TRAINING STRATEGIES FOR AUTOMATIC SONG WRITING: A UNIFIED FRAMEWORK PERSPECTIVE

Tao Qian[1], *Jiatong Shi*[2], *Shuai Guo*[1], *Peter Wu*[2], *Qin Jin*[1]*

[1]School of Information, Renmin University of China, P.R.China
[2]Language Technologies Institute, Carnegie Mellon University, U.S.A.
{qiantao, shuaiguo, qjin}@ruc.edu.cn, {jiatongs, peterw1}@cs.cmu.edu

ABSTRACT

Automatic song writing (ASW) typically involves four tasks: lyric-to-lyric generation, melody-to-melody generation, lyric-to-melody generation, and melody-to-lyric generation. Previous works have mainly focused on individual tasks without considering the correlation between them, and thus a unified framework to solve all four tasks has not yet been explored. In this paper, we propose a unified framework following the pre-training and fine-tuning paradigm to address all four ASW tasks with one model. To alleviate the data scarcity issue of paired lyric-melody data for lyric-to-melody and melody-to-lyric generation, we adopt two pre-training stages with unpaired data. In addition, we introduce a dual transformation loss to fully utilize paired data in the fine-tuning stage to enforce the weak correlation between melody and lyrics. We also design an objective music generation evaluation metric involving the chromatic rule and a more realistic setting, which removes some strict assumptions adopted in previous works. To the best of our knowledge, this work is the first to explore ASW for pop songs in Chinese. Extensive experiments demonstrate the effectiveness of the dual transformation loss and the unified model structure encompassing all four tasks. The experimental results also show that our proposed new evaluation metric aligns better with subjective opinion scores from human listeners.

Index Terms— Automatic song writing; pre-training; dual transformation loss; music objective evaluation

1. INTRODUCTION

Automatic song writing (ASW) aims to allow machines to generate associated melodies and lyrics automatically. In previous literature, four tasks have been investigated for ASW: lyrics-to-lyrics generation (L2L, generating lyrics from initial lyrics) [1], melody-to-melody generation (M2M, generating melody from initial melody) [2, 3], lyric-to-melody generation (L2M, generating melody from corresponding lyrics) [4–7] and melody-to-lyric generation (M2L, generating lyrics from corresponding melody) [5, 7–9]. We identify three main challenges in ASW: 1) **data scarcity**. Paired melody and lyrics data are required for model learning in M2L and L2M tasks, which cannot be easily obtained due to high annotation costs and copyright issues. Consequently, most previous works [4, 8] only use limited paired data. Some efforts have been made to take advantage of the easily obtained large amount of unpaired data recently [7], which is a promising direction to pursue. 2) **correlation between melody and lyrics**. The correlation between melody and lyrics is difficult to learn as there is no strict correspondence between lyrics and melody. Therefore, better learning methods are needed to mine the correlation between melody and lyrics, especially when limited paired data is available. 3) **evaluation metrics**. Previous works [10, 11] have proposed the pitch distribution (PD) metric, which calculates the intersection of the pitch probability distributions for two music pieces at the bar level. PD requires the prediction to be in the same length as the ground truth. However, this requirement is too strict for music generation and may limit the music modeling. In addition, these metrics do not consider semitone shifts that may happen in song writing. According to the chromatic scale [12], the twelve semitones in music can be formulated into a chromatic circle, which presents a strict cyclicality. As music scores generally follow this rule in composing, harmony could be easily maintained with consistent semitone offsets. For the aforementioned reasons, better evaluation metrics are desired in ASW.

