresponding lyric, which means that the melody generator is modeled to learn the following probability distribution

$$p(M|S) = \prod_{t=1}^{T} p\left(m_t \mid m_1, m_2, \ldots, m_{t-1}, S\right). \tag{4.9}$$

Therefore, the melody generator is trained to minimize the negative log conditional probability

$$L_{melody} = -\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \log p_{\theta}\left(m_t \mid m_{i<t}, S\right). \tag{4.10}$$

Combine the loss function of the encoder and decoder, the lyric-melody generator is trained to minimize the total loss defined as

$$L = L_{lyric} + \lambda L_{melody}, \tag{4.11}$$

where λ is a real value hyper-parameter.

4.3 Music Sentiment Classifier

In order to control the music generation process, I train a music sentiment classifier by using the labelled data. It takes a sequence of music, $C = \{c_1, c_2, ..., c_T\}$, as input, each c_i is a (syllable, pitch, duration, rest) four-dimensional tuple. The syllable is converted to a vector and then the three-dimensional music note (pitch, duration, rest) is also embedded as a vector. These two vectors are concatenated to represent a music note c_i. Next step, bidirectional long short-term memory (LSTM) network and multi-head self-attention Transformer [37] encoder are trained to predict the label of the input music sequence C.

4.4 Music Generation with Sentiments

In this section, I describe how to use the music sentiment classifier to control the process of music generation to match a particular sentiment. Beam search is a commonly used algorithm for text generation and neural machine translation [44], which selects the best and most likely words for the target sequence. In this work, the music generator is required to generate music not only harmonious but also perceived to have a specific sentiment. For that I propose sentimental beam search (SBS), a modified beam search algorithm guided by the music sentiment classifier as illustrated in Section 4.3.

The SBS algorithm takes an initial seed lyric with length n, lyric and melody generator G, music sentiment classifier C, beam size $b_1\&b_2\&b_3$ as input, output a piece of music with required sentiment e of length N, where $n < N$.

As shown in Figure 4.3, assuming that a piece of music with length t has been generated, which consists of a piece of lyric $S = \{s_1, s_2, ..., s_t\}$ and a piece of melody $M = \{m_1, m_2, ..., m_t\}$. The probability of x_i being the next lyric token can be calculated by using softmax function to the output of encoder at position t

$$p(s_{t+1} = x_i | S) = \frac{\exp{(e_{ti})}}{\sum_{k=0}^{|V_s|} \exp{(e_{tk})}}, \tag{4.12}$$

where e_{ti} represents the i-th element of the output of encoder at position t, $|V_s|$ is the number of syllables in the vocabulary. The higher the probability, the more fluent lyrics are generated.

Similarly, the probability of y_i being the next melody note can be calculated by

$$p(m_{t+1} = y_i|S, M) = \frac{\exp{(d_{ti})}}{\sum_{k=0}^{|V_m|} \exp{(d_{tk})}}, \tag{4.13}$$

where d_{ti} represents the i-th element of the output of decoder at position t, $|V_m|$ is the length of melody vocabulary. Music note with high probability means the generated melody sound harmonious.

After calculating the probabilities of all tokens by using equation 4.12 and equation 4.13, b_1 lyric tokens and b_2 melody tokens with highest probabilities are selected, therefore, $b_1 * b_2$ candidate lyric-melody pairs are chosen in total, $\{(x_i, y_j)|i = 1, ..., b_1; j = 1, ..., b_2\}$.

Adding every candidate lyric-melody pair (x_i, y_j) to the original music piece (S, M), the probability that the new music piece is perceived to have a specific sentiment e can be computed by the music sentiment classifier

$$p(e|(S, M)||(x_i, y_j)) = \frac{\exp{(e)}}{\sum_{j=1}^{E} \exp{(e_j)}}, \tag{4.14}$$

where E is the number of sentiments in the dataset and $||$ represents the concatenation operation. After calculating the probabilities of all candidate lyric-melody pairs, b_3 music segments with length $t + 1$ that have highest probabilities to represent the required sentiment e are generated.

Therefore, there's b_3 segments of each length, and for every segment, $b_1 * b_2$ candidate lyric-melody pairs should be evaluated. So the computational complexity of SBS is $O(N * b_1 * b_2 * b_3)$ where N is the required length.

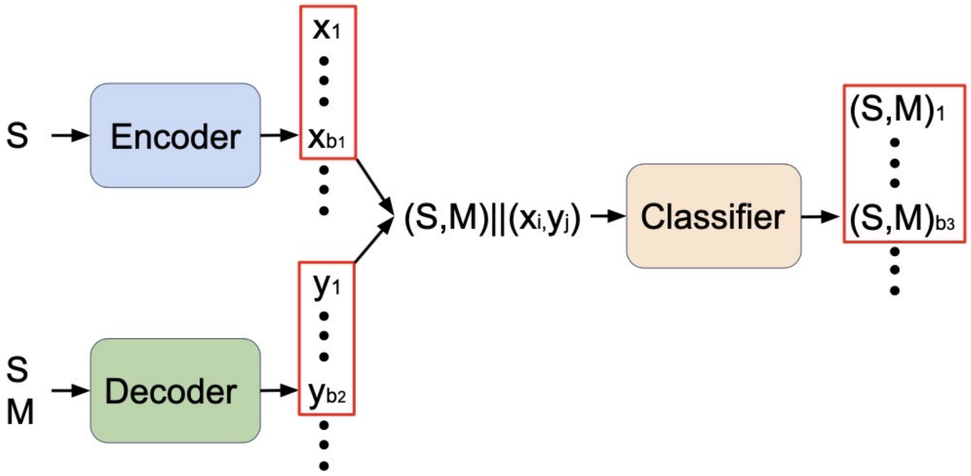


Figure 4.3: The schematic diagram of SBS algorithm. Assuming that a piece of music with length t has been generated, which consists of a piece of lyric $S = \{s_1, s_2, ..., s_t\}$ and a piece of melody $M = \{m_1, m_2, ..., m_t\}$. Input S to encoder, b_1 syllables are selected; input S and M to decoder, b_2 melody notes are selected. Then concatenate each candidate lyric-melody pair (x_i, y_j) to the original music piece (S, M) and input to the classifier, b_3 music segments with length $t + 1$ that have highest probabilities to represent the required sentiment are generated.

Chapter 5

Experiments and Evaluation

Summary

Experimental setup, evaluation methods and experimental results are introduced in this section. The empirical evaluation of the proposed SLMG system is divided into three parts. First, I evaluate the accuracy of the music sentiment classifier in Section 5.1. Then, the experimental setup and objective evaluation of the lyric-melody generator are demonstrated in Section 5.2. Finally, the subjective evaluation of the generated music is shown in Section 5.3. The code of this work can be downloaded at https://github.com/BaoChunhui/Generate-Emotional-Music.

5.1 Sentiment Classifier

As demonstrated in Section 4.3, both bidirectional LSTM and Transformer are trained on labelled data to classify the music sentiment. As shown in Table 3.5, the number of positive samples is larger than the number of

datasets	Length		
	20	50	100
Bidirectional LSTM	99.8	99.9	99.9
Self-attention Transformer	100.0	99.9	99.9

Table 5.1: sentiment classification accuracy (%) of LSTM and Transformer on different datasets with different length.

negative samples, so over-sampling method is used for negative samples.

