Fenfei Guo, Chen Zhang, Zhirui Zhang, Qixin He, Kejun Zhang, Jun Xie, and **Jordan Boyd-Graber**. **Automatic Song Translation for Tonal Languages**. *Findings of the Association for Computational Linguistics*, 2022, 9 pages.


```
    @article{Guo:Zhang:Zhang:He:Zhang:Xie:Boyd-Graber-2022,
Author = {Fenfei Guo and Chen Zhang and Zhirui Zhang and Qixin He and Kejun Zhang and Jun Xie and Jordan B
Title = {Automatic Song Translation for Tonal Languages},
Journal = {Findings of the Association for Computational Linguistics},
Year = {2022},
Location = {Dublin},
Url = {http://cs.umd.edu/~jbg//docs/2022_acl_ast.pdf},
}
```


Links:

- Translation Examples (with sound) [https://gagast.github.io/]
- Code [https://github.com/GagaST/GagaST-code]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Automatic Song Translation for Tonal Languages

Fenfei Guo
University of Maryland
fguo1@umd.edu

Chen Zhang
Zhejiang University
zc99@zju.edu.cn

Zhirui Zhang
Tencent AI Lab
zrustc11@gmail.com

Qixin He
Purdue University
heqixin@purdue.edu

Kejun Zhang
Zhejiang University
zhangkejun@zju.edu.cn

Jun Xie
Alibaba DAMO Academy
qingjing.xj@alibaba-inc.com

Jordan Boyd-Graber
CS, iSchool, UMIACS, LSC
University of Maryland
jbg@umiacs.umd.edu

Abstract

This paper develops automatic song translation (AST) for tonal languages and addresses the unique challenge of aligning words' tones with melody of a song *in addition to* conveying the original meaning. We propose three criteria for effective AST—preserving meaning, singability and intelligibility—and design metrics for these criteria. We develop a new benchmark for English–Mandarin song translation and develop an unsupervised AST system, Guided AliGnment for Automatic Song Translation (GagaST), which combines pre-training with three decoding constraints. Both automatic and human evaluations show GagaST successfully balances semantics and singability.

1 Introduction

Suppose you are asked to translate the lyrics "Let it go" from the Disney musical *Frozen* into Mandarin Chinese. Some good, literal translations of this would be A) "fàng shǒu", B) "fàng shǒu ba" or C) "ràng tā qù ba" (Figure 1); these get the meaning across and are the domain of traditional machine translation. However, what if you needed to sing this song *in Mandarin*? These literal translations simply do not work: Translations A and C do not match the number of notes and break the original rhythm; while the tones of Translation B does not match the pitch flow of the original melody.

Song translation, unlike translation lyrics for understanding (subtitling), aims to translate the lyrics so that it can be *sung* with the original melody. Therefore, the translated lyrics must match the prosody of the pre-existing music in addition to retaining the original meaning. In *Singable Translations of Songs*, Low (2003) says, this is an uncommon and an unusually complex task, a translator consider rhythm, notes' pitches, phrasing, and stress. Nonetheless, there are cultural and commercial incentives for more efficient song translation; *Frozen* alone made over a half a billion dollars in non-English box office receipts[1] and the musical *Les Misérables* has been performed in over a dozen languages on stage.

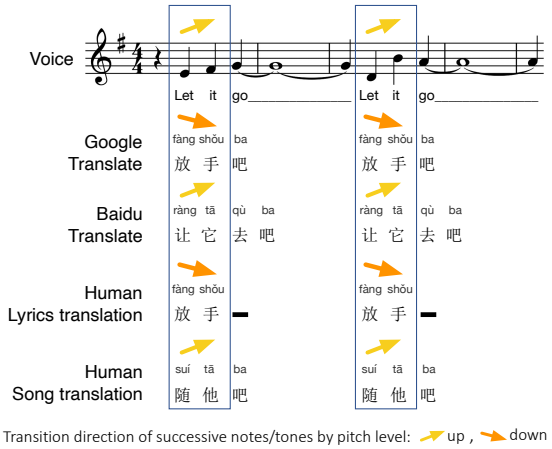

Figure 1: Example Mandarin translations for "Let it go" in *Frozen*. Of these, only the official human song translation is something a singer could actually sing: it fits the length of the notes and matches the tones with the pitch of notes. GagaST finds translations that satisfy these constraints.

As we discuss in Section 2, while translating Western songs resembles poetry translation, translating into *tonal* languages (e.g., Mandarin, Zulu and Vietnamese) introduces new problems. In tonal languages, a word's pitch contributes to its meaning (Figure 2); when singing in tonal languages, the

[1] https://www.the-numbers.com/movie/Frozen-(2013)#tab=international

Findings of the Association for Computational Linguistics: ACL 2022, pages 729 - 743

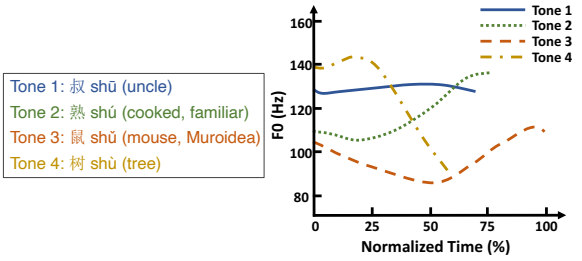

Figure 2: In total languages like Mandarin, the pitch changes the meaning of the words (left). Each of the four tones in Mandarin (right) has a different pitch profile. Figure from Xu (1997).

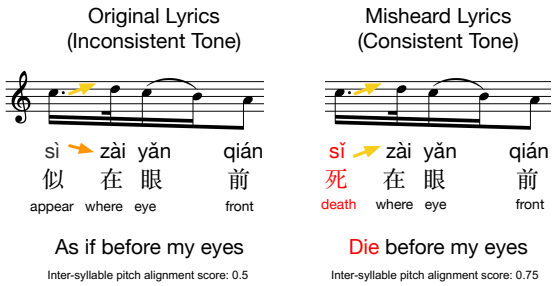

Figure 3: If a song's music doesn't match the tones of the lyrics, it can cause the hearer to misunderstand the lyrics. In this example, someone can hear "sǐ zài" instead of "sì zài", because the notes are going up and "sì zài" is going down.

tones of translated words must align with the "flow" of the pitches in the music (Section 2.1). For example, if "fáng shǒu" were sung instead of "fàng shǒu" (because notes are going up), a listener might hear "defensive" instead of the intended meaning.

This paper builds the first system for automatic song translation (AST) for one tonal language—Mandarin. Section 3 proposes three criteria—*preserving semantics*, *singability* and *intelligibility*—needed in an AST system.

