# Automatic Song Translation for Tonal Languages

**Anonymous ACL submission**

## Abstract

This paper addresses automatic song translation (AST) for tonal languages and 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 semantics, singability and intelligibility—and develop objectives for these criteria. We develop a new benchmark for English–Mandarin song translation and develop an unsupervised AST system, the 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]

## 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 Chinese? These literal translations simply do not work: translation A) and C) do not match the number of notes and break the original rhythm; while the tone of translation B) does not match with the pitch flow of the original melody.

Song translation, unlike lyrics translation (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 must bear in mind the rhythms, note-values, phrasings, and stresses. 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[2] and *Les Misérables* (musical) has been performed in over a dozen languages on stage.

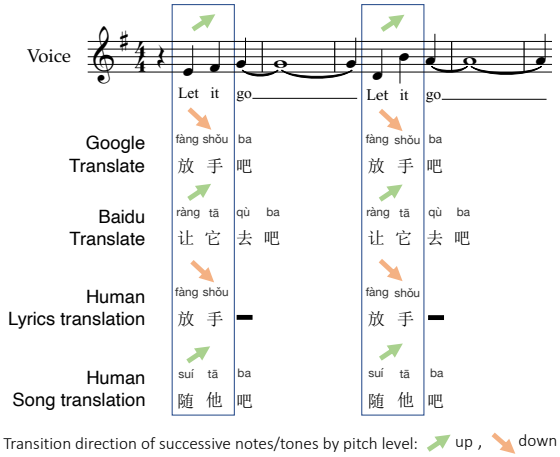



Figure 1: Example Mandarin translations for "Let it go" in *Frozen*. Of these, only the official human song translation considers whether a singer could sing the song: it fits the length of the notes and matches the tones with the pitch of notes.

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

[1] We illustrate the task and examples of translated songs by GagaST on https://gagast.github.io/posts/gagast.

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

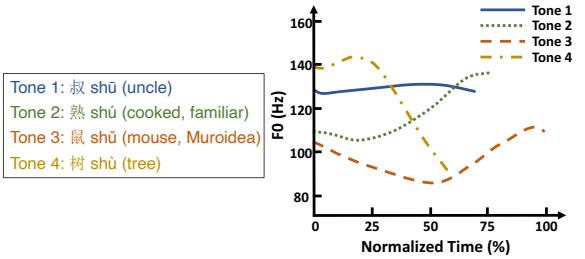



Figure 2: In total languages like Mandarin, the pitch changes the meaning of the words (left). Each of the four tones in Mandarin (right) has a different pitch profile. Figure from Xu (1997).

| | Original lyrics | | | | Misheard lyrics | | | |
|---|---|---|---|---|---|---|---|---|
| Pitch level | 66 | 68 | 66 | 65 | 66 | 68 | 66 | 65 |
| Pronunciation | sì | zài | yǎn | qián | sǐ | zài | yǎn | qián |
| Lyrics | 似 | 在 | 眼 | 前 | 死 | 在 | 眼 | 前 |
| English translation | as | in | front of eyes | | die | in | front of eyes | |
| Pitch alignment score | 0.5 | | | | 0.75 | | | |

Figure 3: A misheard example in Mandarin song caused by a mismatch between music pitch flow and the lyric's tones. The heard word is "sǐ zài" instead of "sì zài", because notes are going up and "sì zài" is going down by the sandhi of Mandarin tone.

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 data (Section 4.1) and adds constrats that favor translations that are the appropriate length and whose tones match the underlying music (Section 4.3). Naturally, such constraints result in a trade-off between semantic meaning and singability/intelligibility. Section 5.4 discusses this trade-off between alignment scores and 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 judgement (Section 5.5).

## 2 Background: Prose, Poetry, and Song Translation

The form of written or spoken language has two divisions: prose, which has a natural flow of speech and grammatical structure; and verse, which is typically rhythmic and has special line breaks, such as traditional poetry and song lyrics.

