Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI, the Arts and Creativity

Intoner: For Chinese Poetry Intoning Synthesis

Heda Zuo[1], **Liyao Sun**[2], **Zeyu Lai**[1], **Weitao You**[1*], **Pei Chen**[1] and **Lingyun Sun**[1]
[1]Zhejiang University
[2]Fudan University
{zuoheda, jerrylai, weitao_you, chenpei, sunly}@zju.edu.cn, sunly23@m.fudan.edu.cn

Abstract

Chinese Poetry Intoning, with improvised melodies devoid of fixed musical scores, is crucial for emotional expression and prosodic rendition. However, this cultural heritage faces challenges in propagation due to scant audio records and a scarcity of domain experts. Existing text-to-speech models lack the ability to generate melodious audio, while singing-voice-synthesis models rely on predetermined musical scores, which are all unsuitable for intoning synthesis. Hence, we introduce Chinese Poetry Intoning Synthesis (PIS) as a novel task to reproduce intoning audio and preserve this age-old cultural art. Corresponding to this task, we summarize three-level principles from poetry metrical patterns and construct a diffusion PIS model Intoner based on them. We also collect a multi-style Chinese poetry intoning dataset of text-audio pairs accompanied by feature annotations. Experimental results show that our model effectively learns diverse intoning styles and contents which can synthesize more melodious and vibrant intoning audio. To the best of our knowledge, we are the first to work on poetry intoning synthesis task.

1 Introduction

Intoning (*yinsong* in Chinese) is a prevalent method for studying Chinese ancient poetry. It diverges from reading through its incorporation of distinct melodies, which enhance emotional expression and augment the prosodic allure of the poetry. Unlike singing either, intoning permits expressive latitude without adherence to a fixed musical score. Figure 1 illustrates an example of famous poetry intoning.

While intoning lacks a predefined musical score, it necessitates adherence to specific principles based on poetry metrical patterns. We summarize these principles into three levels according to the insights of professional scholars.
1) character-level: Ancient Mandarin comprises four distinct tones, each offering unique avenues for conveying emotions. It is imperative to modulate the pitch and duration of each word in harmony with the inherent characteristics of its tone.

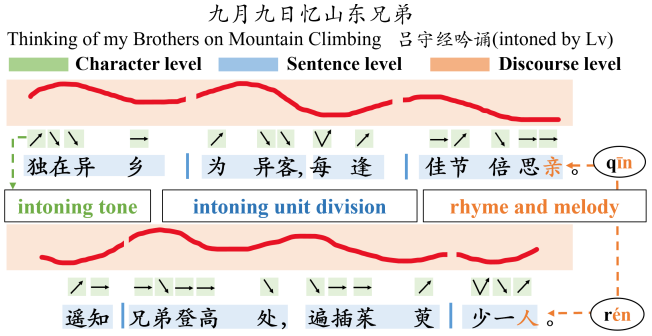


Figure 1: A melody curve derived from Lv's intoning. Characters in orange are rhyming characters (discourse-level), the blue bars stand for intoning unit division (sentence-level), while arrows in green stand for different tones in Mandarin (character-level).

2) sentence-level: A sentence undergoes segmentation into multiple intoning units. These intoning units maintain internal coherence, with transitions between adjacent moves typically signaled by a caesura (pause and breath) in intoning.
3) Discourse-level: Every instance of intoning is accompanied by a graceful melody, often drawn from experts. Furthermore, it is essential to accentuate the rhyming characters to preserve the cohesiveness and integrity of the poem.

Although poetry intoning is a critical oral art, it faces challenges in its cultural dissemination due to the limited availability of intoning audio. Prior investigations into Chinese poetry intoning have predominantly concentrated on fundamental intoning principles [Zhu, 2013; Li, 2017; Xu, 2014], distinctions among various intoning styles, and the educational applications of intoning [Cong, 2021]. While a limited number of studies have ventured into audio synthesis related to poetry, their scope has been confined to regular reading rather than intoning [Zhu and Zhu, 2008].

The advancement in voice synthesis technology presents new prospects for promoting the traditional art of intoning. Utilizing generative models trained on a constrained set of expert intoning data, the synthesis of intoning audio for previously unencountered poems is now achievable. Contemporary voice synthesis methodologies fall into two categories: text-to-speech synthesis [Ju *et al.*, 2024; Ren *et al.*, 2019] and singing voice synthesis [Hwang *et al.*, 2025; Liu *et al.*, 2022]. The former overlooks melody, whereas the

*Corresponding author.

latter requires a specified musical score. Moreover, in the absence of guiding knowledge, audio generated by these models fails to align with the distinct attributes of poetry intoning.

In this paper, we try to solve the Poetry Intoning Synthesis (PIS) task. First, we summarize three-level principles as guiding knowledge, along with evaluation metrics for the task. Then, we train a PIS model named Intoner with a Character-level Feature Extraction(CFE) module based on a multi-decoder transformer [Vaswani *et al.*, 2017] and a PIS module utilizing autoregressive and diffusion strategy. We hierarchically extract features in CFE module and embed them to guide synthesis in PIS module, making it possible to synthesize melodic intoning without any musical inputs. Due to the lack of off-the-shelf Chinese poetry intoning dataset, we also collect a large-scale dataset with text-audio pairs accompanied by annotations of poetic characteristics.