For bidirectional LSTM, The number of layers is set to 6 and the dimension of hidden state is set to 256. The learning rate is set to 0.0001 with gradually decay. The number of epochs is 30 and the dimension of embedding vector is set to 256. For Transformer, The number of Transformer blocks is set to 6, each Transformer block consists of an 8-head self-attention Transformer encoder layer connected with a LayerNorm [45]. The dimension of input is set to 128. The learning rate and number of epochs are the same with bidirectional LSTM.

The classifiers are evaluated by using a 8-fold cross validation approach, in which the testing fold and the training folds have no overlapping data. Table 5.1 shows the sentiment classification accuracy of all datasets created in Section 3.3, from it we can see that both the LSTM and Transformer based models can successfully classify the datasets. Therefore the classifier trained on labelled data of the datasets can be used in SBS algorithm.

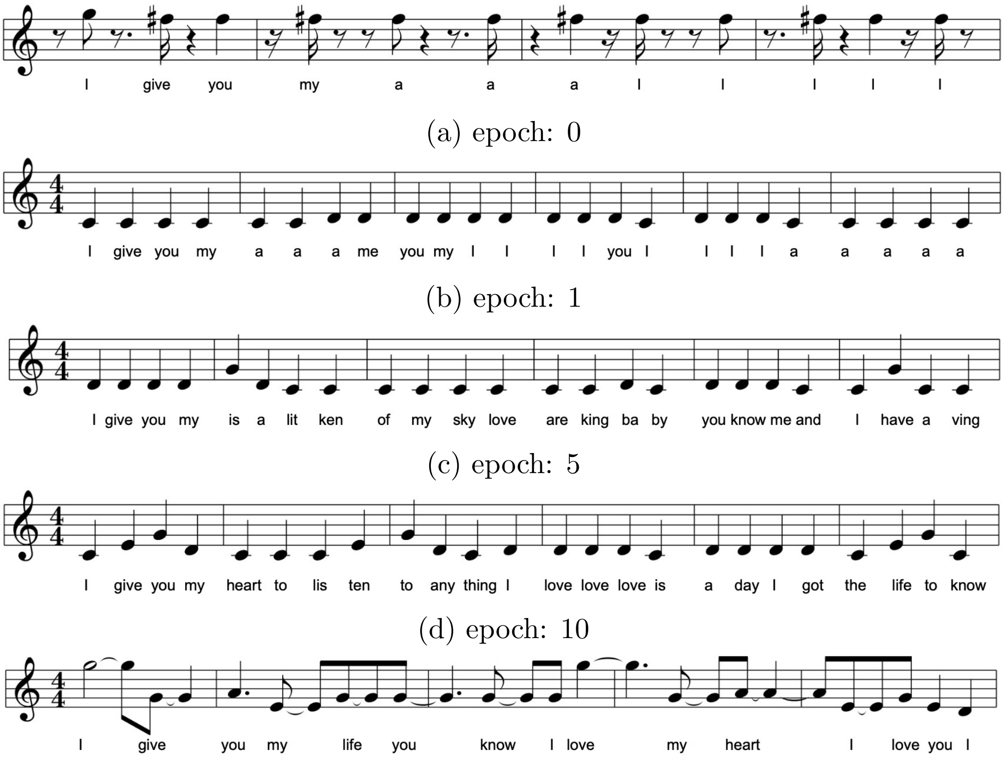
5.2 Music Generation

The lyric-melody generator is an encoder-decoder model trained end-to-end on the unlabelled datasets. 9/10 of them are used in the training process and 1/10 are used to evaluate the trained sequence to sequence model. Both GRU and Transformer based neural networks are trained for lyric-melody generation.

For GRU, the encoder and decoder have the same neural structure. The number of layers is set to 4 and the dimension of hidden state is set to 256. The initial hidden state of encoder is initialized with zero vector, and the initial hidden state of decoder is initialized with the last hidden state of encoder. All parameters are initialized from zero mean, 0.08 variance Gaussian distribution. For Transformer, both the encoder and decoder have 12 Transformer blocks, the number of head is set to 16 and the input dimension is set to 256. The loss function is optimized by Adam optimizer with initial learning rate of 0.0001 and decayed after every epoch. the λ in equation 4.11 is set to 1. The batch size is set to 64, 32, 16 for datasets with length 20, 50, 100 respectively.

Figure 5.1 shows the training process of the GRU based model. When model trained for 0, 1, 5, 10 and 30 epochs, one music segment is generated by using beam search algorithm with beam size 3. We can see that the generated music notes become more and more varied, and the generated lyrics become more and more fluent. In particular, there are dull and repetitive outputs, which is a common phenomenon in text generation tasks [46].

After training, the GRU and Transformer based networks are evaluated by using the test data. Input test data into the sequence to sequence model, melodies can be generated. In order to eliminate the duplication and increase diversity, the predicted probabilities of tokens occurred in the melody are divided by 2 and then the next token are randomly selected from the top 3 tokens. I also implement AutoNLMC on my dataset for com-


I give you my a a a I I I I I
I give you my a a a me you my I I I I you I I I I a a a a a
I give you my is a lit ken of my sky love are king ba by you know me and I have a ving
I give you my heart to lis ten to any thing I love love love is a day I got the life to know
I give you my life you know I love my heart I love you I

(a) epoch: 0

(b) epoch: 1

(c) epoch: 5

(d) epoch: 10

(e) epoch: 30

Figure 5.1: Generated music segments when model trained for 0, 1, 5, 10 and 30 epochs respectively. The generated music notes become more and more varied, and the generated lyrics become more and more fluent.

parison, which is a sequence to sequence model consists of one encoder and multiple decoders proposed in [31]. Different from my method, AutoNLMC regards each attribute of the melody as independent and train decoders separately for each attribute. Then, I compare the melody distributions of ground-truth melodies and melodies generated by AutoNLMC, GRU based model as well as Transformer based model. Detailed quantitative comparison of melody distributions are shown in Table 5.2. The frequency distribution histograms of ground-truth melody are shown in Figure 5.2, the melody distributions generated by AutoNLMC, GRU based generator and Transformer based generator are show in Figure 5.3, Figure 5.4 and

Items	Ground-Truth	AutoNLMC	GRU	Transformer
Mean value of pitch	66.55	66.71	66.62	66.51
Standard deviation of pitch	10.05	9.31	9.40	9.64
Number of unique pitch value	82	81	81	76
Maximum pitch value	111	111	111	101
Minimum pitch value	3	6	6	12
Mode of duration	1.0	1.0	1.0	1.0
Number of unique duration value	19	18	18	18
Maximum duration value	32.5	32.0	32.0	32.0
Minimum duration value	0.25	0.25	0.25	0.25
Percentage of 1.0 (%)	43.63	52.58	74.22	76.83
Mode of rest	0.0	0.0	0.0	0.0
Number of unique rest value	8	8	8	8
Maximum rest value	32.0	32.0	32.0	32.0
Percentage of 0.0 (%)	80.65	81.96	96.67	96.86

Table 5.2: Detailed comparison of ground-truth melody distribution and generated melody distributions.