The vast majority of machine translation research has been focused on prose translation and has made huge progress; while 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 created eighty-nine translation of a single poem to capture various aspects 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, trade-offs are unavoidable. In his "pentathlon principle", Low (2003) argues for prioritizing *singability*: can a performer put the translated lyrics to music. Tonal language (e.g., Mandarin, Zulu and Vietnamese) dramatically increases the complexity of singability, and raises a new issue of intelligibility.

### 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:

[3] Additional misheard examples on demo page https://gagast.github.io/posts/gagast/#misunderstanding_examples

- **The pitch level and shape of tones.** Four Mandarin tones are used since the 19th century. 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). Their pitch level and shape are shown in Figure 2, right.
- **The sandhi of tones.** Some combinations of tones have difficult articulatory patterns, so words that might normally have one tone might take another. For example "nǐ" and "hǎo" are typically both third tone, but when they are together it is pronounced as "ní hǎo", with the first syllable changing to a *second* tone. These changes are called sandhi (Xu, 1997; Hu, 2017).

Mandarin tones interact with singing 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 characters should be consistent with the musical notes they're matched with; for example, in "Love Island" (Figure 4), "shàng" in the blue block has the "falling" shape and the group of notes that it assigned with also goes falling. Second, and a global level, the *pitch contour* of music constrains the tones in a successive sequence of syllables and the sandhi. In practice, we only consider the relative change in pitch of the two successive characters that belongs to the same word (refer to Figure 5).

## 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 major criteria that singable song translation needs to fulfil:

- **Preserve semantics.** The translated lyrics should be faithful to the original source lyrics.
- **Singability.** Low (2003) defines singability as the phonetic suitability of the translated lyrics with music. The translated song needs to be sung without too much difficulty; difficult consonant clusters, cramming too many sylables 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 stage performance, since the audience suppose to understand the content instantly to follow the plot. For pop song covering, not understanding all contents could be acceptable for some audience; however, hilarious misheard lyrics will hurt the experiences (Figure 3).

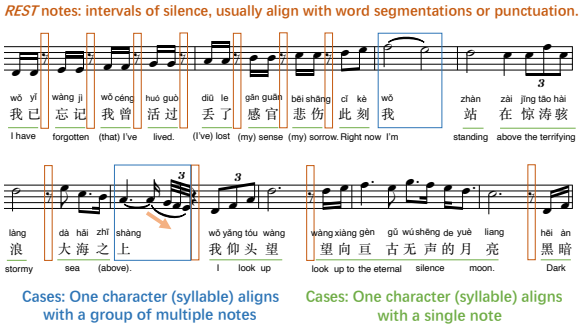



Figure 4: The alignments of syllables in Mandarin to notes in the song "Love Island". Orange: *REST* notes; Blue: cases where one syllable is assigned to a group of multiple notes (need consider *tone shape* alignment, e.g., the down arrow matches with falling tone of "ràng"); Green: cases where one syllable is assigned with one note.

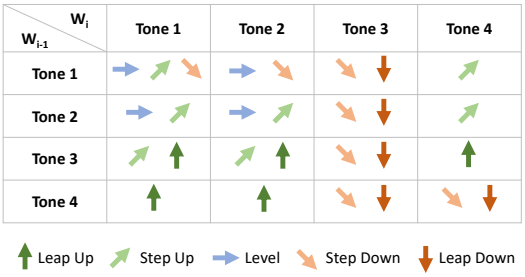



Figure 5: Acceptable notes transitions in music for two successive Mandarin characters $(w_{i-1}, w_i)$.

### 3.2 Task Definition

We define the AST task as follows: given a pair of melody $M$ and source lyrics $X$ (as shown in Table 1), generate text $Y$ in the target language. $X = [x_1, ..., x_L]$ are $L$ syllables of the original source lyrics. The melody $M$ has three sequences:

| 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 piece of song "Seasons of love" from the musical *Rent*. We convert the notes into a normalized numerical pitch level for actual computation.