In this paper, we aim to address these challenges with a unified framework for ASW. First, inspired by recent work in [7], we propose to leverage rich unpaired data to mitigate the scarcity issue of paired data. Different from [7] which only handles L2M and M2L, we propose a unified model framework that can handle L2L, M2M, L2M, and M2L at the same time without additional training parameters. Our model is optimized through different pre-training tasks following a two-stage pre-training strategy to minimize the domain difference and further boost the performance. Second, to make better use of limited paired data during fine-tuning, we propose a dual transformation loss, inspired by the similar strategy used to boost both automatic speech recognition (ASR) and text-to-speech (TTS) performance [13–17]. This loss can enhance the correlation learning between different modalities. Third, we propose a new evaluation metric involving the chromatic rule, dubbed the Soft Pitch Distribution (SPD), to accommodate variable-length melodies. Specifically, we explore the task of automatic pop song writing (paired melody and lyrics) in Chinese. To the best of our knowledge, our work is the first to explore an ASW system in this language [6], providing more insight into AWS performance across different languages. Extensive experiments demonstrate the effectiveness of our proposed training strategies over the unified framework (i.e., a two-stage pre-training and dual transformation loss), which significantly improve the quality of lyric and melody generation. The proposed SPD is shown to align with subjective evaluations from human listeners as well.

2. METHODOLOGY

Figure 1 illustrates the overview of our proposed ASW framework [1]. We apply the Transformer-based [18] encoder-decoder architecture with separated encoders and decoders for melody and lyrics generation. Unlike [19], we modify the decoders in order to handle all four

*Corresponding author.

[1]More details and project code can be found at https://github.com/DrWelles/ASW

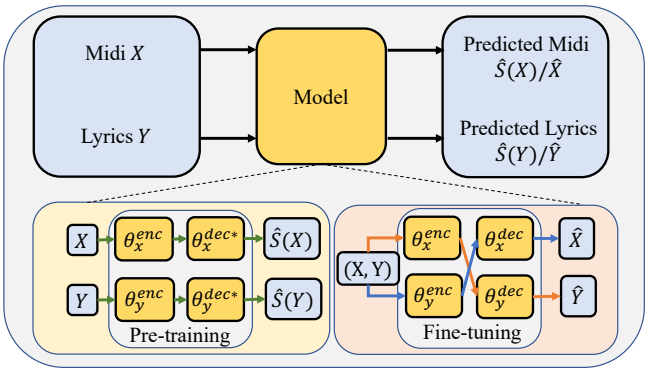

Fig. 1: Overview of our model framework. The green lines illustrate the pre-training stages, including task-specific and domain-specific pre-training (Section 2.1). The orange and blue lines stand for M2L and L2M tasks in fine-tuning with limited paired data (Section 2.2).

tasks in a unified model for ASW (details in Section 2.1). To mitigate the data scarcity issue mentioned in Section 1, two training strategies are proposed: (1) To utilize knowledge from unlabeled lyrics and melody, we introduce two pre-training stages. (2) We further fine-tune the network with dual transformation loss to strengthen the correlation between lyrics and melody based on limited paired data.

2.1. Pre-training with Unpaired Data

We optimize our proposed model with a two-stage pre-training strategy. The network is first trained based on unpaired data from various domains (e.g., jazz, classical music, etc.). It is then further trained based on unpaired pop music data to enable the model to learn in-domain knowledge for the target pop song writing. Related notations are defined in Table 1. The formulation of the M2M task is as follows:[2]

$$\begin{aligned} &L(X|\theta_x^{\text{enc}}, \theta_x^{\text{dec}^*}) \\ =&\Sigma_{x\in X}\Sigma_{i=1}^{l} \log P(S(x_{s_i:t_i})|x_{s_i:t_i}, \theta_x^{\text{enc}}, \theta_x^{\text{dec}^*}) \\ =&\Sigma_{x\in X}\Sigma_{i=1}^{l}\Sigma_{t=s_i}^{t_i} \log P(x_{t+1}|x_{\le t}, \theta_x^{\text{enc}}, \theta_x^{\text{dec}^*}) \end{aligned} \tag{1}$$

We skip the cross attention in the Transformer-based decoders [18], so that the encoder-decoder architecture can be switched to an encoder-only architecture, which enables the model to handle M2M and L2L as well in a unified manner. We maximize $L(X|\theta_x^{\text{enc}}, \theta_x^{\text{dec}^*})$ and $L(Y|\theta_y^{\text{enc}}, \theta_y^{\text{dec}^*})$ for melody and lyric pre-training based on unpaired data as shown in the pre-training procedure in Figure 1.