Guided by those goals, we propose an unsupervised AST system, Guided AliGnment for Automatic Song Translation (GagaST). GagaST begins with an out-of-domain translation system (Section 4.1) and adds song alignment constraints that favor translations that are the appropriate length and whose tones match the underlying music (Section 4.2). Naturally, such constraints trade-off between semantic meaning and singability/intelligibility. Section 5.4 discusses this trade-off between song alignment scores and the standard machine translation metric, BLEU.

These criteria also form the evaluation for our initial evaluation (Section 5.3). However, we go beyond an automatic evaluation through a human-centered evaluation from musicology students. GagaST creates singable songs that make sense given the original text, and our proposed alignment scores correlate with human judgements (Section 5.4.3).[2]

2 Background: Prose, Poetry, and Song Translation

A spoken language can be divided into two forms: prose, which corresponds to natural conversation and conventional grammatical structure; and verse—typically rhythmic and broken into stanzas–such as poetry and song lyrics.

The vast majority of machine translation research has been focused on prose translation and has made huge progress; in contrast verse translation is more difficult as it must obey the rhythmic constraints and is less developed. In his *tour de force* work *Le Ton Beau de Marot*, Douglas Hofstadter presents eighty-nine translations of a single poem to capture the panoply of considerations of what makes the task difficult (Hofstadter, 1997).

In western verse, the rhythmic structure are mostly defined by meter, such as the iambic pentameter for sonnets, which defines the length of each line, the patterns of long syllables versus short ones and the stressed ones versus weak ones. Existing work (Greene et al., 2010; Ghazvininejad et al., 2018) use finite-state constraints to encode both meter and rhyme.

Song translation, on the other hand, can be viewed as a translation where the *melody* defines the constraints. Reproducing *all* of the essential values of a song—perfectly matching the meaning, perfectly singable, and perfectly understandable—is an impossible ideal (Franzon, 2008). Thus, tradeoffs are unavoidable. Low (2003) argues for prioritizing *singability* over other qualities such as *sense* and *rhyme* since "effectiveness on stage" is a practical necessity. Tonal languages (e.g., Mandarin, Zulu and Vietnamese) dramatically increases the complexity of singability, and introduces a new factor that could hamper intelligibility.

[2]Examples of translated songs by GagaST at https://gagast.github.io/posts/gagast.

2.1 Song Translation for Tonal Languages

For tonal languages, pitch contributes to the meaning of words. In a conservative estimation, fifty to sixty percent of the world's languages are tonal (Yip, 2002) and cover over 1.5 billion people. For the lyrics to be *intelligible*, the speech tone and music tone should be correlated (Schneider, 1961). If not, the pitch contour could override the intended tone, which could produce different meanings. This is not just a theoretical consideration; Figure 3 shows how lyrics can be and have been misunderstood.[3]

2.2 Mandarin Tones and how to Sing them

Schellenberg (2013) summarizes the rules of singing with tone with a focus on Chinese dialects. The tonal system of Mandarin has two components:

- **The pitch level and shape of tones.** Four Mandarin tones are used since the 19th century (Figure 2). We denote tones with a diacritic over the vowel whose shape roughly matches the shape of the tone. The four tones are a high level (tone 1, e.g., shūo), rising (tone 2, yú), falling-rising (tone 3, wǒ) and falling (tone 4, huài).
- **The sandhi of tones.** Some combinations of tones have difficult articulatory patterns, so words that might normally have one tone might take another depending on the context. For example "nǐ" (you) and "hǎo" (good) are typically both third tone, but when they are together it is pronounced as "ní hǎo" (hello), with the first syllable changing to a *second* tone. These changes are called sandhi (Xu, 1997; Hu, 2017).

Mandarin tones interact with a sung melody in two ways (Yinliu et al., 1983; Schellenberg, 2013) to ensure lyrics are intelligible. First, at a local level, the *shape of tones* of individual syllables should be consistent with the musical notes they are matched with; for example, in "Love Island" (Figure 4), "shàng" in the blue block has the "falling" shape and the group of notes it assigned to it also falls from an A to a E. Second, and a global level, the music's *pitch contour* should align with the tones of the corresponding syllables (taking sandhi into account). In practice, we align the transitions between successive syllables and successive notes (Figure 5) ensuring that the tone matches the relative pitch change (Schellenberg, 2013).

[3]More examples at https://gagast.github.io/posts/gagast/#misunderstanding_examples

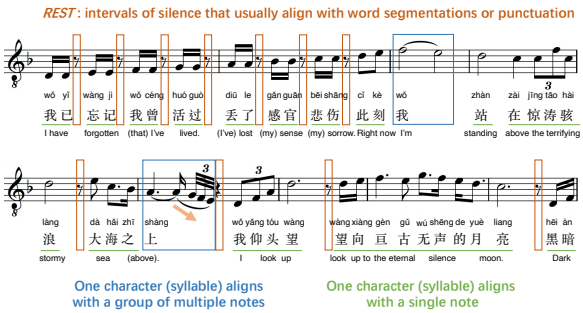

Figure 4: The output of a song translation needs to align syllables to the reference melody. There are several options, as evinced by the song "Love Island (xīn dǎo)". Orange (top): *REST* notes; Blue (bottom left): one syllable is assigned to a group of multiple notes (which needs *tone shape* alignment: the down arrow matches with falling tone of "ràng"); Green (bottom right): one syllable is assigned with one note.

3 AST for Tonal Languages

This section formally defines automatic song translation (AST) for tonal languages and introduce three criteria for what makes for a good song translation. These criteria form the foundation for the quantitative metrics we use in the experiment.

3.1 Criteria

There are three criteria that a singable song translation needs to fulfil.

Preserve meaning. The translated lyrics should be faithful to the original source lyrics.

Singability. Low (2003) defines singability as the phonetic compatability of translated lyrics and music. The translated song needs to be sung without too much difficulty; difficult consonant clusters, cramming too many syllables into a line, or incompatible tones all impair the singability.

Intelligibility. The translated song need to be understood by the listener. This quality has two components. First, could a listener produce *any* transcription of the lyrics. If the lyrics are too fast or garbled because the keywords do not fit well with the music, the lyrics are unintelligible. Beyond this basic test of recognizability, the lyrics must also be accurate: does this transcription match the intended meaning. Both aspects matter for a stage performance, since the audience should understand the content to follow the plot. For pop songs, not understanding all contents could be acceptable for some audiences; for example, Adriano Celentano's

notes	A3	C4	D4	*REST*	F4	G4	F4	F4
pitch level	57	60	62		65	67	65	65
duration	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	1	3	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
syllables	How	a-	bout		love?			

Table 1: A snippet of the song "Seasons of love" from the musical *Rent* that shows the input into GagaST. Notes are converted into integer pitches with a duration, and syllables are aligned to notes: the "a" from "about" has one note but "love" has four.