1. The pitch profile $P = [\mathbf{p}_1, ..., \mathbf{p}_L]$, where $\mathbf{p}_i = [p_i^0, ...]$ are the pitch values of the $i^{\text{th}}$ notes group assigned to syllable $x_i$, and each syllable/character could be assigned to one or multiple successive notes (Figure 4);
2. The durations $D = [\mathbf{d}_1, ..., \mathbf{d}_L]$, where $\mathbf{d}_i = [d_i^0, ...]$ is the real-valued duration of each note in the $i$-th group;
3. $R = [r_1, ..., r_L]$, where $r_i$ 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 types of critical lyric-melody alignments for English-Mandarin AST (c.f., Section 2.1 and 2.2).

#### 3.3.1 Length Alignment

The number of syllables $L_y$ in translated lyrics $Y$ needs to fit the number of musical phrases in the melody $M$, so that it can be sung with the provided music. It is unnecessary to keep the grouping of the original notes in $M$ for the translated song.

#### 3.3.2 Pitch Alignment

For tonal languages, the pitches are required to match the translated songs. There are two types of pitch alignments:

**Tone shape alignment.** It is at single-syllable level. We only consider this alignment for the syllable that assigns to more than one notes. The shape of the tone is predefined in a tonal language (Wee, 2007), i.e., Mandarin tone shape (Xu, 1997) can be viewed in Figure 2. The shape of notes is the pitch contour of corresponding notes group $\mathbf{p}_i$. For computing tone shape alignment score in Mandarin and each group of $\mathbf{p}_i$ that assigns to syllable $x_i$, we estimate its shape by interpolation of the second order on $\mathbf{p}_i$, and classify it into one of the five categories: level, rising, falling, rising-falling, falling-rising. For example, if $p^i_{max} - p^i_{min} > 1.0$ and the estimated curve is convex with axis in the middle of $\mathbf{p}_i$, we fit it into category "falling-rising". Then we compare the shape with that of the syllable $y_i$, and compute the local tone shape match score $S^i_{ns}$:

$$S^i_{ns} = \begin{cases} 1.0 & \text{if the shape matches,} \\ \epsilon & \text{if not match,} \end{cases} \quad (1)$$

where $\epsilon$ is the probability to accept error; "level" can match with any tone, "rising" matches with tone2 (yú), "falling" matches with tone4 (huài), "falling-rising" matches with tone3 (wǒ) while "rising-falling" matches none.

**Pitch contour alignment.** It compares the transitions between tones $(t_{i-1}, t_i)$ of successive syllables $(y_{i-1}, y_i)$ that belong the same word and the pitch contour of corresponding successive 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*, *leap up*, *step down* and *leap down*. For Mandarin, according to Yinliu et al. (1983), we summarize the acceptable notes' transitions of two successive characters as illustrated in Figure 5. Similarly, for each pair of syllables $(y_{i-1}, y_i)$, we compute the local pitch contour $S^i_{pc}$ as follow:

$$S^i_{pc} = \begin{cases} 1.0 & \text{if the contour matches,} \\ \epsilon & \text{if not match,} \end{cases} \quad (2)$$

where $\epsilon$ is the probability to accept error.

#### 3.3.3 Rhythmic Alignment with Word Segmentation in Mandarin

A *REST* note represents the interval of silence. For Mandarin, a word should not be broken up by a *REST*, and sometimes *REST*s indicate the end of a phrase and correlate with the punctuation ([punc]), see Figure 4 for examples. Therefore, when there is a *REST* note before $y_i$ (after $y_{i-1}$), i.e., $r_i > 0.0$, we reward the [punc] and word segmentation between $y_{i-1}$ and $y_i$:

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

$P_{\text{seg}}$ is the probability that $(y_i, y_{i-1})$ are segmented into different words (the higher the probability, the better it is to have a pause between them), and $\epsilon$ is a parameter that represents our tolerance of having a rest within a word.