Our main contributions can be summarized as follows: 1) We are the first to study Chinese poetry intoning synthesis and propose this as a novel task, which helps preserve and propagate this form of art; 2) We summarize three-level principles and metrics for poetry intoning synthesis; 3) We propose a model Intoner which can synthesize high-quality poetry intoning audio that meets all the three-level principles; 4) We construct a large-scale dataset of Chinese poetry intoning with text-audio pairs and labels of character-level features.

2 Related Work

Chinese Poetry Intoning has been highly valued in recent decades [Yang, 2015; Yang, 2021], with a focus on preserving this traditional art. To this end, experts such as [Xu, 2014] have provided clear definitions of poetry intoning, while [Zhu, 2013; Li, 2017] have proposed basic rules for poetry intoning at different levels. In recent times, a number of studies have investigated different styles and dialects of poetry intoning as well as intoning in education [Cong, 2021; Du, 2023]. However, thus far, only [Zhu and Zhu, 2008] has proposed a synthesis method for poetry reading (but not intoning), with other studies neglecting the use of digital techniques to aid in the preservation of this cultural practice.

Text To Speech (TTS) aiming to synthesize human-like voice with text inputs. TTS task is usually seperated by mel-spectrum synthesis and audio synthesis. For general models for mel-spectrum synthesis with texts as input, tacotron series [Wang *et al.*, 2017; Elias *et al.*, 2021] and Transformer-TTS [Li *et al.*, 2019] are end-to-end models for single-speaker TTS synthesis; FastSpeech series [Ren *et al.*, 2019; Ren *et al.*, 2020], VITS series [Kim *et al.*, 2021; Kong *et al.*, 2023] can support multi-speaker and acoustic feature control; [Ju *et al.*, 2024; Liu *et al.*, 2022; Deng *et al.*, 2025] are proposed based on diffusion models to solve over-smoothing and unstable training problems. For models for audio synthesis with mel-spectrums as input, different vocoders are proposed in succession, separately based on convolutional neural network [Oord *et al.*, 2016], flow-based structure [Prenger *et al.*, 2019], GAN [Kumar *et al.*, 2019; Kong *et al.*, 2020]. Nowadays, with the development of large multi-modal models, TTS models can generate more complex speech audio [Huang *et al.*, 2024; Du *et al.*, 2024].

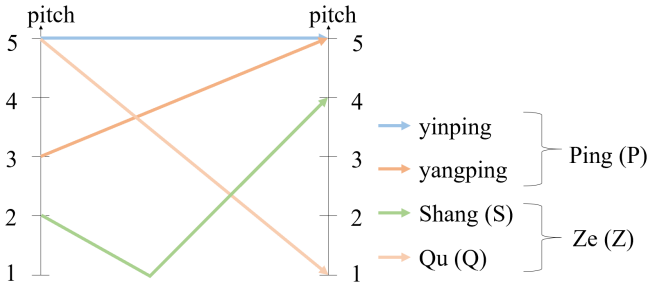


Figure 2: The pitch shape of modern Mandarin. For the lost tone 'ru (R)' in ancient Mandarin, we usually pronounce it shortly by ending with a glottal stop.

Singing Voice Synthesis (SVS) is used to be researched by concating or statistical methods. At present, deep neural networks are widely used in this area [Nishimura *et al.*, 2016]. [Hwang *et al.*, 2025; Liu *et al.*, 2022] employ diffusion strategy for high-quality SVS while [Yu *et al.*, 2024] uses self-supervised learning techniques. More studies focus on the application of SVS models, in which [Dai *et al.*, 2024; Zhang *et al.*, 2024] explore style transformation and controllability while [Gu *et al.*, 2021; Lu *et al.*, 2020] focus on adapting SVS to different languages. Nowadays, models like [Wang *et al.*, 2025] adopt large language models to control the synthesis of singing voice.

However, these models cannot support PIS directly, since they cannot synthesize melodious intoning without predefined musical scores, which do not match the characteristic of poetry intoning.

3 Principles

We summarize three-level principles from professional scholars, which guide the design of evaluation metrics and the structure of Intoner.

3.1 Character-level

In poetry intoning, it is necessary to distinguish the correct tone and pronunciation of each character, as they significantly influence the melody of the intoning. Instead of simply using the modern pronunciation, sometimes we need to imitate ancient pronunciation in intoning to keep consistent with the poetry metrical patterns.

Tone We accord particular significance to tone input as a pivotal guiding element. As a tonal language, different tones in Mandarin have distinct characteristics as illustrated in Figure 2. There are four tones in ancient Mandarin. '**ping (P)**' sounds sad and peaceful, '**shang (S)**' sounds sharp and firm, '**qu (Q)**' sounds clear and remote, while '**ru (R)**' sounds direct and short. 'S', 'Q' and 'R' make up 'ze (Z, meaning oblique)' which is contrary to 'ping (P, meaning flat)'. Generally, 'Z' is characterized by brevity and a heavier tone, whereas 'P' exhibits a prolonged duration and a soft sound. Notably, 'P' further subdivides into 'yinping' and 'yangping,' while 'R' has become a lost tone in modern Mandarin.