Prisencolinensinainciusol sacrifices all intelligibility for singability (Bellos, 2013). However, in more traditional media, hilarious misheard lyrics can ruin the audience's experience (Figure 3).

3.2 Task Definition

We define the AST task as follows: given an aligned pair of melody M and source lyrics X, generate translated text Y in the target language that aligns with the input melody M.

Specifically, $X = [x_1, ..., x_L]$ are the input lyrics with L syllables. Each syllable x_i is aligned to a snippet of the melody (Table 1) represented by a sequence of notes. To represent this to our algorithm, each syllable is aligned to three components of the melody:

1. A sequence of pitch values $\mathbf{p}_i \equiv [p_i^0, \dots]$ with $|\mathbf{p}_i| \geq 1$ where an integer value of 1.0 means a semitone (e.g., between C and C-sharp).
2. The duration of those notes $\mathbf{d}_i \equiv [d_i^0, ...]$, where 1.0 is a quarter note. Because it encodes the duration of each note, the length of $\mathbf{d}_i$ must be the same as the length of $\mathbf{p}_i$.
3. Sometimes there is a rest (pause) before a lyric is sung. We align this to the following syllable i. The scalar r_i is the real-valued duration of the *REST* note before note group $\mathbf{p}_i$. If no *REST* exists before $\mathbf{p}_i$, $r_i = 0.0$.

3.3 Constraints for Aligning Lyrics to Music

To make translated songs singable and intelligible, we summarize three desirable properties of that the AST lyric outputs should have if they are to match the underlying melody. Each of these induces a score function which we will use both in our objective functions for constrained translation and for our evaluation metrics.

3.3.1 Length Alignment

The number of syllables L_y in translated lyrics Y need to match the number of groups of notes $\mathbf{p}_i$ in the melody M, so that it can be sung with the music. Within the scope of this paper, we either keep the original grouping in the melody M and have $L_y = L_x$ for reproducing the original music; or strictly produce one target syllable for each single note in the melody.

3.3.2 Pitch Alignment

For tonal languages, pitch of the music must match the lyrics. As in Section 2.2, there are two types of pitch alignments: 1) *intra-syllable*, the tone shape of each syllable (Figure 4 blue box) should align with the shape of the assigned group of notes; 2) *inter-syllable*, the overall pitch contour of the music phrase should align with the tones of lyrics.

Intra-syllable alignment. For an *individual* syllable, if it is assigned to more than one note (e.g., "love" in Table 1), those notes must be consistent with the shape of the syllable's tone (Wee, 2007). For Mandarin, there are four tones (Xu, 1997, Figure 2). We estimate the shape of the multi-note sequence $\mathbf{p}_i$ by least-square estimation and classify it into one of five categories: level, rising, falling, rising-falling, falling-rising.

Specifically, for each group $\mathbf{p}_i$ that $|\mathbf{p}_i| > 1$, we classify it as,

1. "level", if $p^i_{max} - p^i_{min} \leq 1.0$; otherwise, we fit $\mathbf{p}_i$ into $ax^2 + bx + c$ via least-square estimation, and compute the axis of symmetry $l = -b/2a$,
2. "rising", if ($l \leq p_i^0$ and $a > 0.0$) or ($l \geq p_i^{-1}$ and $a < 0.0$);
3. "falling", if ($l \leq p_i^0$ and $a < 0.0$) or ($l \geq p_i^{-1}$ and $a > 0.0$);
4. "rising-falling", if $p_i^0 < l < p_i^{-1}$ and $a < 0.0$;
5. "falling-rising", if $p_i^0 < l < p_i^{-1}$ and $a > 0.0$;

We compare the shape with that of syllable y_i, and compute the intra-syllable alignment score S^i_{intra}:

$$S^i_{\text{intra}} = \begin{cases} 1.0 & \text{if the shape matches,} \\ \epsilon & \text{otherwise,} \end{cases} \quad (1)$$

where ϵ is a small parameter that allows for mismatches. Of the five patterns, "level" can match with any tone, "rising" matches with tone 2 (yú), "falling" matches with tone 4 (huài), "falling-rising" matches with tone 3 (wǒ) while "rising-falling" matches no Chinese tones.

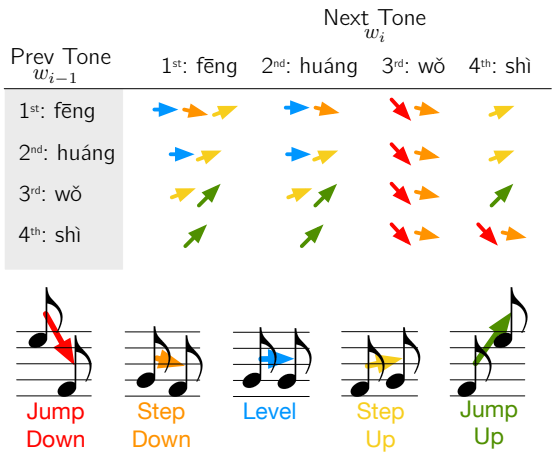

Figure 5: For translated songs in Mandarin to be singable, music notes should align the tones of successive characters; this becomes our **inter-syllable pitch alignment**. The arrows show acceptable transitions in music for two successive Mandarin characters (w_{i-1}, w_i) based on the shape of Mandarin tones including sandhi.

Inter-syllable alignment. The second constraint compares the transition directions between consecutive tones (t_{i-1}, t_i) of successive syllables (y_{i-1}, y_i) that belong to the same word (see arrows in Figure 3). These must match the transition directions of music notes $(\mathbf{p}_{i-1}, \mathbf{p}_i)$.[4] Each transition (the movement from one syllable/note to the next) can be categorized as *level*, *step up*, *jump up*, *step down* and *jump down*. We summarize the acceptable transitions for each pair of successive syllables in Figure 5 based on analysis by Yinliu et al. (1983) and we discuss our choices with more details in Appendix A.2. Given two syllables (y_{i-1}, y_i), we compute the local pitch contour S^i_{inter}:

$$S^i_{\text{inter}} = \begin{cases} 1.0 & \text{if contour matches,} \\ \epsilon & \text{otherwise,} \end{cases} \quad (2)$$

where ϵ again is a small value to allow mismatches.

3.3.3 Rhythmic Alignment with Word Segmentation in Mandarin

A musical *REST* is a silence separating music. Recall that in our setup of the data, a scalar r_i denotes if a note precedes syllable i. In any language, it is uncommon for a rest to break up a word's syllables. Thus a good translation should avoid this. For Mandarin, creating metrics that capture this are slightly more complicated because translation systems typically do not explicitly generate word boundaries. Thus, we must rely on the output of segmentation systems to know where word boundaries are.