Figure 5.5 respectively. In addition, in order to further compare these three generators, the training and testing loss, training and testing perplexity, as well as Jensen-Shannon divergence between the ground-truth distribution and generated distributions are given in Table 5.3. Compared with AutoNLMC, the quality of pitches generated by my models is better, since the pitch distributions generated by my models have higher standard deviation and lower Jensen-Shannon divergence, which means that the pitches generated by my models are more diverse and closer to the ground-truth data. But the disadvantage of my models is that the generated duration and rest have lower diversity than ground-truth data and AutoNLMC-generated data. Moreover, the training and testing loss, training and testing perplexity of Transformer based generator are much lower than AutoNLMC and GRU based generator. It demonstrates the Transformer has stronger learning ability and can better fit the dataset.

Then, I generate lyrics and melodies by using the SBS algorithm introduced in Section 4.4. I use 5 different seed lyrics: “I give you my”, “but

Items	AutoNLMC	GRU	Trans.
Training loss	7.60	7.75	4.79
Testing loss	8.69	8.59	5.21
Training perplexity	2004.85	2316.11	120.77
Testing perplexity	5967.03	5369.26	183.74
Pitch JSD vs GD	.0186	.0140	.0111
Duration JSD vs GD	.0069	.1174	.1499
Rest JSD vs GD	.0025	.0694	.0720

Table 5.3: Detailed comparison of AutoNLMC, GRU based generator and transformer based generator. Here “trans.” stands for transformer based generator, “JSD” and “GD” are the abbreviation of Jensen-Shannon divergence and ground-truth respectively.

when you told me”, “if I was your man”, “I have a dream”, “when I got the” and different generators trained on various datasets (length = 20, 50, 100) with various skip-gram models (dimension = 10, 50, 100, 128). For the SBS algorithm, the beam size is set to $(b_1 = 3, b_2 = 3, b_3 = 5)$ and the maximum length is set to 25. I generate 180 segments by using the GRU based generator and LSTM based classifier, in which 60 are positive, 60 are negative and 60 are uncontrolled. Similarly, 180 segments are generated by using the Transformer based generator and Transformer based classifier. Generated samples with required sentiment are shown in Figure 5.6 and Figure 5.7. Then I use the fine-tuned Bert model introduced in Section 3.3 and the classifier used in SBS to objectively evaluate them. The evaluation results are shown in Table 5.4, which shows that the SBS algorithm successfully controlled the generation process.

Without control, the generator tends to generate more positive samples since the unbalanced distribution in training dataset. By using SBS algorithm, the generation process is controlled by the music sentiment classifier. From the perspective of the classifier used in SBS, all music segments are

	Annotator			Classifier		Total
	P	N	U	P	N	
GRU Positive	45	0	15	60	0	60
GRU Negative	2	33	25	0	60	60
GRU Uncontrolled	23	17	20	33	27	60
Transformer Positive	47	0	13	60	0	60
Transformer Negative	1	37	22	0	60	60
Transformer Uncontrolled	25	14	21	34	26	60

Table 5.4: Objective evaluation of the generated music pieces. The classifier used in SBS and fine-tuned Bert annotator are utilized to evaluate the generated lyrics. “P”, “N” and “U” represent positive, negative and unlabelled respectively.

correctly generated to convey the required sentiment. Even use the Bert annotator to measure the generated segments, we can see that SBS algorithm obviously bias the generation process towards the given sentiment, most of the generated segments convey the required sentiment from the perspective of the generator. We can also observe that SBS algorithm can applied to both traditional GRU or LSTM based model and Transformer based model.

5.3 Subjective Evaluation

Although statistical and objective evaluation indicate that the model is able to generate harmonious lyric and melody to capture the required sentiment, it is still difficult to conclude that the generated music pieces please human ears and evoke sentiments in listeners’ hearts. Music composition is a human creative process, so I adapt the subjective evaluation method to evaluate generated lyrics and melodies. I invited volunteers to evaluate the music data selected from the ground-truth dataset, music segments generated by GRU based model and Transformer based model.

Firstly, the participants should offer their basic information, include their name, age, gender and musicianship experience. Musicianship experience was assessed using a 5-point scale where 1 to 5 means “I’ve never studied music theory or practice”, “I’ve studied music theory or practice within two years”, “I’ve studied music theory or practice for two to five years”, “I’ve studied music theory or practice for more than five years” and “I have an academic degree in music” respectively. Then, each participant needs to evaluate 18 music pieces. For each piece of music, first play the melody to the participants and ask the participants to classify the sentiment conveyed by the melody (positive or negative). Next, the lyric of the this music piece is given to participants. Participants should to classify the sentiment conveyed by this music segment again according to the lyric, in this step, participants are not allowed to change the classification answer of previous question but can make a different decision about the sentiment conveyed by this music segment. Finally, we ask the following questions to participants

- Is this melody agreeable to the ears?
- Is this lyric meaningful?
- How well does the melody fit the lyric of this music segment?

Participants answer the above questions on a five point discrete scale where 1 to 5 corresponds to “Very bad”, “Bad”, “Ok”, “Good” and “Very good” respectively.

I invited 20 participants for the subjective evaluation, where 10 are male and 10 are female. They have an average age of approximately 24.5

	GT	GRU	Transformer
Positive lyrics	100.0	95.0	96.7
Negative lyrics	96.7	91.7	93.3
Positive melodies	48.3	51.7	50.0
Negative melodies	53.3	46.7	55.0

Table 5.5: Classification accuracy for ground-truth (GT) and generated music segments (%).

Questions	**Ground-truth**					
	1	**2**	**3**	**4**	**5**	**Average**
How meaningful are the lyrics	0	7	17	31	5	3.6
How sounds good are the melodies	1	8	26	22	3	3.3
How well does the melodies fit the lyrics	0	7	23	25	5	3.5

Table 5.6: Answers of questions given by 20 Participants of ground-truth music segments.