[4]We considers only the first note $p_i^0$ in group $\mathbf{p}_i$ if has more than one notes in each group

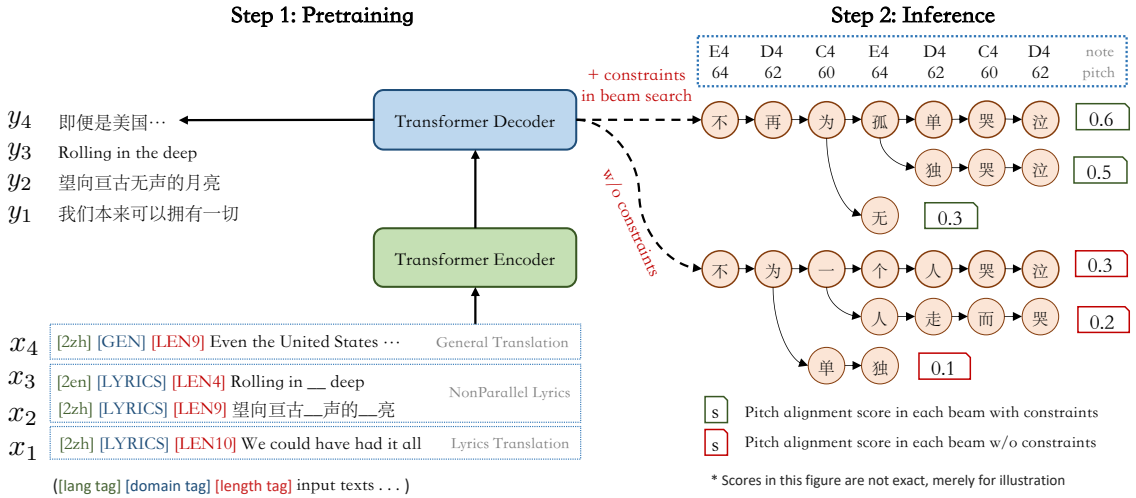



Figure 6: Overview of GagaST for English–Mandarin song translation. We first pre-train a lyrics translation model with mixture domain data (left); and then add alignment constraints in decoding scoring function during inference (right), we use unconstrained version as our baseline in the experiment.

## 4 GagaST

To build an AST system for English-Mandarin song translation, ideally, we can learn all alignments by data-driven models with large amount of parallel data, i.e., the aligned triples ($M$, $X$, $Y$). However, let alone triples, we do not have sufficient accurate parallel data for Mandarin[5]. In this case, we leverage cross-domain pre-training and propose an unsupervised AST system baseline, Guided AliGnment for Automatic Song Translation (GagaST).

For the pre-training, we collect a large amount of non-parallel lyrics data in both English and Mandarin, as well as a small set of lyrics translation (subtitling) data[6]; details about training dataset are in 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 general domain, the collected monolingual lyrics data and a small set of lyrics translation data. We adopt mix-domain training to optimize a translation model that fits into the lyrics domain. We append domain tags (Figure 6) before each input entry to control the model to produce translations merely in lyrics domain during song translation. For monolingual lyrics data, we adopt BART pre-training strategy (Lewis et al., 2020).

[5]The only parallel dataset in Mandarin parsed from web contains lots errors in notes and mismatches between syllables and notes; we need accurate alignments for intelligibility

[6]The translations are in plain text, not in lyrics style

### 4.2 Length Control

To meet the length alignments, we pre-define the syllable-notes assignments with two strategies[7]: 1) *one-to-one*, i.e., for each note, we produce one syllable; 2) *one-to-many*, we use the original notes grouping in the input melody, and assigns 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.

### 4.3 Music Guided Alignment Constraints

There is no available parallel data to learn the lyric-melody alignments with data-driven models, thus we leverage rules and metrics for each type of alignment (Section 3.3). Then we follow the idea in lexical constrained MT (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019) to impose constraints in the decoding phase. More specifically, since all constraints that we design are either unigram (tone shape, *REST*) or bi-gram (pitch contour, *REST*), we directly apply the lyric-melody alignment constraints at each step of beam search as rewards and penalties in the scoring function :

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

[7]Dynamic mapping between the note sequence and the syllables to be generated increase the search space exponentially.

where $S_{pc}$, $S_{ns}$, and $S_R$ refer to the alignment scores for pitch contour, note shape and the rhythm, $\lambda_{pc}$, $\lambda_{ns}$, and $\lambda_R$ represent the corresponding hyper-parameters that controls the influence of each constraints.