An exception to this is punctuation (Figure 4). If a comma, period, or other punctuation is attached to the previous syllable y_{i-1}, then that is a clear signal that it's fine to pause between them. Thus, our rest score a syllable y_i following y_{i-1} that are part of different words with probability P_{seg}[5], the rest score is:

$$S^i_R = \begin{cases} 1.0 & \text{if } r_i > 0.0 \text{ and [punc] after } y_{i-1}, \\ 1.0 & \text{if } r_i = 0.0, \\ P_{\text{seg}} & \text{if } r_i > 0.0 \text{ and not [punc]}, \\ \epsilon & \text{otherwise,} \end{cases} \quad (3)$$

where ϵ is a parameter that represents our tolerance of having a rest within a word.

4 GagaST

Ideally, we would build an AST system for English–Mandarin song translation with data-driven models from parallel data, i.e., aligned triples (M, X, Y). However, these data are not available in the quantity or quality necessary for Mandarin: there is not enough data of any quality, and those that do exist have errors in the syllable-notes alignment. Thus, we propose an unsupervised AST system, Guided AliGnment for Automatic Song Translation (GagaST). For the pre-training, we collect non-parallel lyrics data in both English and Mandarin, as well as a small set of lyrics translation data (Section 5.1).

4.1 Song-Text Style Translation

To produce faithful translations in song-text style, we pre-train a transformer-based translation model with cross-domain data: translation data in the general domain, the collected monolingual lyrics data, and a small set of lyric translation data. We append domain tags (Figure 6) before each input example to control the model to produce translations merely in lyrics domain during song translation. For monolingual lyrics data, we adopt BART pre-training (Lewis et al., 2020).

4.2 Music Guided Alignment Constraints

Without available parallel data to learn the lyric-melody alignments, we impose constraints (Sec-

[4] We compute the directions of two notes group $(\mathbf{p}_{i-1}, \mathbf{p}_i)$ by the first notes (p^0_{i-1}, p^0_i) for simplicity.

[5] In practice, we use the cut output by the Jieba toolkit.

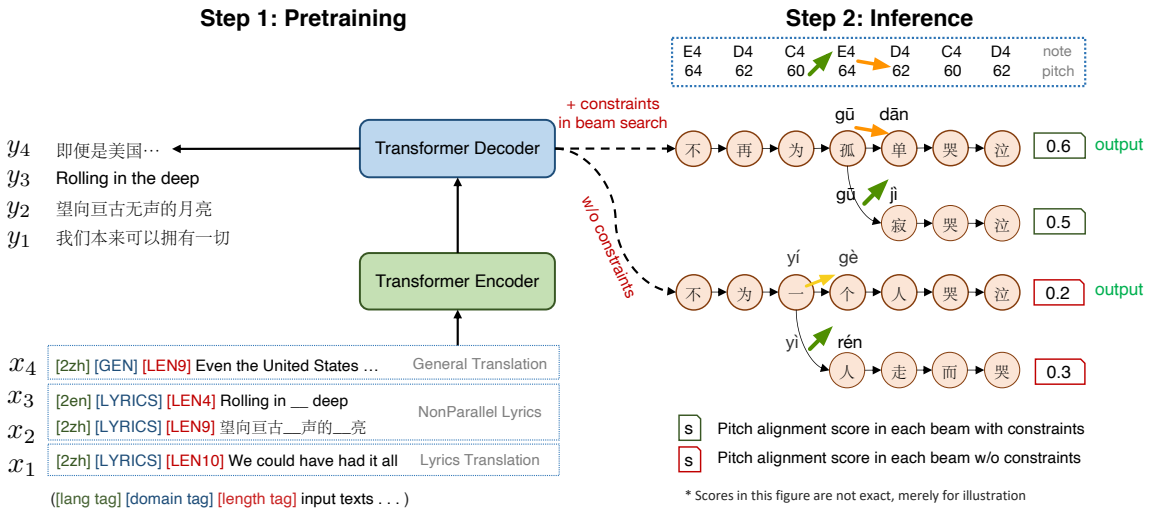

Figure 6: Overview of GagaST for English–Mandarin song translation. We first pre-train a lyrics translation model with mixture domain data (left); and then add alignment constraints in decoding scoring function during inference (right), we use unconstrained version as our baseline in the experiment.

tion 3.3) in the decoding phase. Specifically, since all constraints are applied at the unigram (intra-syllable, *REST*) or bigram (inter-syllable, *REST*) level, we apply them at each step of beam search as rewards and penalties in the scoring function:

$$\log P(Y \mid X, M) = \sum_{i=0}^{L} [\log P(y_i \mid y_{i-1:0}, X) + \lambda_{\text{inter}} \log S^i_{\text{inter}} + \lambda_{\text{intra}} \log S^i_{\text{intra}} + \lambda_R \log S^i_R], \quad (4)$$

where S_{inter}, S_{intra}, and S_R refer to the alignment scores for inter-syllable pitch alignment, intra-syllable pitch alignment and the rhythm alignment by *REST*. We introduce three tunable parameters—λ_{inter}, λ_{intra}, and λ_R—that control the importance of each of the song-specific constraints.

4.3 Length Control in Pre-training

To meet the length constraints, we pre-define the syllable-notes assignments with two strategies:[6] 1) *note-to-syllable*, i.e., for each note, we produce one syllable; 2) *syllable-to-syllable*, we use the original notes grouping in the input melody, and assign one syllable to each note group. In this case, the length of target translation is known. Following Lakew et al. (2019), we use length tag "[LEN$i]" to control the length of outputs during pre-training, where $i refers to the length of the target sequence.

[6] A dynamic mapping between the note sequence and the syllables changes the original rhythm and increases the search space exponentially. We leave this to future work.

5 Generating Melody-constrained Lyrics and Validating Singability

This section details data sets, model configuration, and proposed evaluation metrics. Then we analyze the results and the trade-offs inherent in song translation. Our code and data are open-sourced at https://github.com/GagaST.

5.1 Training Datasets and Model Configuration

WMT dataset: news commentary and back-translated news datatsets from WMT14 (29.6 million en2zh sentence pairs). No Cantonese texts included and the official Chinese texts can be pronounced in Mandarin by default.

Monolingual lyrics data: monolingual lyrics in both Mandarin and English collected from the web (12.4 million lines of lyrics for Mandarin and 109.5 million for English after removing duplicates).

Lyrics translation data: a small set of lyrics translation data crawled from the web [7] (140 thousands pairs of English-to-Mandarin lines). These translations are not singable.

We preprocess all training data with fastBPE (Sennrich et al., 2016) and a code size of 50,000. We use encoder-decoder Transformer (Vaswani et al., 2017) with 768 hidden units, 12 heads, GELU activation, 512 max input length, 12-12 layers structure (Appendix B for more details).