Stage 1, task-specific pre-training. There are some universal theories in the field of music, such as harmony theory, that is, music has some common elements regardless of theme and style. Therefore, in the first pre-training stage, we optimize the model to learn universal patterns based on all available unpaired data. Previous work [7] adopts the encoder-decoder pre-training task, which is suitable for M2L and L2M, but not for M2M and L2L. In our unified model framework, we apply the auto-regressive training in the first task-specific pre-training stage, so that it can handle all four ASW tasks.

Stage 2, domain-specific pre-training. As the goal of this work is pop song writing, the model pre-trained in the first stage based on general domain music data still has some significant domain gaps. Thus, we further conduct domain-specific pre-training only based on unpaired data in the pop music domain to reduce these gaps.

[2]The L2L task shares a similar formulation with lyric input Y.

Notation	Definition
$L(t\|c)$	The log likelihood of t condition on c
$X=\{x_i\}_{i=1}^{n}$	MIDI token sequence, x_i means i[th] token in X
$Y=\{y_i\}_{i=1}^{n}$	Lyrics sequence, y_i means i[th] token in Y
X_i	Pitches in i[th] bar in X
$X_{l:r} = \{x_i\}_{i=1}^{k}$	Pitches in i[th]-(j-1)[th] bar in X, k is the number of notes
$z_{l:r}=\{z_i\}_{i=l}^{r}$	Subsequence, $z \in \{x, y\}$
$z_{\le t}=\{z_i\}_{i\le t}$	Subsequence, $z \in \{x, y\}$, similar for $z_{<t}$
$S(Z)=\{z_i\}_{i=2}^{n+1}$	Shifted one step of Z, $Z \in \{X, Y\}$
$\hat{Z}$	Reconstructed Z, $Z \in \{X, Y, S(X), S(Y)\}$
θ_z^{enc}	Parameters of z domain encoder, $z \in \{x, y\}$
θ_z^{dec}	Parameters of z domain decoder, $z \in \{x, y\}$
$\theta_z^{\text{dec}^*}$	Parameters of z domain decoder without cross attention, $z \in \{x, y\}$

Table 1: Definition of formula symbols for Section 2 and 3. MIDI means midi file format, MIDI token means REMI [2] token in our work.

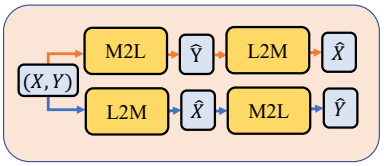

Fig. 2: Illustration of dual transformation loss defined in Section 2.2.

2.2. Fine-tuning with Paired Data

During the pre-training stages, the encoder-decoder for lyrics and melody are trained separately, and thus the model cannot yet handle M2L and L2M tasks. Moreover, separate training may lead to misalignment in the same latent space for θ_x^{enc} and θ_y^{enc}. Further processing may be needed to prevent the M2L and L2M parts from deviating from each other and enable the model to perform both of these tasks. Thus, we fine-tune the model with lyric-melody paired data following the two-stage pre-training. Given paired data (X, Y), the formulation of M2L is:

$$\begin{aligned} &L(Y|X, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) \\ =&\Sigma_{(x,y)\in(X,Y)}\Sigma_{t=1}^{|y|} \log P(y_t|y_{<t}, x, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) \end{aligned} \tag{2}$$

The supervised fine-tuning is applied to both L2M and M2L. The basic fine-tuning loss is:

$$L_{ftb} = -L(X|Y, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) - L(Y|X, \theta_y^{\text{enc}}, \theta_x^{\text{dec}}) \tag{3}$$