Questions	**GRU**					
	1	**2**	**3**	**4**	**5**	**Average**
How meaningful are the lyrics	2	4	25	29	0	3.4
How sounds good are the melodies	4	6	16	29	5	3.4
How well does the melodies fit the lyrics	0	4	20	32	4	3.6

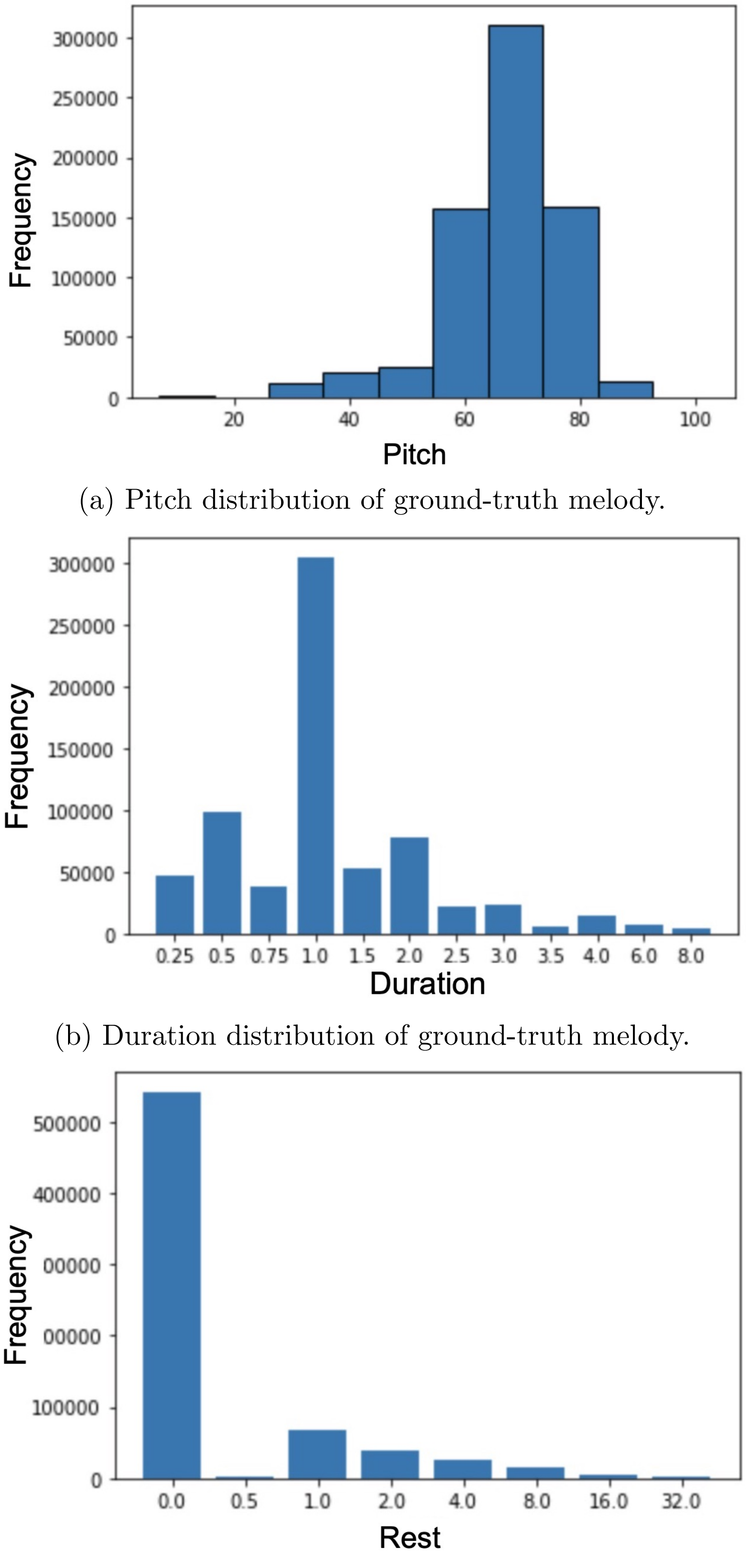
Table 5.7: Answers of questions given by 20 Participants of GRU generated music segments.

Questions	**Transformer**					
	1	**2**	**3**	**4**	**5**	**Average**
How meaningful are the lyrics	0	5	29	26	0	3.4
How sounds good are the melodies	2	6	27	25	0	3.3
How well does the melodies fit the lyrics	1	4	17	30	8	3.7

Table 5.8: Answers of questions given by 20 Participants of Transformer generated music segments.

years and the average musicianship experience is 2.45. Detailed subjective classification results are shown in Table 5.5. We can see that only by listening to the melodies, participants cannot distinguish the sentiment of the music segments. Even on the ground-truth data, the classification accuracy is about 50%. This shows that sentiments in the new dataset are mainly conveyed by lyrics. After reading the lyrics, the classification accuracy has increased to more than 90%, which demonstrates that the SLMG system proposed in this work successfully learned to generate lyric and melody with a required sentiment. In addition, we can also find that the sentiments conveyed by generated music segments are more ambiguous than ground-truth data. I investigate the quality of music segments by asking questions, such as “Is this melody agreeable to the ears?”, “Is this lyric meaningful?” and “How well does the melody fit the lyric of this music segment?”. The results are shown in Figure 5.6, Figure 5.7 and Figure 5.8. We can see that both GRU based generator and Transformer based generator can successfully generate music segments of almost the same high quality as the training dataset. Even though Transformer has stronger learning ability and can better fit the training data, the quality of music segments generated by Transformer dose not obviously beyond GRU.

I interviewed some participants to ask them how they classify the melodies and lyrics, how they feel about the quality of the melodies and lyrics. The participants told me that there’s no obvious difference between these melodies, they classified the sentiments of the melodies by their own feelings. The sentiments of the lyrics can be classified by short sentences such as “I love you”, “I don’t love you”, and keywords such as ”good”, ”bad”. One participant told me that some lyrics have obvious grammatical errors and typos, he gave low scores to these lyrics. I think this is the reason why the quality of generated lyrics is slightly lower than ground-truth lyrics (3.4 vs 3.6). Another participant told me that he thought all the lyrics (including ground-truth data) are very low quality. There are no punctuations, and many short sentences have no sentiment. I think that the quality of the dataset is the bottleneck of the SLMG system. The SLMG system has the potential to generate music with higher quality if a better dataset is given.