[7] https://lyricstranslate.com/

5.2 Evaluation Datasets

For evaluation, we need aligned triples (melody M, source lyrics X, target reference lyrics Y), where two conditions hold: 1) M and X are syllable-to-note aligned; 2) the reference Y should be singable and intelligible. With the confluence of digitization and copyright making such resources rare, we choose fifty songs from the lyrics translation dataset that have open-source music sheets on the web and create aligned triples manually. However, the reference lyrics in this dataset are not singable (our primary goal!), we use them to validate that the translations preserve the original meaning. Twenty songs comprise the validation set (464 lines) and thirty songs comprise the test set (713 lines).

5.3 Evaluation Metrics

An AST system for tonal languages should generate translated songs that are singable and intelligible while conveying the original meaning. Evaluating such system is an intrinsically hard task since all three qualities can be qualitative. Especially for preserving meaning, the lack of gold references and the greater tolerance for a loose translation in songs make it difficult to say how much semantic divergence is acceptable. Therefore, we first establish evaluations based on the relationship between lyrics and music and then design human annotations for more qualitative evaluations.

5.3.1 Objective Evaluation

Section 3.3 outlines three constraints inspired by music and linguistic theory. Because these constraints are directly incorporated into the decoding objective (Equation 4), these will be better than an unconstrained translation. However, we want to understand the trade-off between these new objectives and traditional translation evaluations.

To control for the length of the sentence, we normalize the score to 0–1.0 by the length of alignment pairs L_i, that is, based on Equation 1,2 and 3,

$$s_{[\cdot]} = \sum_{1}^{L_i} S^i_{[\cdot]} / L_i, \tag{5}$$

For the **length constraint**, we compute: 1) N_l, the number of samples that has length longer than the predefined length L_i; 2) N_s, that are shorter than L_i. For each case we compute the average error ratio of $\{\Delta l_i / L_i\}_1^{N_{[\cdot]}}$. For **meaning**, although we lack gold singable translations, we follow the common practice and calculate BLEU (Papineni et al., 2002) between the translated songs and the prose translation.

5.4 Trade-offs between Meaning and Melody-lyric Alignments

GagaST adds constraints in the decoding scoring functions to enforce lyric-music alignments; however, there are trade-offs between preserving meaning and adhering to these constraints. To select the importance of these constraints in decoding, we vary the value of the corresponding parameter λ (Equation 4) and analyze how much the BLEU score falls on the validation set as we increase the influence of the parameter. We set the hyper-parameters where the alignment scores increase fast while the BLEU decreases slowly. The *REST* constraint does not affect the BLEU (Table 2) but does alter ammount of punctuation. Working off the assumption that excessive punctuation is bad, we select a parameter that minimizes the mismatches between the *REST* and word boundaries. We choose (Figure 7) $\lambda_{\text{inter}} = 0.5$; $\lambda_{\text{intra}} = 1.0$; $\lambda_R = 1.5$ for all subsequent experiments.

5.4.1 BLEU Evaluation

Table 2 compares GagaST as we ablate constraints with our two syllable to note alignment strategies (Section 4): note-to-syllable and syllable-to-syllable. As in previous work, the length tag "[LEN$i]" helps lyrics fit the notes available. In all cases, less than 30 out of 713 lines produces a longer sentence with ratio less than 0.22; and no short cases. Thus, because it most closely resembles prior work in controlled translation and works well in this task, we adopt GagaST with only length tags and no other constraints as our baseline. With all of the constraints, GagaST indeed increases both pitch and rhythm alignments. It almost doubles the pitch contour alignment score, which affect the intelligibility the most.

However, these gains come at the cost of BLEU score.[8] While we believe that the audience would be more accepting of a less-than-literal translation in a song if it sounds better, we need a qualitative evaluation to validate that hypothesis.

[8] Due the paucity of reliable references, BLEU scores do not correlate with human judgement. For example, three official Disney Mandarin song translations have a lower BLEU score (12.3) than our more literal but demonstrably worse automatic translations.

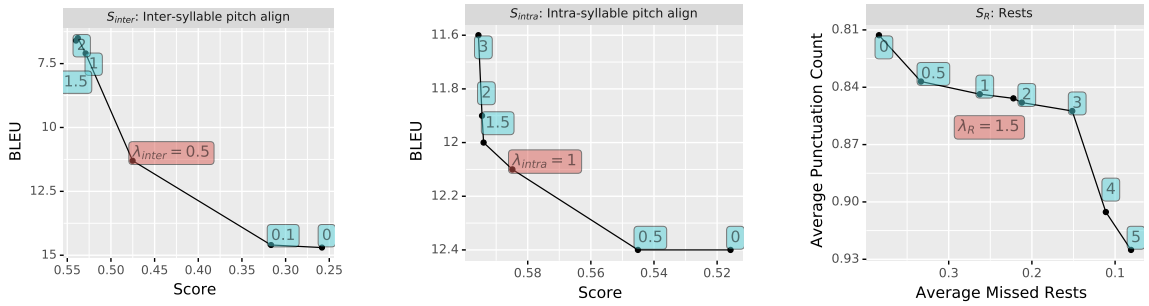

Figure 7: Trade-off between meaning (y-axis) and lyric-music alignments (x-axis) while adjusting the tuning parameter λ on the validation set. The selected value for the tuning parameter λ for downstream experiments is shown in red (preceeded by $\lambda =$). REST constraints do not affect BLEUs, but increase the number of [punc]s, which impairs the fluency of the lyrics, thus we select its parameter based on number of [punc]s.

Syllable-notes Assignment	Model	Pitch inter ↑	Pitch intra ↑	Rhythm avg # of missed rests ↓	Length longer ↓	Length shorter ↓	Meaning BLEU ↑
note-to-syllable	GagaST w/o constraints	0.28	-	0.53	9 (0.09)	0	24.0
	GagaST	0.51	-	0.31	26 (0.21)	0	16.9
	–only inter-syllable	0.51	-	0.45	26 (0.21)	0	16.8
	–only rest	0.28	-	0.31	11 (0.09)	0	23.8
syllable-to-syllable	GagaST w/o constraints	0.29	0.49	0.62	4 (0.12)	0	22.1
	GagaST	0.50	0.55	0.28	13 (0.13)	0	15.9
	–only inter-syllable	0.51	0.50	0.42	7 (0.12)	0	15.8
	–only intra-syllable	0.29	0.56	0.44	4 (0.12)	0	21.6
	–only rest	0.29	0.49	0.28	5 (0.12)	0	21.6

Table 2: Our song-specific constraints with two syllable alignment techniques. All results here use the same pre-training checkpoint and length tags are applied. For length score, 9 (0.09) means that 9 out of 713 samples are longer than the predefined length with an average ratio 0.09. All constraints have an effect, but inter-syllable pitch alignment has the largest.