Dual transformation loss. Different from other sequence-to-sequence tasks (e.g., machine translation), melody and lyrics show weak correlation. To strengthen the relationship between the two modalities, we further introduce the dual transformation loss [13–17] in the fine-tuning phase. As illustrated by the orange lines in Figure 2, the M2L model transforms melody X to predicted lyrics $\hat{Y}$, and then the L2M model leverages the transformed pair $(X, \hat{Y})$ for reconstruction training. The formulation is as follows:

$$\begin{aligned} &L(X|\hat{Y}, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) \\ =&\Sigma_{(x,y)\in(X,\hat{Y})}\Sigma_{t=1}^{|y|} \log P(x^t|x_{<t}, \hat{Y}, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) \end{aligned} \tag{4}$$

Similarly, as illustrated by the blue lines in Fig 2, the L2M model transforms lyrics Y to predicted melody $\hat{X}$, and then the M2L model leverages the pair $(\hat{X}, Y)$ for training. We use a hyper-parameter α to weight the reconstruction loss. The final fine-tuning loss is:

$$L_{ft} = L_{ftb} - \alpha[L(Y|\hat{X}, \theta_x^{\text{enc}}, \theta_y^{\text{dec}}) + L(X|\hat{Y}, \theta_y^{\text{enc}}, \theta_x^{\text{dec}})] \tag{5}$$

3. EXPERIMENTS

In this section, we first describe the data processing pipeline, then introduce the experimental setup, and finally present the experimental results. In this work, our task is to automatically write pop songs in Chinese. As most related works focus on English song writing [6], to the best of our knowledge, our work is the first to explore ASW in Chinese, providing more insight into AWS across different languages.

3.1. Data Processing Pipeline

To obtain paired data for two modalities, we mine data from the Internet. The procedure of data collection can be roughly divided into data acquisition, singing separation, and representation extraction.

Data acquisition. For pre-training, we collect 17,699 songs (1,764h) in MIDI format for melody generation (Giant-MIDI [20], Maestro [21], Classic piano,[3] Pop piano [22]) and 189,456 songs for lyrics generation.[4] To get paired data, we crawl thousands of paired songs and corresponding lyrics from open domains. Then, we filter the songs that are longer than five minutes or shorter than one minute. We remove meta-information (such as composer, singer, etc.) in lyrics and unusual genre data. The resulting corpus includes 3,524 songs (237h) from 959 Chinese singers.

Singing and accompaniment separation. Almost all the crawled songs are mixed with singing and accompaniments, which introduce extra noise to the system. Therefore, we use Spleeter [23][5] to separate songs into singing voice and accompaniment tracks.

Preparing data for training. We utilize audio-to-midi [24] to convert the singing voice to MIDI format. Then, we use the melody extraction algorithm skyline [25] to convert polyphonic music to the monophonic melody. For the input feature of the network, we adopt the REMI [2] representation to convert the MIDI to token sequences.

3.2. Experimental Setup

Melody and Lyrics Generation. We use a Transformer-based encoder-decoder architecture [18] as our baseline [6], which consists of six encoder/decoder layers.[7] The hidden size of each layer is 128. The number of attention heads is 8. The model is trained on an NVIDIA 1080Ti, and the batch size is 32 with 300 tokens for each sequence in the batch. Dropout with the rate of 0.2 is used for training and dual transformation loss weight α is 0.05. Adam [26] optimizer with Noam learning-rate decay strategy and 10,000 warm-up steps are employed during the training. We select 20 songs for validation and test sets, respectively, and use the remaining data for training. The same split setting is adopted in the pre-training and fine-tuning phases. Our baseline (noted as B in experimental results) is the same aforementioned model with losses in all four tasks directly over our paired dataset without any pre-training and dual transformation loss.

Search strategy. According to the REMI [2] representation, the notes consist of three consecutive tokens ('velocity', 'pitch', and 'duration'), while tempos consist of 'class' and 'value' tokens. Instead of filtering invalid tokens as [2], we limit the search space during decoding. Specifically, notes and tempos in REMI sequences require certain orders of tokens. For example, the position token can be followed with note tokens, tempo tokens, and their attributes should appear consecutively.