# 5 Experiments

## 5.1 Training Datasets and Model Configuration

**WMT translation data.** We use the news commentary and back-translated news datatsets from WMT14, which consisting of about 29.6 million en2zh sentence pairs.

**Monolingual lyrics data.** We collect monolingual lyrics in both Mandarin and English from the web, which contains about 12.4 million lines of lyrics for Mandarin and 109.5 million for English after removing the duplication.

**Lyrics translation data.** We crawl a small set of lyrics translation data from the web[8], which contains 140 thousands pairs of English-to-Mandarin lines. These translations are not singable.

We preprocess all data with fastBPE (Sennrich et al., 2016) and a code size of 50,000. We choose standard encoder-decoder Transformer (Vaswani et al., 2017) model with an architecture of 768 hidden units, 12 heads, GELU activation, a dropout rate of 0.1, 512 max input length, 12 layers in encoder, 12 layers in decoder.

## 5.2 Evaluation Dataset

For evaluation, we need aligned triples (melody $M$, source lyrics $X$, target reference lyrics $Y$), where $M$ and $X$ are syllable-to-notes aligned; and the reference $Y$ should be singable and intelligible. Without copyright and accessibility to the singable translated songs, we chose 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 do not necessarily resemble song-text style and are not singable, we use them merely to provide a coarse estimation for semantic changes. Twenty songs are used as the valid set (464 lines) and thirty songs as the test set (713 lines).

## 5.3 Evaluation Metrics

To evaluate the ability of given AST system in preserving semantics, singability and intelligibility, we design both objective and subjective evaluations.

[8]https://lyricstranslate.com/

### 5.3.1 Objective Evaluation

For **semantics**, we follow the common practice and calculate the BLEU scores (Papineni et al., 2002) between the translated lyrics and corresponding reference. For singability and intelligibility, we use the following metrics:

For **length alignment**, we computes: 1) $N_l$, the number of samples that has length longer than the predefined length $L_i$; 2) $N_s$, that are shorter than $L_i$. And for each case, we show the average error ratio of $\{\Delta l_i/L_i\}_1^{N_{[\cdot]}}$.

For the three scores for **pitch** and **rhythm alignment**, 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,

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

### 5.3.2 Subjective Evaluation

To demonstrate whether the proposed metrics align with actual human experiences and examine the quality of the translated songs by GagaST, we conduct human evaluations. We randomly select five songs from the test set and show the music sheets[9] of the first ten sentences of each translated song by GagaST to five annotators from Music School.

Following mean opinion score (MOS) (Rec, 1994) in speech synthesize task, we use five-point scales (1 for bad and 5 for excellent) in four aspects: 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 latter two qualities require the annotators to sing the song by themselves.

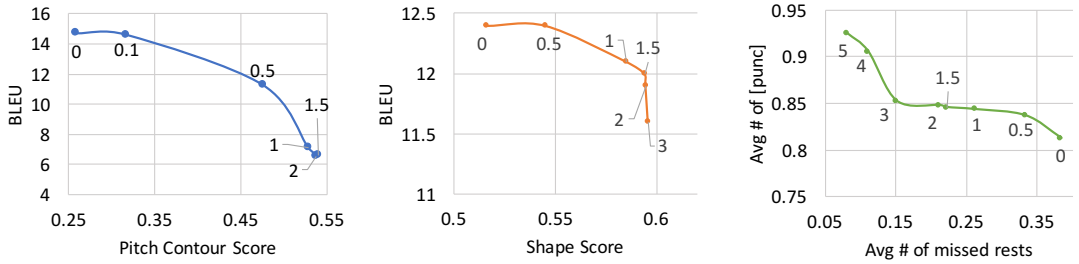
## 5.4 Hyper-parameters and Trade-offs

The GagaST adds constraints in the decoding scoring functions to enforce lyric-music alignments. There are trade-offs between semantics and other alignments. We analyze the increasing curves of pitch alignment scores against BLEU on valid set, and choose the hyper-parameters where the alignment scores increase fast while the BLEU decrease slow. The *REST* constraint does not affect the