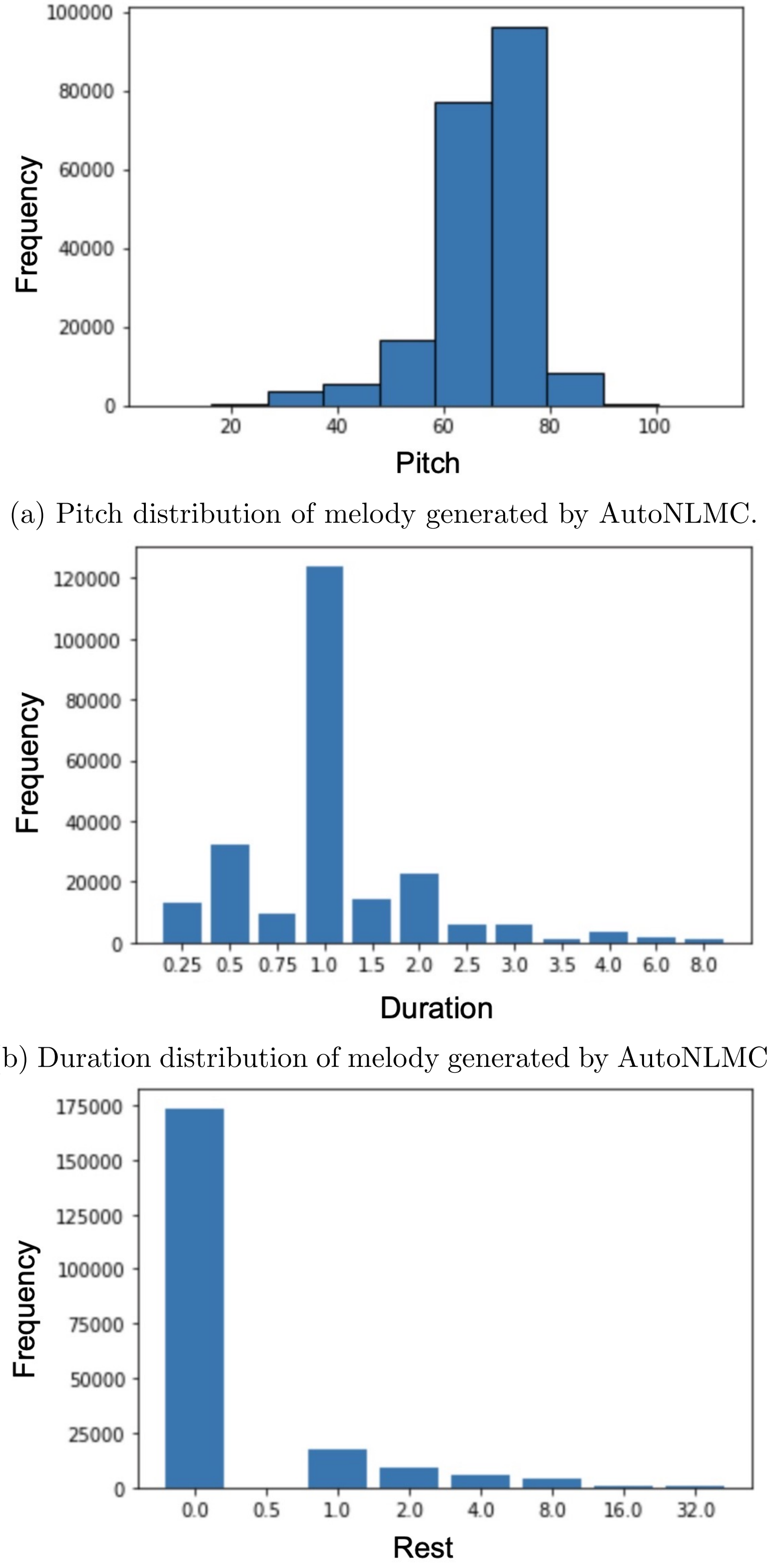


(a) Pitch distribution of ground-truth melody.

(b) Duration distribution of ground-truth melody.

(c) Rest distribution of gd-truth melody.

Figure 5.2: Distributions of ground-truth melody. (a), (b), (c) show the distribution of pitch, duration and rest respectively.

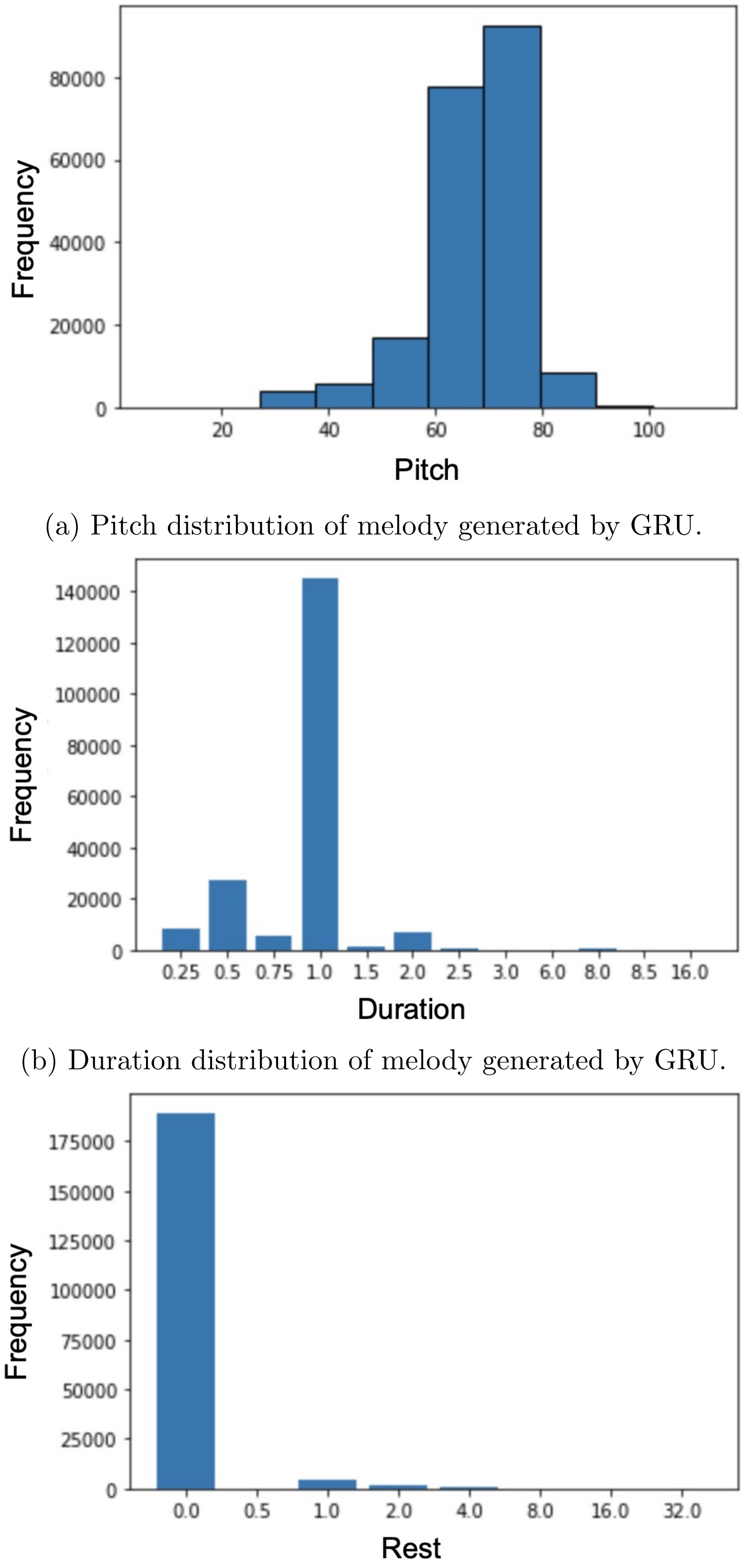


(a) Pitch distribution of melody generated by AutoNLMC.

(b) Duration distribution of melody generated by AutoNLMC.

(c) Rest distribution of melody generated by AutoNLMC.

Figure 5.3: Distributions of melody generated by AutoNLMC. (a), (b), (c) show the distribution of pitch, duration and rest respectively.