3.3. Evaluation

Both objective and subjective metrics are used for evaluation. For M2M, the first 150 tokens from ground truth melody are used as the initial decoding condition, from which the model decodes for 800 steps. Similarly, L2M models decode 800 steps conditioned on the first 150 tokens from both ground truth melody and lyrics for each song. We then remove the initial condition from the generated samples and chunk each sample and ground truth into 15 seconds for a fair comparison.

3.3.1. Subjective Evaluation Metrics

For subjective evaluation, we invite 45 annotators with basic knowledge in music and singing to evaluate seven songs. We require each annotator to evaluate melody with respect to four aspects, with each aspect rated with an opinion score from one to five (bad to excellent). These aspects are: (1) Similarity: the overall similarity of the melody, including rhythm, genre, etc. (2) Continuity: is the melody stumbling? (3) Singability: is the melody easy to sing or not? And (4) Rhythm [27]: is the duration and pause of melody natural and in line with the genre?

3.3.2. Objective Evaluation Metrics

Different objective metrics have been proposed to evaluate the harmony, quality, and similarity between two musical pieces, for example, the pitch distribution (PD) [10, 11] and melody distance (MD) [28]. To summarize, we use perplexity (PPL) to measure the performance for four generation tasks (L2L, M2M, L2M, and M2L) roughly and adopt MD, PD and our metric, soft pitch distribution (SPD), as fine-grained objective metrics for melody objective experiments.

Pitch and Duration Distribution Similarity. The pitch distribution similarity (PD) [11] measures the similarity using average Overlapped Area (OA) between two distributions (normalized frequency histogram) of pitches in melodies. In [11], PD adapts hard alignment between music pieces, requiring the length of generated results to be equal to ground truth. In the following section, we refer it as hard PD (HPD). The formulation of HPD [11] is as following:

$$\begin{aligned} \text{HPD} =& \frac{1}{N_X} \Sigma_{i=1}^{N_X} \text{OA}(\text{Dis}(X_i), \text{Dis}(\hat{X}_i)) \\ =& \frac{1}{N_X} \Sigma_{i=1}^{N_X} \Sigma_{j \in T} \text{min}(\text{Dis}(X_i)_j, \text{Dis}(\hat{X}_i)_j) \end{aligned} \tag{6}$$

where $\text{OA}(X, \hat{X})$ is defined as the averaging pitch overlap between ground truth X and generation piece $\hat{X}$. The notation T means the possible values of pitch in REMI representation [2]. We define notation N_X as bar numbers of X and function Dis(S) returns the normalized frequency histogram of pitches of S (S can be X_i, $\hat{X}_i$ or other bars notation).

The aforementioned HPD does not consider the chromatic rule and it has a strict assumption that two music pieces should match strictly. To alleviate these problems, we consider chromatic rule [12] in our metric so as to allow consistent tonality shift (i.e., semitone shift). Moreover, dynamic time warping (DTW) [28] is applied to get bar-level alignment, which allows flexible alignments between music pieces. The new metric, namely soft pitch distribution (SPD) is formulated as follows:

[3] http://www.piano-midi.de

[4] https://www.datafountain.cn/datasets/46

[5] An open-source project of music separation that achieves state-of-the-art separation quality, https://github.com/snkirusi/spleeter

[6] The model structure is same as SongMASS [10].

[7] For L2L or M2M generation, we skip the cross attention in the corresponding decoder.

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.36	2.66	2.28	2.79
B*	2.54	2.78	2.59	2.95
B*+ C	**2.79**	**3.02**	**2.67**	**3.09**

(a) Subjective evaluation results of L2M

	Similarity	Continuity	Singability	Rhythm
GT	–	3.81	3.54	3.80
B	2.11	2.52	2.24	2.72
B*	2.39	2.66	2.49	2.85
B*+ C	**2.59**	**2.97**	**2.69**	**3.11**

(b) Subjective evaluation results of M2M.

Table 2: Subjective evaluation results of L2M and M2M, where the number represents mean opinion scores from all listeners. B stands for baseline (Section 3.2), C stands for dual transformation loss, and * stands for model with pre-training stages.