Pronunciation In Chinese poetry intoning, discerning the correct pronunciation poses three distinct challenges, as il-

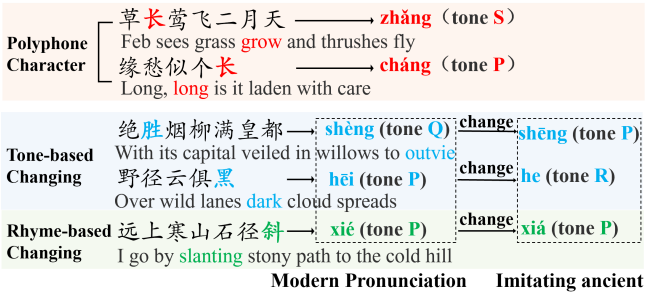


Figure 3: Three kinds of distinguishing pronunciation, including polyphone character, tone-based and rhyme-based changing.

lustrated in Figure 3: 1) polyphone characters: a single character could be pronounced differently when it stands for different meanings; 2) tone-based changing: some characters at rhythm points deviate from the metrical patterns dictated by modern Mandarin. Consequently, their tones must be adjusted to align with the metrical patterns; 3) rhyme-based changing: some rhyming characters exhibit a lack of rhyme when pronounced in modern Mandarin. To address this, adjustments to the finals of these characters become necessary, ensuring a harmonious rhyme in the context of intoning.

3.2 Sentence-level

In our approach, we explicitly predict **caesura**s (represented by stop tokens) and incorporate them into the phoneme sequence. Each sentence consists of multiple characters, which should not be intoned in one breath, therefore, it becomes imperative to divide the sentence into several intoning units with caesuras between them. This segmentation enhances the natural flow and rhythmic cadence of intoning. In poetry intoning, the division of intoning units typically aligns with either metrical patterns or semantic nuances of the content.

3.3 Discourse-level

To ensure the coherence and reflect the interconnectedness of the intoning, we focus primarily on rhymes and the order of sentences at the discourse level.

Rhyme The syllabic structure of Chinese characters consists of initials and finals, with rhyming characters necessitating identical finals. In Chinese poetry, the incorporation of rhyming characters, typically positioned at the end of a sentence, ensures a harmonious rhyme scheme. In intoning, it is crucial to appropriately extend and accentuate rhyming characters, thereby fortifying the overall integrity of the poetic expression. Typically, a poem features one rhyme. If a poem undergoes a rhyme change, the distinctions between different rhymes should be discernible in intoning.

Order Typically, there are two sentences in one line of a Chinese poem. The lower sentence has the opposite metrical pattern to the upper but the same as the upper one of the next line. Consequently, varying orders of sentences are intoned with distinct melodies, reflecting the differing metrical patterns between them. However, if two sentences share the same order and metrical patterns but belong to different poems, their melodies exhibit a notable similarity. Hence, the order of sentences plays a pivotal role. However, in most voice synthesize tasks, models are trained using cut sentences below ten seconds to ensure the quality of the audio. As a compromise, we embed the orders and add it to the sentence before decoding. This can distinguish the melody of different sentences but consider their contextual relationship.

4 Method

As shown in Figure 4, our model is composed of the Character-level Feature Extraction (CFE) module and the Poetry Intoning Synthesis (PIS) module. In the CFE module, we meticulously extract essential features such as phonemes, rhymes, tones, and stop tokens from the input raw text. Subsequently, in the PIS module, we leverage all these extracted features as textual input to synthesize poetry intoning audio. This structured approach ensures a comprehensive and effective transformation from text to intoning audio.

4.1 CFE module

In CFE module, we aim to derive accurate acoustic feature from raw texts according to the guiding principles. Transformer is employed as the backbone architecture with text encoder and feature decoders. Texts, tones, phonemes, stop tokens and orders are denoted by $x, t, p, s, r, o \in \mathbb{R}^n$.

Attention mechanism is employed in the encoder layer and decoder layer of transformer, which can be calculated as shown in Equation 1, where Q, K and V represents query, key and value.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

In poetry intoning, the determination of tones and rhymes is straightforward by characters and metrical patterns. But the phonemes and caesuras are also impacted by tones and rhymes. To encapsulate these dependencies, we adopt a two-step approach. Initially, we directly extract tones and rhymes and subsequently incorporate them with distinct weights when predicting phonemes and caesuras, as Equations 2 and 3. $h_t, h_p, h_r \in \mathbb{R}^{n \times H}$ where H stands for hidden size.

$$h_t^i = D_t(h_t^{i-1}, E(x_i)) \tag{2}$$

$$h_p^i = D_p(h_p^{i-1}, \lambda_x E(x_i) + \lambda_t h_t^i + \lambda_r h_r^i) \tag{3}$$

D and E denotes decoder and encoder where $D(a, b) = Attention(a, b, b)$ while $E(a) = Attention(a, a, a)$.

Moreover, in order to ensure the correctness of our method, we prepare a dictionary $\mathbb{D}$ including all possible pronunciations and tones for each character. When the model predict a pronunciation or tone that not in $\mathbb{D}(p_i)$, we will revise it by choosing one from the dictionary. This strategy can reduce error propagation in autoregressive progress.

When we obtain sequences of tones, rhymes, phonemes and stops, we combine phonemes and stops by inserting 'spn' symbols after the p_i when $st_i = 1$ to indicate that there should be a caesura like Equation 4.

$$\boldsymbol{p} = \begin{cases} \{p_1, p_2, \cdots, p_i, p_{i+1}, \cdots, p_n\}, & s_i = 0 \\ \{p_1, p_2, \cdots, p_i, \text{'spn'}, p_{i+1}, \cdots, p_n\}, & s_i = 1 \end{cases} \tag{4}$$