5.4.2 Qualitative Evaluation

The true test of whether AST works is whether the songs can be sung, understood, and enjoyed. Thus, we follow Sheng et al. (2021) and show annotator from a music school students the resulting sheet music, ask their opinion, and ask them to sing the songs. We randomly select five songs from the test set and show the music sheets (see Appendix C) of the first ten sentences of each translated song to five annotators.

Following mean opinion score (Rec, 1994, MOS) in speech synthesis, we use five-point Likert scales (1 for bad and 5 for excellent). And we evaluate the songs on four dimensions: 1) *sense*, fidelity to the meaning of the source lyric; 2) *style*, whether the translated lyric resembles song-text style; 3) *listenability*, whether the translated lyric sounds melodious with the given melody; 4) *intelligibility*, whether the audience can easily comprehend the translated lyrics if sung with provided melody. The last two dimensions require the annotators to sing the song.

Model	Song	*sense*	*style*	*listenability*	*intelligibility*
GagaST w/o constraints	Song1	3.4	3.0	3.2	3.4
	Song2	3.6	3.9	3.4	3.8
	Song3	3.7	3.6	3.4	3.5
	Song4	3.2	3.0	2.8	3.0
	Song5	3.7	3.6	3.4	3.8
	Average	**3.5** ±0.14	3.4 ±0.14	3.2 ±0.12	3.5 ±0.13
GagaST	Song1	3.5	3.1	3.3	3.5
	Song2	3.4	3.7	3.5	4.0
	Song3	3.2	3.6	3.3	3.6
	Song4	2.9	3.0	3.1	3.5
	Song5	3.4	3.6	3.2	3.9
	Average	3.3 ±0.15	3.4 ±0.15	**3.3** ± 0.12	**3.7** ±0.13

Table 3: Qualitative evaluation results for GagaST w/o constraints and GagaST.

5.4.3 Qualitative Evaluation Results

To examine whether the proposed constraints improve the singability and intelligibility, our qualitative evaluation compares GagaST with only length constraints to fully constrained GagaST (Table 3) with *syllable-to-syllable* assignment. While the constraints significantly improve the intelligibility and slightly improve the singability (listening experience), these constraints make it harder for the original meaning to come through. Overall, the annotators are satisfied with the translated

songs by the proposed baseline GagaST. All aspects receive an average score around 3.5 out of 5. These case studies and three translated songs by GagaST sung by an amateur singer are available on https://gagast.github.io/posts/gagast.

6 Related Work

Verse Generation and Translation. Generating verse text began through rule-based implementations (Milic, 1970) and developed through the next forty years. Manurung (1999) design a chart system that generate strings that match a given stress pattern. Gervás (2000) build a forward reasoning rule-based system. Manurung (2004) address poetry generation with stochastic search based on evolutionary algorithms. Oliveira (2012) create a template-based platform that allows user to define features and create templates. He et al. (2012) adopt statistical machine translation models for Chinese poetry generation. Yan et al. (2013) compose poetry based on generative summarization framework. Zhang and Lapata (2014), Wang et al. (2016), and Hopkins and Kiela (2017) adopt recurrent neural networks for poetry generation and incorporate rhythmic constraints. Ghazvininejad et al. (2016, 2017) represent rhythm and rhyme with finite-state machines. Poetry translation using these frameworks and statistical machine translation thus offers elegant solutions: Genzel et al. (2010) intersect the finite state representation of the meter and rhyme scheme with the synchronous context-free grammar of the translation model under the phrase-based machine translation framework. Ghazvininejad et al. (2018) apply the finite-state constraints to neural translation model. However, these representations of the rhythmic and lexical constraints are not flexible enough to encode the real-valued representation of a *song* as required for translation in tonal languages.

Constrained Text Generation. Most natural language generation tasks, including machine translation (Bahdanau et al., 2015; Vaswani et al., 2017; Hassan et al., 2018), dialogue system (Shang et al., 2015; Li et al., 2016; Wang et al., 2021) and abstractive summarization (Rush et al., 2015; Paulus et al., 2018), are free text generation. However, there is a need to generate text with constraints for special tasks (Lakew et al., 2019; Li et al., 2020; Zou et al., 2021). Hokamp and Liu (2017); Post and Vilar (2018); Hu et al. (2019) attempt to constrain the beam search with dictionary. In the training procedure, Li et al. (2020) add format embedding. Lakew et al. (2019) introduce length tag. Saboo and Baumann (2019) address length control via rescoring the results of beam search for machine translation under dubbing constraints.

Lyrics Generation. As one of the most important tasks in automatic songwriting, lyrics generation has received more attention recently. Sheng et al. (2021), Lee et al. (2019) and Chen and Lerch (2020) generate lyrics via pure data driven models without adding constraints based on expert knowledge. Oliveira et al. (2007) build a rule-based lyrics generation system to handle rhyme and rhythm with designed heuristics. Malmi et al. (2016) address rap lyrics generation via information-retrieval approach and propose a rhyme-density measure. Watanabe et al. (2018) add conditions in standard RNNLM with a featurized input melody for rhythmic alignment. Ma et al. (2021) develop a SeqGAN-based lyrics generator to address various properties, such as rhythmic alignment, theme and genre. Xue et al. (2021) use transformer-based model to generate rap lyrics with a reverse order, address rhymes with vowel embeddings and add extra beat tokens for rthymic alignment. We are the first paper that formally address the importance of aligning melody pitch with languages tones in lyrics generation for tonal languages. We introduce two vital qualities of songs, singability and intelligibility, and design three types of melody-lyric alignment scores to improve the two qualities.

7 Conclusion

This paper addresses automatic song translation (AST) for tonal languages and the unique challenge of aligning words' tones with melody. And we build the first English-Mandarin AST system – GagaST. Both objective and subjective evaluations demonstrate that GagaST successfully improves the singability and intelligibility of translated songs.

More constraints are left in the future work such as rhymes and style. We aim to build a systematic framework that address all constraints. With the help of newly developed singing voice synthesize tools such as X Studio,[9] we can perform human evaluation with actual singing voice with a larger scale to provide more reliable analysis. Moreover, our system can also be applied in lyrics and song generation applications without translation input.

[9] https://singer.xiaoice.com

Ethical Considerations

GagaST improves singability and intelligibility of the translated songs in Mandarin via constraining the decoding of a pretrained lyrics translation model. This methodology has limitations by imposing a direct trade-offs between the original objective and the constraints. In terms of negative impact or risks, the inaccurate translations may cause misunderstandings in applications like Musical Theatre.