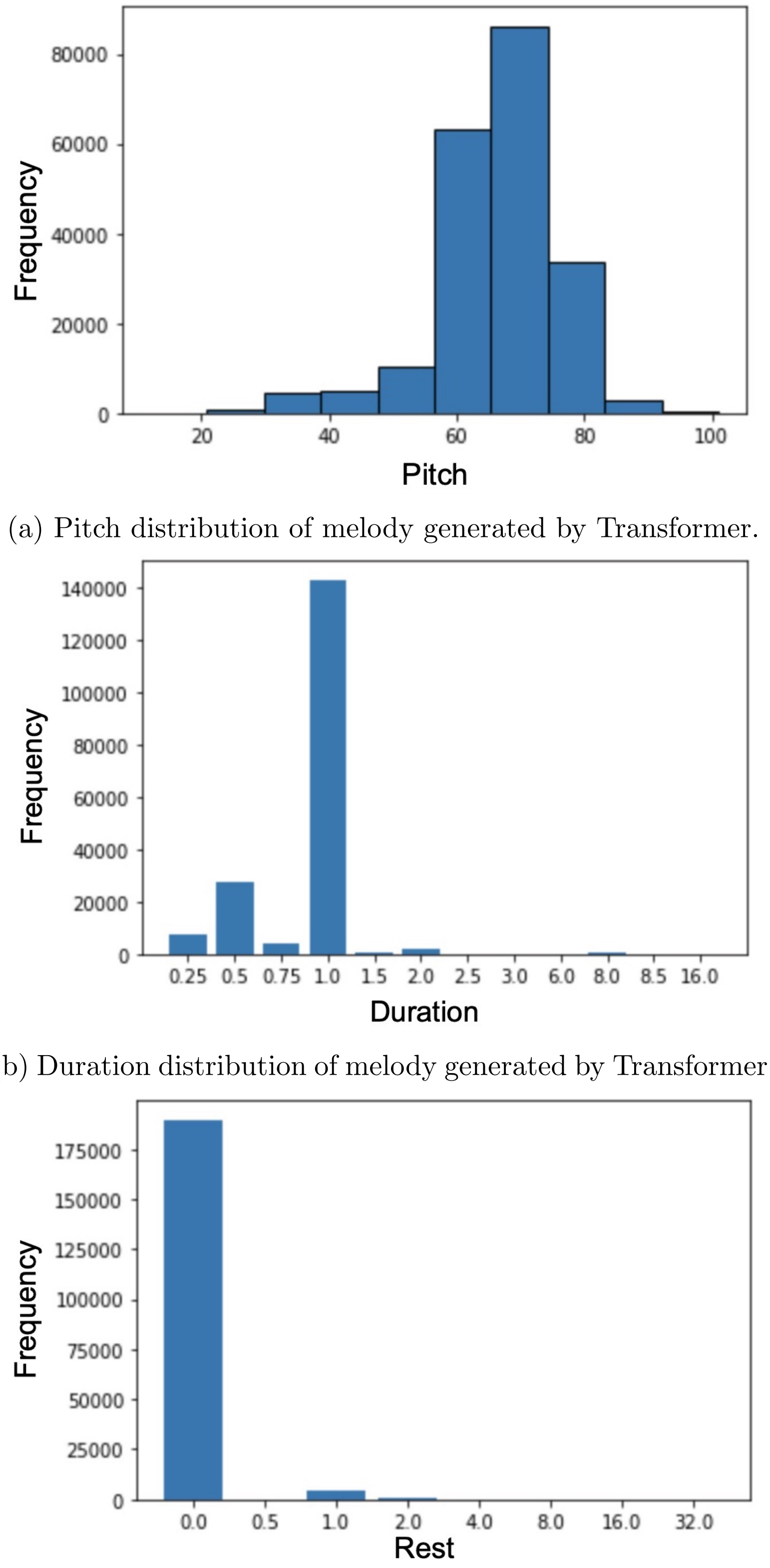


(a) Pitch distribution of melody generated by GRU.

(b) Duration distribution of melody generated by GRU.

(c) Rest distribution of melody generated by GRU.

Figure 5.4: Distributions of melody generated by GRU. (a), (b), (c) show the distribution of pitch, duration and rest respectively.

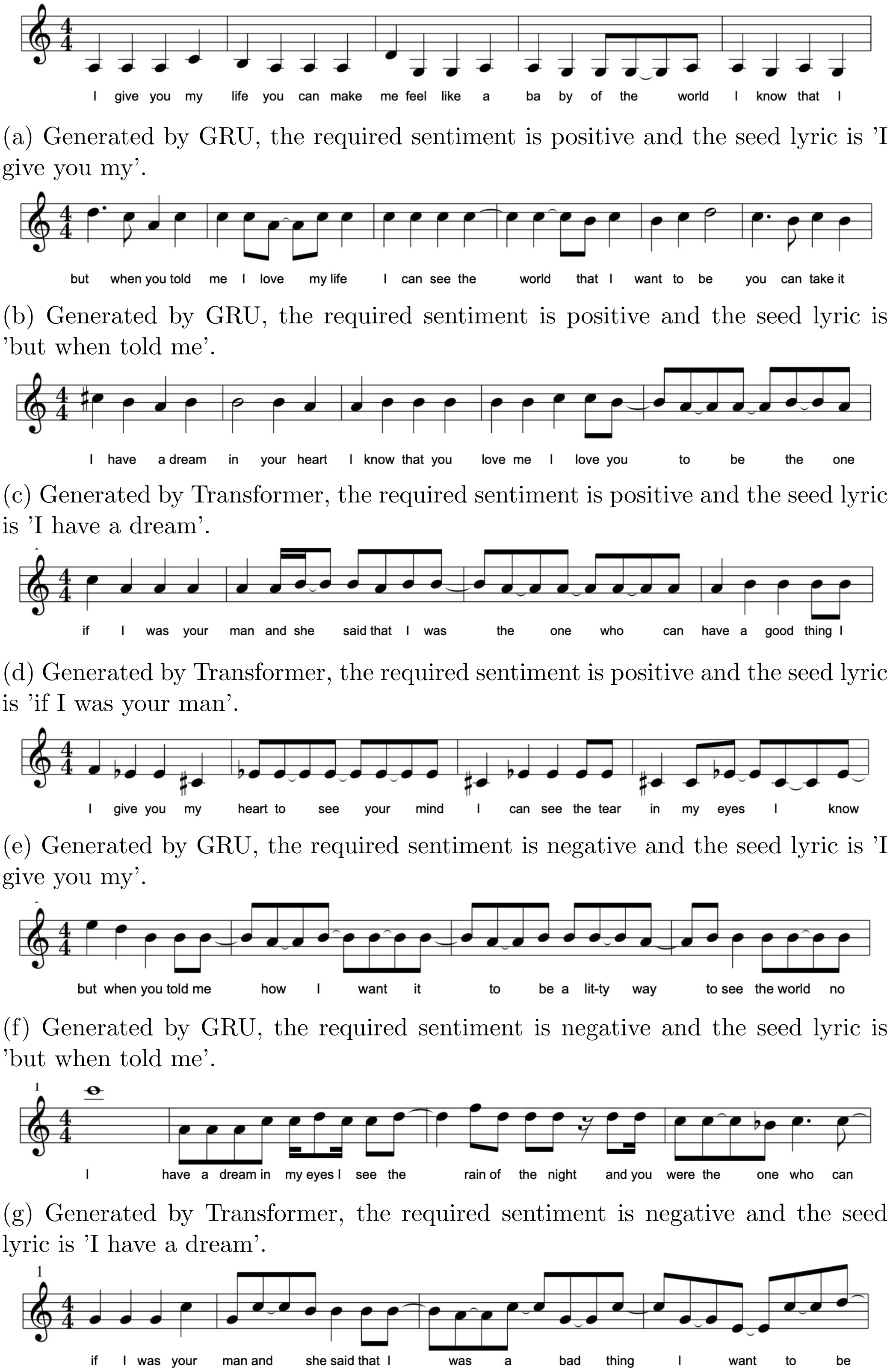


(a) Pitch distribution of melody generated by Transformer.