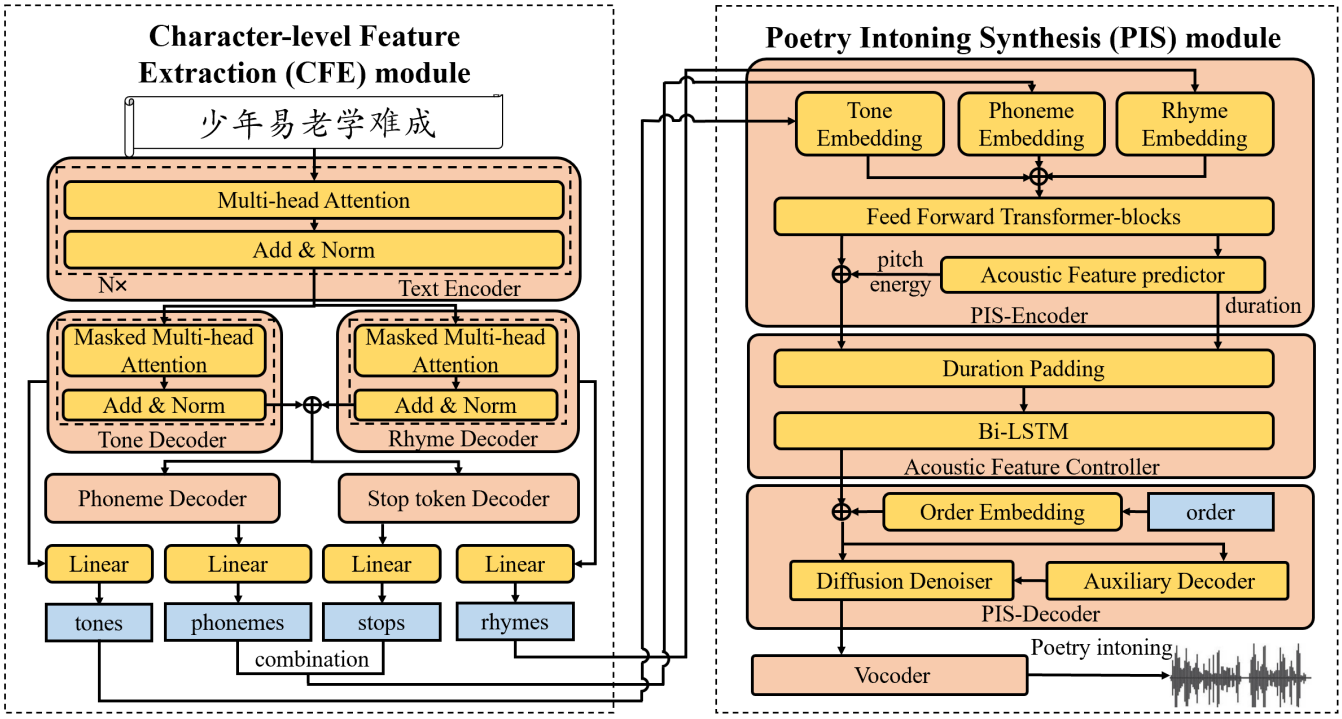


Figure 4: The structure of Intoner. The left part is the CFE module with characters in poem as input and character-level features as output. The right part is the PIS module with character-level features as input and poetry intoning audio as output.

4.2 PIS module

After CFE, we get sequence of tones, rhymes and phonemes with caesuras, which are parallel inputs of PIS. We use diffusion strategy like Figure 5 (b) to construct the PIS module. Given the sample from distribution $\mathbf{y}_0 \sim q(\mathbf{y}_0)$, the diffusion process is a Markov chain with fixed parameters [Ho *et al.*, 2020] that gradually adds Gaussian noise to the data with given variance schedule $\beta = \{\beta_1, ..., \beta_T\}$ in T steps:

$$q(\mathbf{y}_t|\mathbf{y}_0) = \mathcal{N}(\mathbf{y}_t; \sqrt{\bar{\alpha}_t}\mathbf{y}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{5}$$

where $\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$

Then, the reverse process is a Markov chain with learnable parameters θ to approximate intractable reverse transition distributions $q(\mathbf{y}_{t-1}|\mathbf{y}_t)$. To be specific:

$$p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) := \mathcal{N}(\mathbf{y}_{t-1}; \mu_\theta(\mathbf{y}_t, t), \sigma_t^2\mathbf{I}) \tag{6}$$

where θ is shared at every t step, and σ_t^2 are set to untrained time dependent constants. Thus the whole reverse process can be defined as:

$$p_\theta(\mathbf{y}_{0:T}) := p(\mathbf{y}_T) \prod_{t=1}^{T} p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) \tag{7}$$

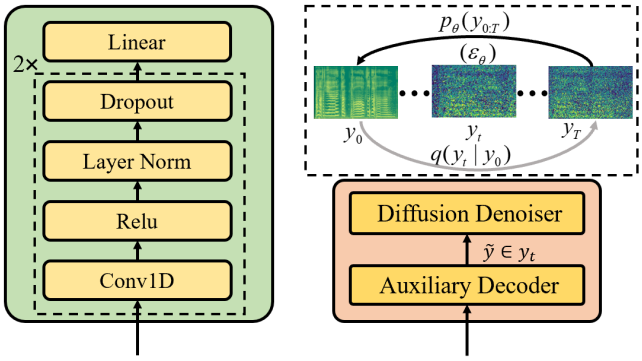
Reparameterize and choose the parameterization:

$$\mu_\theta(y_t, t) = \frac{1}{\sqrt{\alpha_t}}(y_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(y_t, t)) \tag{8}$$