This paper collects lyrics data that are publicly available and are parsed from the Web. We use these data for research purposes only. To prevent any abuse or piracy of these data, we chose the dataset license Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Acknowledgements

This work was initially supported by Alibaba DAMO Academy through Alibaba Innovative Research Program. We thank the anonymous reviewers for helpful feedback on previous versions of this work. We thank Carol, Xin Xu and IGP for the helpful discussions about music. Finally, we thank the singer/songwriter Ayanga for his professional discussion about the difficulties in musical translations, which was the genesis of this paper. Boyd-Graber and Guo's work is supported in part by a gift from Adobe.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

David Bellos. 2013. *Fictions of the Foreign: The Paradox of "Foreign-Soundingness"*, pages 31–43. Columbia University Press.

Yihao Chen and Alexander Lerch. 2020. Melody-conditioned lyrics generation with seqgans. In *2020 IEEE International Symposium on Multimedia (ISM)*.

Johan Franzon. 2008. Choices in song translation: Singability in print, subtitles and sung performance. *The Translator*, 14(2):373–399.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Pablo Gervás. 2000. WASP: Evaluation of different strategies for the automatic generation of spanish verse. In *Proceedings of the AISB-00 symposium on creative & cultural aspects of AI*.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48. Association for Computational Linguistics.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William D. Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *ArXiv*, abs/1803.05567.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating Chinese classical poems with statistical machine translation models. In *Association for the Advancement of Artificial Intelligence*.

Douglas R Hofstadter. 1997. *Le ton beau de Marot: In praise of the music of language*. Basic Books New York.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the Association for Computational Linguistics*, pages 1535–1546. Association for Computational Linguistics.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178. Association for Computational Linguistics.

Fangzhou Hu. 2017. Lexical tones in mandarin sung words: A phonetic and psycholiguistic investigation. Master's thesis, Shanghai International Studies University.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme.

2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 839–850. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*. Association for Computational Linguistics.

Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. iComposer: An automatic songwriting system for Chinese popular music. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 84–88. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. *ArXiv*, abs/1603.06155.

Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Rigid formats controlled text generation. In *Proceedings of the Association for Computational Linguistics*, pages 742–751. Association for Computational Linguistics.

Peter Low. 2003. Singable translations of songs. *Perspectives: Studies in Translatology*, 11(2):87–103.

Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. 2021. Ai-lyricist: Generating music and vocabulary constrained lyrics. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. 2016. Dopelearning: A computational approach to rap lyrics generation. In *Knowledge Discovery and Data Mining*.

Hisar Manurung. 2004. *An evolutionary algorithm approach to poetry generation*. Ph.D. thesis, University of Edinburgh.

Hisar Maruli Manurung. 1999. A chart generator for rhythm patterned text. In *Proceedings of the First International Workshop on Literature in Cognition and Computer*.

Louis T. Milic. 1970. The possible usefulness of poetry generation. In *Symposium on the Uses of Computers in Literary Research*.

Hugo Gonçalo Oliveira. 2012. Poetryme: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence*, 1:21.

Hugo R Gonçalo Oliveira, F Amilcar Cardoso, and Francisco C Pereira. 2007. Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th. International Joint Workshop on Computational Creativity*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the International Conference on Learning Representations*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1314–1324. Association for Computational Linguistics.

ITUT Rec. 1994. P. 85. a method for subjective performance assessment of the quality of speech voice output devices. *International Telecommunication Union, Geneva*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Ashutosh Saboo and Timo Baumann. 2019. Integration of dubbing constraints into machine translation. In *Proceedings of the Fourth Conference on Machine Translation*.

Murray Henry Schellenberg. 2013. *The realization of tone in singing in Cantonese and Mandarin*. Ph.D. thesis, University of British Columbia.

Marius Schneider. 1961. Tone and tune in West African music. *Ethnomusicology*, 5(3):204–215.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the Association for Computational Linguistics*, pages 1577–1586. Association for Computational Linguistics.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Association for the Advancement of Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

In *Proceedings of Advances in Neural Information Processing Systems*.

Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. 2016. Chinese song iambics generation with neural attention-based model. In *International Joint Conference on Artificial Intelligence*.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2021. Task-oriented dialogue system as natural language generation. *ArXiv*, abs/2108.13679.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 163–172. Association for Computational Linguistics.

Lian Hee Wee. 2007. Unraveling the relation between mandarin tones and musical melody. *Journal of Chinese Linguistics*, 35(1):128.

Yi Xu. 1997. Contextual tonal variations in mandarin. *Journal of phonetics*, 25(1):61–83.

Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. Deeprapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the Association for Computational Linguistics*.

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. 2013. I, poet: automatic Chinese poetry composition through a generative summarization framework under constrained optimization. In *International Joint Conference on Artificial Intelligence*.

Yang Yinliu, Sun Congyin, and Wu Junda. 1983. *Language and Music*. People's Music Publishing House.

Moira Yip. 2002. *Tone*. Cambridge University Press.

Weimin Yu. 2021. The study of the fourth tone in kunqu opera. *Journal of National Academy of Chinese Theatre Art*.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Yongping Zhuang. 1982. The relation of the sung words in peking opera.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. In *Knowledge Discovery and Data Mining*, page 2450–2460. Association for Computing Machinery.

Appendix A

A.1 Illustration of Tonal Alignment by Frequency

Translating songs into tonal languages faces a unique challenge,i.e., the tones of the translated lyrics should align with the music pitch for singability and intelligibility (Section 2.1). Figure 8 provides visual illustration of the main problem.

To help researchers who speak non-tonal languages understand better how the tones of lyrics in tonal languages should align with the music/sung voice, we record both sung and spoken voice of a piece of lyrics from one of the most popular songs in Mandarin, transform the sound into the frequency space, and compare the shape of the sound with that of the music in Figure 9. The original music of the chosen song is from an American song "Dreaming of Home and Mother", and was rewritten in Mandarin. Despite that this is not a translation task and do not have to convey the original meaning, we can see how the tonal contour of the lyrics in Mandarin align with that of music.

A.2 Acceptable Pitch Transition Directions Table

In Section 2.2, we explain that in practice, the relative relationship of the pitch of the tones of the successive syllables/characters that belongs to the same word affect the most to the singability and intelligibility. And we summarize the acceptable transition directions in Figure 5 under the assumption that only relative relationship of successive notes matters. It should be noted that we intend merely to provide a workable solution but not a perfect one. For example, the handle of the fourth notes of Mandarin is actually very tricky. It is a continuous fall with a large range (see Figure 2), therefore it doesn't represent one note. If it were to be represented by one note, it might represent the onset or offset part of the tone, and the falling trend is hinted by the pitch contour with proceeding and/or following note (Zhuang, 1982; Yu, 2021).