(b) Duration distribution of melody generated by Transformer.

(c) Rest distribution of melody generated by Transformer.

Figure 5.5: Distributions of melody generated by Transformer. (a), (b), (c) show the distribution of pitch, duration and rest respectively.


I give you my life you can make me feel like a ba by of the world I know that I
but when you told me I love my life I can see the world that I want to be you can take it
I have a dream in your heart I know that you love me I love you to be the one
if I was your man and she said that I was the one who can have a good thing I
I give you my heart to see your mind I can see the tear in my eyes I know
but when you told me how I want it to be a lit-ty way to see the world no
I have a dream in my eyes I see the rain of the night and you were the one who can
if I was your man and she said that I was a bad thing I want to be

(a) Generated by GRU, the required sentiment is positive and the seed lyric is 'I give you my'.

(b) Generated by GRU, the required sentiment is positive and the seed lyric is 'but when told me'.

(c) Generated by Transformer, the required sentiment is positive and the seed lyric is 'I have a dream'.

(d) Generated by Transformer, the required sentiment is positive and the seed lyric is 'if I was your man'.

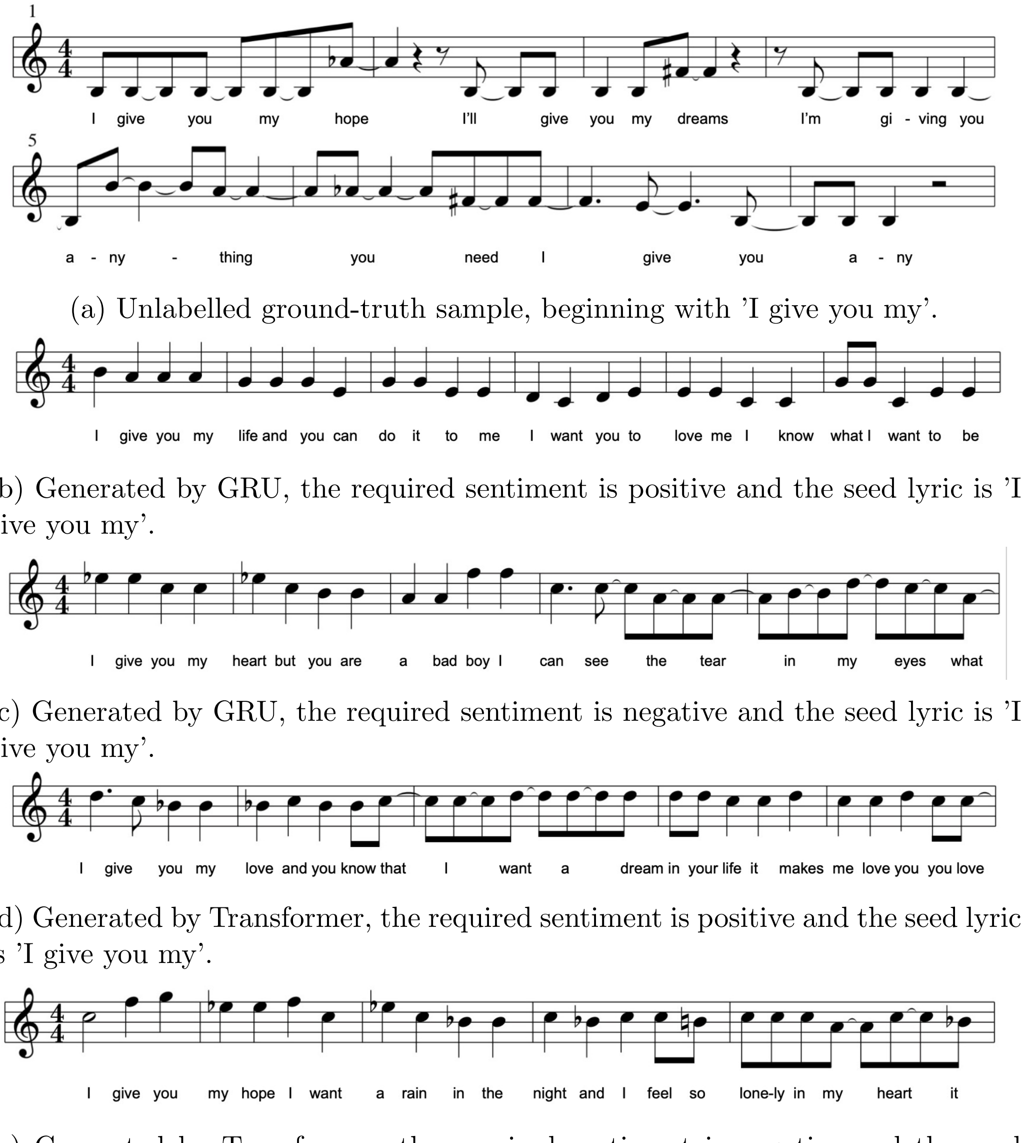
(e) Generated by GRU, the required sentiment is negative and the seed lyric is 'I give you my'.

(f) Generated by GRU, the required sentiment is negative and the seed lyric is 'but when told me'.

(g) Generated by Transformer, the required sentiment is negative and the seed lyric is 'I have a dream'.

(h) Generated by Transformer, the required sentiment is negative and the seed lyric is 'if I was your man'.

Figure 5.6: Generated samples of the SLMG system.


1
I give you my hope I'll give you my dreams I'm gi - ving you
5
a - ny - thing you need I give you a - ny
I give you my life and you can do it to me I want you to love me I know what I want to be
I give you my heart but you are a bad boy I can see the tear in my eyes what
I give you my love and you know that I want a dream in your life it makes me love you you love
I give you my hope I want a rain in the night and I feel so lone-ly in my heart it

(a) Unlabelled ground-truth sample, beginning with 'I give you my'.

(b) Generated by GRU, the required sentiment is positive and the seed lyric is 'I give you my'.

(c) Generated by GRU, the required sentiment is negative and the seed lyric is 'I give you my'.

(d) Generated by Transformer, the required sentiment is positive and the seed lyric is 'I give you my'.

(e) Generated by Transformer, the required sentiment is negative and the seed lyric is 'I give you my'.

Figure 5.7: Ground-truth and generated samples with the same beginning 'I give you my'.

Chapter 6

Conclusion

In this thesis, I construct a large-scale paired lyric-melody dataset with sentiment labels and propose Sentimental Lyric and Melody Generator (SLMG) system for sentiment-conditioned music generation. Firstly, I find that dataset annotators trained on in-domain data are more reliable than models trained on out-of-domain data. Then, both GRU and Transformer based encoder-decoder network trained on the new dataset successfully learned to compose lyric and melody. Next, sentimental beam search (SBS) algorithm is designed to control the generation process by using a music sentiment classifier, which let the generated music segments represent the specific given sentiment. Finally, subjective and objective evaluations demonstrate that the SBS algorithm can bias the generation process to required sentiments.