Inspired by [Liu *et al.*, 2022], we employ shallow diffusion strategy to improve the efficiency and effectiveness of our model. The word "shallow" means that the reverse stage starts from a rough sample y_t rather than Gaussian noise y_T. Firstly, we construct a autoregressive model to predict a rough mel-spectrum $\tilde{y}$ of the intoning. Then, assume that $\tilde{y}$ is a type of y_t which includes t-step noise, when inference, we can get the denoised mel-spectrum step by step with the reverse step:

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}}(y_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(y_t, x, t)) + \sigma_t\mathbf{z} \tag{9}$$

In detail, we embed the tones, phonemes and rhymes as parallel inputs and then add the embeddings together in the PIS-Encoder. Feed Forward Transformer (FFT) blocks are em-



(a) Acoustic Feature predictor (b) shallow diffusion process

Figure 5: Details of the Acoustic Feature Predictor and shallow diffusion process.

ployed to derive self-attention like Equation 1. We also construct a feature predictor like Figure 5 (a) to explicitly control some acoustic variables including pitch pi' and duration d'.

After prediction, we lengthen each phoneme to the predicted duration by padding in Acoustic Feature Controller. Since padding only will cause distortion, we use layers of Bi-LSTM in order to reduce the noise and improve the overall quality. Before decoding, we first add an order embedding module to make the model aware of the contextual relationship. Then, we use an auxiliary decoder to synthesize a rough mel-spectrum $\tilde{y}$ which will be seen as the initial state of reverse step of the shallow diffusion. After denoising with diffusion strategy, the mel-spectrum will be sent into a pretrained vocoder to generate the output intoning audio.

4.3 Training Process

1) In CFE module, losses are calculated by cross-entropy loss separately. The prediction sequence should correspond with the raw text sequence one by one. However, the model will predict a character repeatedly or neglect a character sometimes. This error will propagate to the back and then impact the prediction accuracy sharply. In order to improve the accuracy, we set a higher weight to punctuation prediction loss to block the error propagation.

$$\mathcal{L}_k = -\frac{1}{\boldsymbol{N}} \sum_i \sum_c \lambda_c y_{c,i} log(k_{c,i}|x_i) \tag{10}$$

where c stands for character and punctuation, k stands for tones, rhymes, phonemes and stop tokens, $\mathcal{L}_{CFE} = \sum \mathcal{L}_k$.
2) In PIS module, our training loss contains three parts: mel-spectrum losses, pitch losses, duration losses. Assume y_{mel} and $y_{postmel}$ as target mel-spectrum and postnet mel-spectrum, while pi, d as target pitch and duration.

We train Intoner by warmup and main stage. At warmup stage, we separately train an auxiliary Decoder with the PIS-Encoder and Acoustic Feature Controller. L1 loss and MSE loss are used as:

$$\mathcal{L}_{mel} = \sum |y'_{mel} - y_{mel}| \tag{11}$$

$$\mathcal{L}_d = \sum \sqrt{(d - d')^2} \tag{12}$$

$$\mathcal{L}_{warmup} = \mathcal{L}_{mel} + \mathcal{L}_{postmel} + \mathcal{L}_{pi} + \mathcal{L}_d \tag{13}$$

where $\mathcal{L}_{postmel}$ and $\mathcal{L}_{pi}$ are calculated similarly to $\mathcal{L}_{mel}$.

The main stage aims at making the random noise ϵ_θ predicted by denoiser close to the Gaussian noise $\epsilon \sim \boldsymbol{N}(0, \mathbf{I})$, the denoiser loss can be written as:

$$\mathcal{L}_{main} = \mathbb{E}_{y_0,\epsilon}[C(t)||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t})y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2] \tag{14}$$

5 Chinese Poetry Intoning Dataset

To the best of our knowledge, there is no off-the-shelf Chinese poetry intoning dataset with labels of tones and rhymes. So we collect a large-scale Chinese poetry intoning dataset, including one for training CFE module (CFESet) and the other for training PIS module (PISSet).

type	CFE	PIS
five-character quatrains	1220	435
seven-character quatrains	4576	1743
five-character metrical poetry	1538	679
seven-character metrical poetry	1312	407
others	2119	375

Table 1: Distribution of different poetry types in our dataset.

5.1 CFESet

To extract character-level features, we collected 10,765 poems in different poetry type as shown in Table 1. Each piece of data includes the texts of the poem, as well as their corresponding pronunciations, tones, rhymes and stop tokens. It covers most of the commonly used characters in poetry and all their pronunciations and tones.

5.2 PISSet

We choose 3,639 poems from CFESet to collect text-audio pairs from Ximalaya platform[1] to build the PISSet. The intonings are from 24 intoners who are famous poetry scholars. Each intoner exhibits a distinct style, with relatively consistent melodies within their repertoire. We specifically focus on seven-character and five-character poems in PISSet since they represent the typical structure of ancient Chinese poetry.

First, considered that PIS module can reach a better performance when trained by short audio, we cut the data into pieces of sentence, resulting in a total of 20,112 sentences, each lasting less than 20 seconds. Then, we use a MFA [McAuliffe *et al.*, 2017] model trained by AISHELL3 dataset [Shi *et al.*, 2020] to align the phonemes and audio segments. During preprocessing, we get pitch and duration as target features of each segment according to the results of alignment. All audio samples are resampled to 22050 Hz.

6 Experiment

6.1 Character-level Feature Extraction

Implementation Details In CFE, the dimension of embedding and hidden state is 256 and 1024 separately, while the dimension of keys and values is 64. The encoder and decoder possess 3 layers of FFT blocks with 4-head attention. The weights are chosen as $\lambda_x = 3$, $\lambda_t = 2$, $\lambda_r = 1$, while the loss weights are $\lambda_{punc} = 8$ and $\lambda_{char} = 1$.