		MD (↓)	HPD (%, ↑)	SPD (%, ↑)
L2M	B	**20.20**	7.08	31.63
	B*	22.76	**15.07**	34.51
	B*+ C	30.7	10.58	**40.04**
M2M	B	38.09	6.30	28.57
	B*	23.29	**14.64**	31.48
	B*+ C	36.32	11.53	**38.35**

Table 3: Result of objective metrics for melody evaluation. B stands for baseline (Section 3.2), C stands for dual transformation loss, and * stands for model with pre-training stages. Detailed metrics are discussed in Section 3.3.2.

$$\begin{aligned}\text{SPD} =& \frac{1}{N_S + N_{\hat{S}}} \Sigma_{i=1}^{\text{m}} \text{OA}(\text{Dis}(\mathcal{D}_i), \text{Dis}(\hat{\mathcal{D}}_i)) \\ =& \frac{1}{N_S + N_{\hat{S}}} \Sigma_{i=1}^{\text{m}} \Sigma_{j \in \text{T}} \text{min}(\text{Dis}(\mathcal{D}_i)_j, \text{Dis}(\hat{\mathcal{D}}_i)_j)\end{aligned} \tag{7}$$

The matching results of DTW is two aligned sequence with length m: $P = \{p_i\}_{i=1}^{\text{m}}$ and $Q = \{q_i\}_{i=1}^{\text{m}}$, which means $X_{p_i:p_{i+1}}$ match $\hat{X}_{q_i:q_{i+1}}$. And $\mathcal{D}_i$= $\Delta X_{p_i:p_{i+1}}$ and $\Delta X_{l:r} = \{x_{i+1} - x_i\}_{i=1}^{k-1}$ (k means total number of notes in $X_{l:r}$ as Table 1). Similarly, $\hat{\mathcal{D}}_i = \Delta \hat{X}_{q_i:q_{i+1}}$ and $\Delta \hat{X}_{l:r} = \{\hat{x}_{i+1} - \hat{x}_i\}_{i=1}^{k-1}$.

Melody Distance. For Melody Distance (MD), we split notes into a time series of pitch according to the duration, with a granularity of 1/16 note. Considering the chromatic rule, we normalize the sequence by subtracting the average pitch of the entire sequence. We use DTW to measure the similarity between the generated result and ground truth with different lengths.

3.4. Results and Discussion

Subjective and objective evaluations. We perform subjective and objective experiments for M2M and L2M to verify the effectiveness of the proposed methods.

The subjective results in Table 2 show that both pre-training and dual transformation loss (i.e., B* and B*+ C) improve the opinion scores for both L2M and M2M, which demonstrates the effectiveness of our proposed framework and fine-tuning method.We also notice that L2M (Table 2a) shows better subjective scores than M2M (Table 2b). We think the reason is that L2M is given more context than M2M in the decoding step (details in Section 3.3).

	L2L	M2L	M2M	L2M	Average
B	16.85	17.18	2.19	2.12	9.59
B*	11.17	11.89	2.21	2.15	6.86
B*+ C	**11.10**	**11.84**	**2.18**	**2.00**	**6.78**

Table 4: The perplexity results of four generation tasks. B stands for baseline (Section 3.2), C stands for dual transformation loss, and * stands for model with pre-training stages.

	L2L/M2L	M2M/L2M	Average
Baseline	16.85/17.18	2.19/2.12	9.59
+ S1	11.49/11.85	2.30/2.29	6.98
+ S2	11.34/12.14	2.28/2.25	7.01
+ S1 + S2	**11.10/11.84**	**2.18/2.00**	**6.78**

Table 5: The perplexity results with different pre-training setting. S1 and S2 stand for the two pre-training stages in Section 2.1.

Table 3 shows the comparison between our proposed SPD with two existing evaluation metrics (i.e., MD and HPD). The MD trend of L2M and M2M in Table 3 do not align with the subjective results in Table 2. For example, there is an abnormal jitter of HPD (B* is much higher than B*+ C for both L2M and M2M) in Table 3, which does not agree with the increasing trend of subjective scores. Our proposed evaluation metric SPD shows nice agreement with the subjective results, which indicates that the flexible setting and chromatic rule [12] do improve the evaluation quality.