In addition, music generation with sentiments is still unexplored well and a challenging problem in deep learning area. The new dataset created in this work only has single track in the melody and the sentiment annotator only focus on the lyric. The quality of the dataset limits the effectiveness of SLMG. Building large-scale polyphonic music dataset with sentiment labels is a valuable further work for me.

Appendix A

Questionnaire

Hello, this is a questionnaire about my music generation research. Please answer the following questions, they will take you about 20 minutes.

What's your name?

How old are you?

What is your gender?

(1) Male (2) Female

Can you tell me your musicianship experience?

(1) I've never studied music theory or practice

(2) I've studied music theory or practice within two years

(3) I've studied music theory or practice for two to five years

(4) I've studied music theory or practice for more than five years

(5) I have an academic degree in music

Listen the melody of music segment 1, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 1, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 1.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 1?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 2, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 2, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 2.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 2?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 3, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 3, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 3.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 3?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 4, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 4, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 4.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 4?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 5, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 5, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 5.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 5?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 6, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 6, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 6.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 6?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 7, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 7, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 7.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 7?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 8, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 8, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 8.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 8?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 9, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 9, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 9.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 9?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 10, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 10, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 10.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 10?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 11, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 11, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 11.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 11?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 12, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 12, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 12.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 12?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 13, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 13, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 13.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 13?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 14, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 14, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 14.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 14?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 15, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 15, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 15.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 15?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 16, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 16, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 16.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 16?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 17, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 17, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 17.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 17?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Listen the melody of music segment 18, what kind of sentiment is expressed?

(1) Positive (2) Negative

Read the lyric of music segment 18, what kind of sentiment is expressed? Please do not change the answer of previous question and you can make a new decision of the sentiment expressed in music segment 18.

(1) Positive (2) Negative

Is this melody agreeable to the ears?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Is this lyric meaningful?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

How well does the melody fit the lyric of music segment 18?

(1) Very bad (2) Bad (3) Ok (4) Good (5) Very good

Thanks for your participation.

References

[1] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University-Computer and Information Sciences*, 2020. 2

[2] Y. Zhao, X. Xia, and R. Togneri, "Applications of deep learning to audio generation," *IEEE Circuits and Systems Magazine*, vol. 19, no. 4, pp. 19–38, 2019. 2

[3] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017. 2

[4] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801*, 2020. 2

[5] `https://colinraffel.com/projects/lmd/`. 2, 12, 14

[6] `https://www.reddit.com/r/datasets/`. 2, 12, 14

[7] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018. 2, 7

[8] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *International Conference on Machine Learning*, pp. 4364–4373, PMLR, 2018. 2

[9] 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. 2, 11

[10] J. Ens and P. Pasquier, "Mmm: Exploring conditional multitrack music generation with the transformer," *arXiv preprint arXiv:2008.06048*, 2020. 2

[11] M. Czyz and M. Kedziora, "Automated music generation using recurrent neural networks," in *International Conference on Dependability and Complex Systems*, pp. 22–31, Springer, 2021. 2

[12] L. N. Ferreira and J. Whitehead, "Learning to generate music with sentiment," *arXiv preprint arXiv:2103.06125*, 2021. 3, 8, 15

[13] L. Ferreira, L. Lelis, and J. Whitehead, "Computer-generated music for tabletop role-playing games," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, pp. 59–65, 2020. 4, 8

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. 4

[15] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "Emopia: A multi-modal pop piano dataset for emotion

recognition and emotion-based music generation," *arXiv preprint arXiv:2108.01374*, 2021. 4, 9, 15, 18

[16] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics," *arXiv preprint arXiv:2101.02051*, 2021. 4

[17] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-relevant features for classification and regression of music lyrics," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240–254, 2016. 4, 15

[18] Y. Yu, A. Srivastava, and S. Canales, "Conditional lstm-gan for melody generation from lyrics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–20, 2021. 4, 10, 12, 13, 14, 27, 29

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 4, 17

[20] D. Edmonds and J. Sedoc, "Multi-emotion classification for song lyrics," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 221–235, 2021. 4, 15, 16

[21] A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018. 5

[22] W. Kool, H. Van Hoof, and M. Welling, "Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement," in *International Conference on Machine Learning*, pp. 3499–3508, PMLR, 2019. 5

[23] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach: a steerable model for bach chorales generation," in *International Conference on Machine Learning*, pp. 1362–1371, PMLR, 2017. 6

[24] A. Roberts, J. Engel, and D. Eck, "Hierarchical variational autoencoders for music," 2017. 7

[25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014. 7

[26] O. Mogren, "C-rnn-gan: Continuous recurrent neural networks with adversarial training," *arXiv preprint arXiv:1611.09904*, 2016. 7

[27] C.-F. Huang and C.-Y. Huang, "Emotion-based ai music generation system with cvae-gan," in *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pp. 220–222, IEEE, 2020. 9

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 9

[29] H. Bao, S. Huang, F. Wei, L. Cui, Y. Wu, C. Tan, S. Piao, and M. Zhou, "Neural melody composition from lyrics," in *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 499–511, Springer, 2019. 10

[30] Y. Yu, F. Harscoët, S. Canales, G. Reddy, S. Tang, and J. Jiang, "Lyrics-conditioned neural melody generation," in *International Conference on Multimedia Modeling*, pp. 709–714, Springer, 2020. 10, 12

[31] G. R. Madhumani, Y. Yu, F. Harscoët, S. Canales, and S. Tang, "Automatic neural lyrics and melody composition," *arXiv preprint arXiv:2011.06380*, 2020. 10, 12, 13, 38

[32] Y. Yu, A. Srivastava, and R. R. Shah, "Conditional hybrid gan for sequence generation," *arXiv preprint arXiv:2009.08616*, 2020. 13

[33] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020. 15, 16

[34] E. Çano and M. Morisio, "Moodylyrics: A sentiment annotated lyrics dataset," in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 118–124, 2017. 15

[35] P. Ekman and W. V. Friesen, "A new pan-cultural facial expression of emotion," *Motivation and emotion*, vol. 10, no. 2, pp. 159–168, 1986. 16

[36] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions.* Oxford University Press, 1994. 16

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. 17, 30, 31

[38] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980. 19

[39] C. L. Krumhansl and E. J. Kessler, "Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys.," *Psychological review*, vol. 89, no. 4, p. 334, 1982. 22

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. 24, 29

[41] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016. 25

[42] H. Caselles-Dupré, F. Lesaint, and J. Royo-Letelier, "Word2vec applied to recommendation: Hyperparameters matter," in *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 352–356, 2018. 27

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. 30

[44] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," *arXiv preprint arXiv:1702.01806*, 2017. 32

[45] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *arXiv preprint arXiv:1911.07013*, 2019. 36

[46] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," *arXiv preprint arXiv:1908.04319*, 2019. 37