Metrics To evaluate the effect of the models, we choose two kinds of accuracy as metrics. One is the accuracy of the predicted features of the entire sentence, the other is of the polyphone characters. We choose 213 poems of different forms from test set of CFESet to construct the evaluation.

Comparison models Since there is no existing model to synthesize these features, we have tried different methods to do the prediction. 1) rule based, where phonemes are predicted by *lazy_pinyin* from *pypinyin*[2], and tones are predicted according to the dictionary $\mathbb{D}$; 2) random, phonemes and

[1]https://www.ximalaya.com

[2]https://pypi.org/project/pypinyin

Method	Total				Five-character				Seven-character			
	MOS ↑	PC↑	IUD ↑	IMS↑	MOS ↑	PC ↑	IUD ↑	IMS ↑	MOS ↑	PC ↑	IUD ↑	IMS ↑
Ground Truth (GT)	4.53±0.07	4.44	4.49	4.49	4.56	4.46	4.56	4.52	4.51	4.43	4.43	4.47
GT (vocoder)	4.19±0.07	4.35	4.34	4.35	4.25	4.38	4.44	4.37	4.13	4.31	4.25	4.34
VITS2	1.58±0.14	1.48	1.76	1.89	1.64	1.47	1.81	1.93	1.52	1.49	1.72	1.84
PromptSinger	2.72±0.11	2.73	2.50	2.39	2.81	2.81	2.53	2.41	2.64	2.65	2.46	2.38
FastSpeech2	2.38±0.10	2.85	2.82	2.57	2.28	2.88	2.88	2.44	2.47	2.82	2.76	2.69
DiffSpeech	3.00±0.10	3.22	3.25	2.76	3.04	3.33	3.34	2.76	2.97	3.10	3.15	2.77
Our model	**3.18±0.10**	**3.47**	**3.46**	**3.02**	**3.27**	**3.67**	**3.51**	**3.08**	**3.10**	**3.26**	**3.41**	**2.96**

Table 2: Experimental result of our experiment (with 95% confidence intervals). The results of Tacotron2 and Transformer-TTS are not included since they cannot synthesize any normal voice when adapted to PIS.

Method	phoneme		tone ↑	rhyme ↑	stop ↑
	all ↑	poly ↑			
Random	77.89%	42.51%	89.18%	94.81%	66.27%
Rule based	96.40%	85.75%	97.17%	96.37%	73.35%
Parallel	89.11%	87.56%	97.23%	98.19%	96.79%
CFE w/o. $\mathbb{D}$	93.35%	91.11%	96.13%	98.78%	97.12%
CFE	**98.69%**	**96.49%**	**98.35%**	**99.19%**	**97.49%**

Table 3: Accuracy Evaluation on CFE.

tones are randomly chosen, while rhymes and stop tokens are decided randomly yes or no; 3) parallel-decoder transformer, which set 4 decoders in parallel; 4) CFE without $\mathbb{D}$.

Experimental Results Table 3 shows that, based on the CFE module and the dictionary $\mathbb{D}$, our model outperforms other baselines, especially in predicting pronunciations and stop tokens. Since many characters have only one pronunciation, our model don't outperform the rule-based method in entire prediction much significantly. However, it showcases substantial advantages in predicting polyphone characters. This can greatly improve the quality of the synthesized audio in the later PIS module.

6.2 Poetry Intoning Synthesis

Implementation Details The encoder and decoder possess 4 layers with 2-head attention. All the dropout rates are set to 0.1, and the hidden sizes are 256. We use Adam optimizer with betas of 0.9 and 0.98, batch size of 48 and train with 120k warmup steps and 160k main steps. All the models are trained on NVIDIA 4090 (single card) for 17 hours (8 hours for warmup and 9 for main).

Metrics We choose MOS [Loizou, 2011] as a primary metric, which is commonly used in TTS and SVS tasks to evaluate the subjective quality of audio. 12 native speakers with poetry intoning knowledge are invited to listen to the audio samples while 20 normal native speakers are invited to participate in ablation study and universality study. Furthermore, we propose some novel metrics like MOS for poetry intoning content according to three-level principles:

Pronunciation Correctness (PC) metric is for character-level features, which evaluates the correctness of pronunciation and tone. Intoning unit division (IUD) metric is for sentence-level features, which evaluates whether the intoning unit is divided correctly and the caesura is rhythmic. Intoning melody style (IMS) is for discourse-level features which evaluates the melody and integrity.

All of our test inputs are the whole poem of raw texts in Chinese characters from test sets rather than phonemes of cut sentences in training. This means that all the test inputs are obviously different from training inputs. We choose 10 five-character poems and 10 seven-character poems (4 sentences for each) as test set.

Comparison models Current models are unable to perform PIS directly, among which most SVS models will fail due to the lack of predefined musical notes in intoning. Thus, for comparison, we select some models which can be adapted to PIS and train them on our dataset. Tacotron2 [Elias *et al.*, 2021], Transformer-TTS [Okamoto *et al.*, 2020] and FastSpeech2 [Ren *et al.*, 2020] are autoregressive models while DiffSpeech [Liu *et al.*, 2022] adopts diffusion and VITS2 [Kong *et al.*, 2023] integrates GAN and implicit alignment. Prompt-Singer [Wang *et al.*, 2025] is an SVS model based on large language models.