Ablation study of the model. Table 4 shows the general objective evaluation of the four tasks in ASW. We can see that pre-training significantly outperforms the baseline in L2L and M2L and slightly improves the performance in M2M and L2M. Dual transformation loss is shown to be effective as well.

Ablation study of pre-training stages. To investigate the impact of pre-training, we conduct ablation experiments on different pre-training settings. The results are shown in Table 5. The results show that the general knowledge from various domain data is beneficial. The difference between S1 and S1+S2 indicates that the domain gap between genres may also limit the model performance.

4. CONCLUSION

In this paper, we propose a unified ASW framework for M2M, L2L, M2L, and L2M without additional training parameters. Due to the scarcity of paired data, we adopt pre-training and fine-tuning paradigms to take advantage of a large amount of unpaired data. To better use the limited paired data, we introduce dual transformation loss to further boost the performance in the fine-tuning stage. Furthermore, we propose an objective music evaluation metric, SPD, which removes some strict assumptions adopted in previous works. Our work is the first to explore ASW in Chinese, providing more insight into AWS performance across different languages. Extensive experiments demonstrate that the proposed training strategies over the unified framework outperforms the baseline, and the SPD metric is more stable and aligned with subjective opinion scores from human listeners.

5. ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China (No. 62072462) and the National Key R&D Program of China under Grant No.2020AAA0108600.

6. REFERENCES

[1] Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi, "Rigid formats controlled text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 742–751.

[2] Yu-Siang Huang and Yi-Hsuan Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1180–1188.

[3] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, et al., "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[4] Hangbo Bao, Shaohan Huang, Furu Wei, Lei Cui, Yu Wu, et al., "Neural melody composition from lyrics," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2019, pp. 499–511.

[5] Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma, "icomposer: An automatic songwriting system for chinese popular music," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 84–88.

[6] Yi Yu, Abhishek Srivastava, and Simon 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.

[7] Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, et al., "Songmass: Automatic song writing with pre-training and alignment constraint," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 13798–13805.

[8] Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, et al., "A melody-conditioned lyrics language model," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 163–172.

[9] Yihao Chen and Alexander Lerch, "Melody-conditioned lyrics generation with seqgans," in *2020 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2020, pp. 189–196.

[10] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu, "Mass: Masked sequence to sequence pre-training for language generation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5926–5936.

[11] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, et al., "Popmag: Pop music accompaniment generation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1198–1206.

[12] Ian Quinn, "Tonal harmony," in *The Oxford Handbook of Critical Concepts in Music Theory*. Oxford University Press, 2019.

[13] Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Speech chain for semi-supervised learning of japanese-english code-switching asr and tts," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 182–189.

[14] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, et al., "Almost unsupervised text to speech and automatic speech recognition," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5410–5419.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96.

[16] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, et al., "Dual learning for machine translation," *Advances in neural information processing systems*, vol. 29, pp. 820–828, 2016.

[17] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 301–308.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[20] Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang, "Giantmidi-piano: A large-scale midi dataset for classical piano music," *arXiv preprint arXiv:2010.07061*, 2020.

[21] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, et al., "Enabling factorized piano music modeling and generation with the maestro dataset," *arXiv preprint arXiv:1810.12247*, 2018.

[22] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 178–186.

[23] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, pp. 2154, 2020.

[24] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang, "A streamlined encoder/decoder architecture for melody extraction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 156–160.

[25] Alexandra Uitdenbogerd and Justin Zobel, "Melodic matching techniques for large music databases," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 1999, pp. 57–66.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[27] Hongyuan Zhu, Qi Liu, Nicholas Jing Yuan, Chuan Qin, Jiawei Li, et al., "Xiaoice band: A melody and arrangement generation framework for pop music," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2837–2846.

[28] Donald J Berndt and James Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*. Seattle, WA, USA:, 1994, vol. 10, pp. 359–370.