[9]Without singing voice synthesize tools, following Sheng et al. (2021), we show the annotators the music sheets without singing

| Syllable-notes Assignment | Model | Pitch contour ↑ | Pitch shape ↑ | Rhythm avg # of missed rests ↓ | Length longer ↓ | Length shorter ↓ | Semantics BLEU ↑ |
|---|---|---|---|---|---|---|---|
| one-to-one | 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 contour | 0.51 | - | 0.45 | 26 (0.21) | 0 | 16.8 |
| | –only rest | 0.28 | - | 0.31 | 11 (0.09) | 0 | 23.8 |
| one-to-many | 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 contour | 0.51 | 0.50 | 0.42 | 7 (0.12) | 0 | 15.8 |
| | –only shape | 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: Objective results on test set of GagaST with different constraints under one-to-one and one-to-many assignments. 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.




(a) Different contour constraints. (b) Different shape constraints. (c) Different rest constraints.

Figure 7: Trade-off between semantics and lyric-music alignments; all curves are drawn for the valid set.

BLEU (Table 2) but the number of punctuation. We should prevent a large increase in the number of punctuation while reducing the mismatches between the *REST* and semantic segmentation. Based on Figure 7, we chose: $\lambda_{pc} = 0.5$; $\lambda_{ns} = 1.0$; $\lambda_R = 1.5$.

## 5.5 Evaluation Results

### 5.5.1 Objective Evaluation Results

Table 2 compares the performance of GagaST with different constraints. As described in Section 4, we pre-define the note(s) groups and use two syllable-notes assignments: *one-to-one* and *one-to-many*. From results in Table 2, we can see,

- The proposed length tag "[LEN$i]" helps to produce lyrics that fit into the predefined note(s) groups. In all cases, less than 30 out of 713 lines produces a longer sentence with ratio less than 0.22; and no short cases.
- We adopt the GagaST w/o constraints except for length tags as our baseline. Compared to which, GagaST with full constraints is able to increase both pitch and rhythm alignments significantly with a fairly slow drop in BLEU[10]. It almost doubles the pitch contour alignment score, which affect the intelligibility the most.
- Each constraint applied in decoding process is able to increase the corresponding alignment performance.

### 5.5.2 Subjective Evaluation Results

To examine whether the proposed constraints are able to improve the singability and intelligibility, and to evaluate the quality of translated songs by GagaST, we conduct subjective evaluation and compare the GagaST w/o constraints to fully constrained GagaST in Table 3. All songs for subjective evaluation are generated with *one-to-many* assignment. And we compute the confidence intervals for all aspects. Results show that,

- The proposed constraints is able to significantly improve the intelligibility for audience.
- The proposed constraints is able to improve the listening experiences for human with minor significance. The listenability for audience reflects the singability for performer.
- Add constraints do cause a trade-off between the semantics (sense) and other qualities.

[10]For references, we found three officially translated Disney songs in Mandarin and computes the BLEU among the human translated singable lyrics with the lyrics translation from our dataset, the average BLEU is only 12.3.

| 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: Subjective evaluation results for GagaST w/o constraints and GagaST.

- Overall, the annotators are satisfy with the translated songs by the proposed baseline GagaST. All aspects receive an average score around 3.5 out of 5.

The subjective evaluation demonstrates that the proposed alignments, constraints and the acceptable notes transitions for Mandarin (Figure 5) are reasonable. They are able to improve the singability and intelligibility. Although there's a trade-off, due to the lack of singing voice synthesize tools, the subjective evaluation is actually in favour of the sense/style evaluation compared to listenability/intelligibility. We add case studies and post three translated songs by GagaST sung by an amateur singer on https://gagast.github.io/posts/gagast.