Experimental Results We evaluate the models through both statistic data and synthesized mel-spectrum figures. As shown in Table 2, we can find that: 1) Tacotron2 and Transformer-TTS cannot be adapted to PIS task since it cannot synthesize any normal voice. 2) VITS2 performs poorly since it cannot pronounce correctly, perhaps due to the bad alignment of long-duration phonemes. 3) DiffSpeech can outperform other baseline models since the diffusion denoiser can reduce noise and make the audio sound more clear; 4) Our model performs best whether in five-character or seven-character poems since all the metrics are promoted compared with other methods, especially in sentence-level and discourse-level which consider the integrity and contextual relationship. However, there is still disparity between the synthesized intoning and human records (GT) especially in melody. As some listeners have noted, human intoning is more emotional and natural than current synthesized audio, which could be an area for improvement in future research.

The effectiveness can also be demonstrated through the mel-spectrum images of the audio. Figure 6 illustrates that the mel-spectrum synthesized by our model closely matches all three-level principles. In character-level, the melody curves keep highly consistent with the tones of the characters, and the duration of each character is also appropriate according to Section 3.1 (characters in 'P' should be lengthened than others while characters in 'R' should be intone significantly

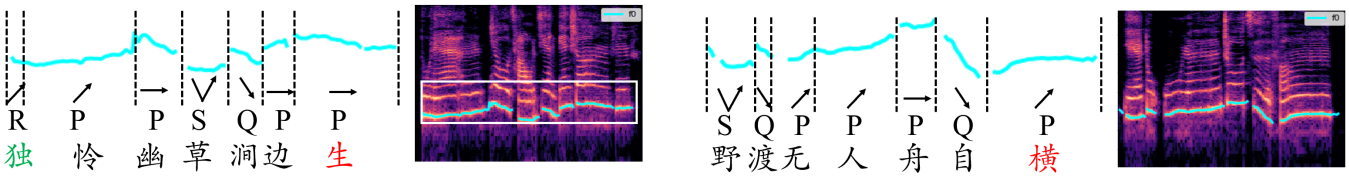


Figure 6: Samples of mel-spectrum synthesized by Intoner. The blue lines represent melody while arrows and capital letters stand for tones, which are highly consistent. Characters in green stand for tone in "R" which should be intoned shortly while in red are rhymes which should be lengthened in intoning.

Method	Character		Sentence	Discourse	Total	Other	
	P ↓	D ↓	SD ↓	Mel ↓	Sum ↓	MOSnet ↑	SRMR ↑
Our model(with Bi-LSTM)	**1.756**	0.712	**0.053**	0.161	**2.682**	2.943±0.033	**11.298±0.456**
Our model(with CNN)	2.000	0.711	0.065	**0.157**	2.933	2.930±0.045	11.178±0.514
Our model(with LSTM)	1.896	**0.699**	0.056	0.161	2.812	**2.957±0.051**	11.275±0.545

Table 4: Objective result of ablation study (with 95% confidence intervals), where P, D, SD, Mel represents for pitch loss, phoneme duration loss, sentence duration loss, mel-spectrum loss, Sum is the total loss of all, MOSnet and SRMR are judged by published pretrained models.

Method	CMOS ↑
Intoner	0
w/o. PE	-1.026
w/o. OE	-0.398
w/o. AFC	-0.003

Table 5: Subjective result of ablation study

Type	Method	CMOS ↑
Song iambics	Intoner	0
	FastSpeech2	-1.404
	DiffSpeech	-1.384
Shijing poetry	Intoner	0
	FastSpeech2	-0.668
	DiffSpeech	-0.389

Table 6: Universality study results

short). In sentence-level, it is obvious that the curve is intermittent which contains correct caesuras. Finally, the rhyming characters in red are lengthened significantly, which meets the principles at discourse-level. Overall, our model showcases substantial advancements in synthesizing coherent and natural intoning audio.

Ablation Study To further study the effectiveness of the different components of Intoner, we conduct an ablation study which consists of both subjective and objective evaluation.

Firstly, in Acoustic Feature Controller (AFC), we have tried different networks, including a 2-layer LSTM, Bi-LSTM, and Conv1D (CNN) network. We consider objective metrics [Chu and Peng, 2006] including different losses in test step as well as SRMR [Falk *et al.*, 2010] and MOSnet [Lo *et al.*, 2019] metrics across 50 complete poem generation. As shown in Table 4, although CNN and LSTM can perform best in some level, but Bi-LSTM network reaches the lowest overall loss. Generally, we consider Bi-LSTM has the best effectiveness, so we choose it to build the AFC.

Secondly, we remove PIS-Encoder (PE), Order Embedding (OE) and AFC separately to assess their individual contributions. We use CMOS [Loizou, 2011] as a subjective metric to compare the voice quality. Table 5 illustrates that their absence significantly impacts the overall integrity and performance of the model.

Universality of our approach In ancient China, intoning is not only appropriate for poetry, but also for *Song* iambics, *Shijing* poetry, and even prose. The primary distinction is that these forms do not consistently adhere to the 5 or 7 characters per sentence, nor do they consistently comprise 4 or 8 sentences. It is important and interesting to explore the applicability of Intoner to other literary forms. Here, we conduct further experiments on some *Song* iambics and *Shijing* poetry. As shown in Table 6, Intoner can also outperform other models significantly in terms of CMOS. This indicates that Intoner can be used as a universal intoning tools in various ancient Chinese literary forms.