Appendix B

B.1 Training Details

We pretrain our transformer-based model with reconstruction objective and corrupt our input sequence with **text infilling** (Lewis et al., 2020). More detailed pretraining hyper-parameters can be found in Table 4.

Parameter	Value
encoder layer	12
decoder layer	12
max source position	512
max target position	512
layernorm embedding	True
criterion	label smoothed cross entropy
learning rate	3e-4
label smoothing	0.2
min lr	1e-9
lr scheduler	inverse sqrt
warmup updates	4000
warmup initial lr	1e-7
optimizer	adam
adam epsilon	1e-6
adam betas	(0.9, 0.98)
weight decay	0.01
dropout	0.1
attention dropout	0.1
text infilling	
mask rate	0.3
poisson lambda	3.5
replace length	1

Table 4: Pretraining hyper-parameters

Appendix C

C.1 Human Evaluation Instruction

In this paper, we conduct subjective evaluations by collecting annotations about the qualities of the translated songs from music school students (Section 5.4.2).

C.2 Music Sheets

As describe in Section 5.4.2 and shown in the instructions (Figure 10), we distributes music sheets of the translated songs to the annotators. All music sheets can be found on https://gagast.github.io/posts/human_eval.

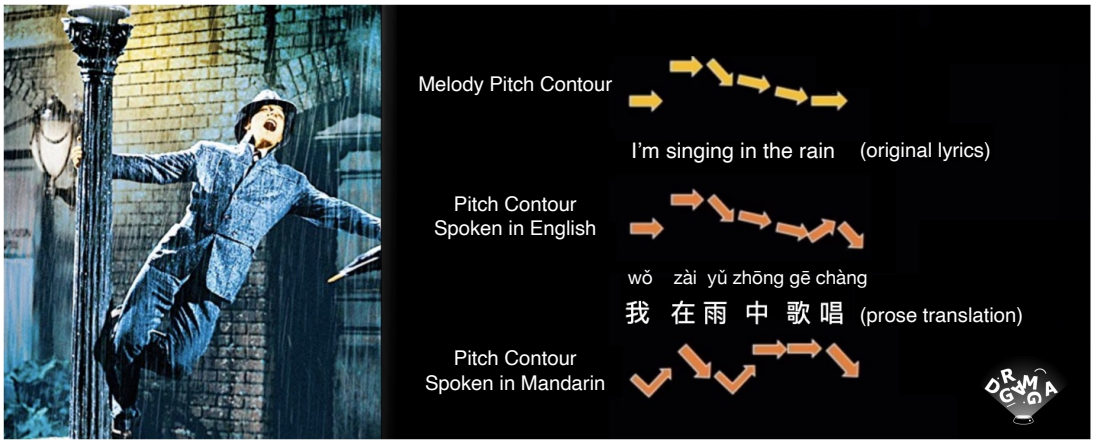

Figure 8: The pitch contour of the prose translation (bottom line, in Mandarin) of lyrics do not match that of the original music (upper line). The directions showed in figure is estimated by the base frequency of spoken sound by text-to-speech tools. Such mismatch in pitch contour makes the sung lyrics sound unnatural and hard to understand.

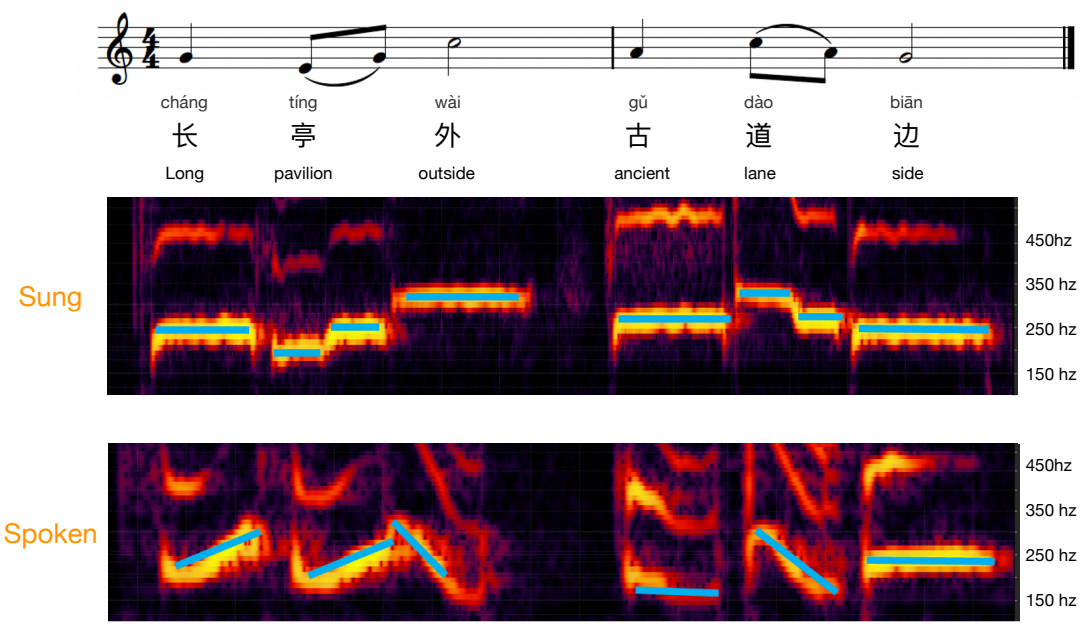

Figure 9: An example of a piece of a popular rewritten song in Mandarin "Farewell (sòng bié)". The original music is from an American song "Dreaming of Home and Mother". We record the sung and spoken voice and plot the actual base frequency of the sound. We can see how the tone shape and overall tonal contour aligns with the sung voice (by the music pitch).

Instruction

1) The sheet music for songs to be evaluated can be found in：
./[Song_Name]
— 1.pdf : sheet music for translation 1
— 2.pdf : sheet music for translation 2

Remark: We numerate the results from different translation models randomly. You *should not* assume that all "translation 1" are generated from the same model.

2) Criteria (1: bad 2: poor 3: fair 4: good 5: excellent)

a. ***sense***: How close the meaning of the translated line is to that the original line of lyric. The translation could be *paraphrase*.
Remark: You should look into the context to identify the actual meaning of each line.
[The lines with successive numeration are successive lines in lyrics]

b. ***style***: Whether the translated line looks like song-text style, as opposed to prose text.

c. ***listenability***: How well does it sound if the translated line is sung with given melody.
[See corresponding sheet music]

d. ***intelligibility***: Can you understand the sung words? (Would you misheard the lyrics when it is sung)
[See corresponding sheet music]

Remark：We accept errors. Try not give bad if minor error occurred.
For example, 1-bad would be "do not fit at all"; 5-excellent would be "fit almost perfectly";
3-fair would be "fit 50%-70%"

Figure 10: Instructions for human evaluation