# 6 Related Work

## 6.1 Verse Generation and Translation

Generating verse text began through rule-based implementations (Milic, 1970) and developed through the next forty years (Gervás, 2000; Levy, 2001; Manurung, 2004; Oliveira, 2012; He et al., 2012; Yan et al., 2013; Zhang and Lapata, 2014; Wang et al., 2016; Ghazvininejad et al., 2016, 2017; Hopkins and Kiela, 2017), incorporating formalisms such as grammars and finite-state machines as reviewed by Gonçalo Oliveira (2017). Poetry translation using these frameworks and statistical machine translation thus offers elegant solutions: Genzel et al. (2010) use phrase-based machine translation technique; while they simply intersect the finite state representation of the meter and rhyme scheme with the synchronous context-free grammar of the translation model. 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.

## 6.2 Constrained Text Generation

Most natural language generation tasks, including machine translation (Bahdanau et al., 2014; Vaswani et al., 2017; Hassan et al., 2018), dialogue system (Shang et al., 2015; Li et al., 2016) and abstractive summarization (Rush et al., 2015; Paulus et al., 2018), are free text generation. However, there is a need to generate text with some constraints for some 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) attempted to constrain the beam search with dictionary. In the training procedure, Li et al. (2020) added format embedding. Lakew et al. (2019) introduced length tag.

## 6.3 Lyrics Generation

Automatic song translation is a challenging task that involves two fields: machine translation and lyrics generation. As one of the most important tasks in automatic songwriting, lyrics generation has received more attention recently (Malmi et al., 2016; Watanabe et al., 2018; Bao et al., 2019; Lu et al., 2019; Lee et al., 2019; Sheng et al., 2021). Malmi et al. (2016) generated lyrics without melody information; Lee et al. (2019); Bao et al. (2019) attempted to deal with the melody-to-lyrics generation with sequence-to-sequence model. Sheng et al. (2021) use pre-training for melody-to-lyrics generation, but does not take knowledge in the music domain into account.

# 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.

In the future, we would like to build human-machine collaborated song translation system. Song translation is a hard task that requires rich music background knowledge including complex rules that most human translators lack; while competent human translator for prose translations can help to provide much more diverse and faithful translations. One can leverage the diversity of human translations to enrich the searching space and the encoded complex rules by AI systems to ensure singability and intelligibility.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hangbo Bao, Shaohan Huang, Furu Wei, Lei Cui, Yu Wu, Chuanqi Tan, Songhao Piao, and Ming Zhou. 2019. Neural melody composition from lyrics. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 499–511. Springer.

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 Empirical Methods in Natural Language Processing*.

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 *Conference of the North American Chapter of the Association for Computational Linguistics*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of Empirical Methods in Natural Language Processing*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of the Association for Computational Linguistics*.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. 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*.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 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*.

Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *16th International Workshop on Spoken Language Translation*.

Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. 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 (Demonstrations)*, pages 84–88.

Robert P Levy. 2001. A computational model of poetic creativity with neural network as measure of adaptive fitness. In *Proceedings of the ICCBR-01 Workshop on Creative Systems*.

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. *ArXiv*, abs/1910.13461.

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.

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

Xu Lu, Jie Wang, Bojin Zhuang, Shaojun Wang, and Jing Xiao. 2019. A syllable-structured, contextually-based conditionally generation of chinese lyrics. In *Pacific Rim International Conference on Artificial Intelligence*, pages 257–265. Springer.

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. 2016. Dopelearning: A computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–204.

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

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. *ArXiv*, abs/1705.04304.

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*.

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 *EMNLP*.

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*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*.

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 *Proceedings of the Association for Computational Linguistics*.

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*.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. 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, Volume 1 (Long Papers)*, pages 163–172.

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.

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.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. *arXiv preprint arXiv:2103.10685*.