7 Conclusion And Future Work

In this paper, we propose a novel and challenging task which is Chinese poetry intoning synthesis. Due to the scarcity of existing intoning audio, our work contributes to the preservation of this cultural art form by autonomous synthesis. We summarize three-level principles according to poetry metrical patterns, employ a CFE module to extract intoning features and then integrate them into a diffusion PIS module to synthesize poetry intoning. Moreover, we conduct three-level metrics for PIS and collect a sizable dataset encompassing text-audio pairs with feature annotations. Experimental results demonstrate that our model can produce correctly-pronounced, rhyming and natural poetry intoning. Our future works aim to enhance the emotional resonance of synthesized audio and explore intoning synthesis in other literary forms.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2023YFF0904900).

References

[Chu and Peng, 2006] Min Chu and Hu Peng. Objective measure for estimating mean opinion score of synthesized speech, April 4 2006. US Patent 7,024,362.

[Cong, 2021] Longmei Cong. Intoning teaching: Using sound to inherit classics(in chinese). *People's Education*, No.846(79), 2021.

[Dai *et al.*, 2024] Shuqi Dai, Ming-Yu Liu, Rafael Valle, and Siddharth Gururani. Expressivesinger: Multilingual and multi-style score-based singing voice synthesis with expressive performance control. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24. Association for Computing Machinery, 2024.

[Deng *et al.*, 2025] Junlin Deng, Ruihan Hou, Yan Deng, Yongqiu Long, and Ning Wu. High-quality text-to-speech implementation via active shallow diffusion mechanism. *Sensors*, 2025.

[Du *et al.*, 2024] Chenpeng Du, Yiwei Guo, Feiyu Shen, Zhijun Liu, Zheng Liang, Xie Chen, Shuai Wang, Hui Zhang, and Kai Yu. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Du, 2023] Xuejing Du. The necessity and feasibility of teaching ancient poetry intoning(in chinese). *Advances in Education*, 2023.

[Elias *et al.*, 2021] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*, 2021.

[Falk *et al.*, 2010] Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:1766–1774, 2010.

[Gu *et al.*, 2021] Yu Gu, Xiang Yin, Yonghui Rao, Yuan Wan, Benlai Tang, Yang Zhang, Jitong Chen, Yuxuan Wang, and Zejun Ma. Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[Huang *et al.*, 2024] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Hwang *et al.*, 2025] Ji-Sang Hwang, Sang-Hoon Lee, and Seong-Whan Lee. Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models. *Neural Networks*, 2025.

[Ju *et al.*, 2024] Zeqian Ju, Yuancheng Wang, and et al. Shen Kai. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

[Kim *et al.*, 2021] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[Kong *et al.*, 2020] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[Kong *et al.*, 2023] Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*, 2023.

[Kumar *et al.*, 2019] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[Li *et al.*, 2019] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713, 2019.

[Li, 2017] Changji Li. The historical tradition and rules of poetry intoning(in chinese). *Journal of Jiangsu Normal University(Philosophy and Social Sciences Edition)*, 43:1–14, 2017.

[Liu *et al.*, 2022] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11020–11028, 2022.

[Lo *et al.*, 2019] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

[Loizou, 2011] Philipos C Loizou. Speech quality assessment. *Multimedia analysis, processing and communications*, pages 623–654, 2011.

[Lu *et al.*, 2020] Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou. Xiaoicesing: A high-quality and inte-

grated singing voice synthesis system. *arXiv preprint arXiv:2006.06261*, 2020.

[McAuliffe *et al.*, 2017] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.

[Nishimura *et al.*, 2016] Masanari Nishimura, Kei Hashimoto, and Keiichiro et al. Oura. Singing voice synthesis based on deep neural networks. *Interspeech*, 2016.

[Okamoto *et al.*, 2020] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. Transformer-based text-to-speech with weighted forced attention. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6729–6733. IEEE, 2020.

[Oord *et al.*, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Prenger *et al.*, 2019] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

[Ren *et al.*, 2019] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

[Ren *et al.*, 2020] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[Shi *et al.*, 2020] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2017] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[Wang *et al.*, 2025] Yongqi Wang, Ruofan Hu, Rongjie Huang, Zhiqing Hong, Ruiqi Li, Wenrui Liu, Fuming You, Tao Jin, and Zhou Zhao. Prompt-singer: Controllable singing-voice-synthesis with natural language prompt, 2025.

[Xu, 2014] Jianshun Xu. What is intoning(in chinese). *Chinese Language Teaching for Primary Schools*, 3, 2014.

[Yang, 2015] Mei Yang. Review of intoning(2011-2015)(in chinese). *People's Music*, 11, 2015.

[Yang, 2021] Mei Yang. Review of researches of intoning(2016-2020)(in chinese). *Music Life*, 2021.

[Yu *et al.*, 2024] Yifeng Yu, Jiatong Shi, Yuning Wu, Yuxun Tang, and Shinji Watanabe. Visinger2+: End-to-end singing voice synthesis augmented by self-supervised learning representation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[Zhang *et al.*, 2024] Yu Zhang, Rongjie Huang, Ruiqi Li, JinZheng He, Yan Xia, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Stylesinger: Style transfer for out-of-domain singing voice synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Zhu and Zhu, 2008] Chengsong Zhu and Yaoting Zhu. A new chinese speech synthesis method apply in chinese poetry learning. In *Advances in Web Based Learning-ICWL 2008: 7th International Conference, Jinhua, China, August 20-22, 2008. Proceedings 7*, pages 356–365. Springer, 2008.

[Zhu, 2013] Lixia Zhu. Research on intoning rules based on recording(in chinese). *Research Trends in Chinese Poetry*, pages 59–63, 2013.