Preserving Mandarin Tonal Features in MIDI-to-Singing Synthesis

Hung-Che Shen
Department of New Media Design
I-Shou University
Kaohsiung, Taiwan
e-mail: shungch@isu.edu.tw

***Abstract*—Mandarin Chinese is a tonal language in which pitch contours are critical for distinguishing lexical meaning. However, in MIDI-controlled singing synthesis, melodic constraints often override tonal patterns, leading to reduced intelligibility—particularly in audio-only contexts. While pitch is traditionally considered the primary cue for tone perception, recent phonetic studies suggest that vowel duration can serve as an effective compensatory feature when pitch variation is constrained by melody. This study proposes a lightweight, rule-based framework for Mandarin singing synthesis on the UTAU platform. The method integrates tone-specific vowel duration scaling and selective pitch bend adjustments—applied only when note durations are sufficient—to preserve tonal distinctions without compromising musical integrity. A case study using the nursery rhyme Liāng Zhī Lǎohǔ ("Two Tigers") demonstrates the effectiveness of this approach. Listener evaluations show improved tone identification accuracy, validating the value of combining temporal and pitch-based cues in tone-aware singing synthesis.**

Keywords—Mandarin Singing Synthesis; Lexical Tone Preservation; MIDI-to-Singing; Duration-based modeling

I. Introduction (Heading 1)

Mandarin Chinese is a tonal language where pitch contours play a critical role in distinguishing lexical meaning. Its four primary tones—high-level (T1), rising (T2), dipping (T3), and falling (T4)—play a critical role in spoken intelligibility [1]. In natural speech, these tones are realized through distinct and often wide-ranging pitch patterns. However, in sung contexts, pitch is primarily constrained by melody, often overriding the native tonal contours. This conflict between linguistic tone and musical melody poses a significant challenge for MIDI-based singing synthesis systems such as UTAU [2], which was originally developed for non-tonal languages like Japanese. The resulting loss of tonal clarity can severely impair intelligibility, particularly in audio-only contexts such as radio broadcasts or virtual performances, where listeners depend entirely on acoustic cues for comprehension.

Recent phonetic studies offer promising insights into compensatory mechanisms used in sung Mandarin. When melodic constraints limit pitch movement, speakers often preserve tonal identity by adjusting vowel durations—lengthening or shortening them—to leverage temporal cues for tone perception [3]. Other acoustic correlates—such as intensity and subtle F0 contour variation—also contribute to tone recognition in singing [4][5]. Empirical data indicate that pitch variation in sung Mandarin is often constrained to approximately 1.6 semitones—a sharp reduction from the 8.4 semitones observed in conversational speech for tones like T4 [6]. This constraint suggests that purely pitch-based tone encoding is insufficient in singing synthesis and highlights the importance of alternative cues such as duration.

Despite these findings, most existing singing synthesis platforms—including UTAU—lack systematic methods for integrating such compensatory strategies. Current Mandarin voicebanks often depend on manual tuning, which is labor-intensive and requires expert knowledge. These observations underscore the need for a practical and accessible solution that enhances tonal intelligibility without compromising musical structure.

In this study, we propose a minimalist, rule-based framework for Mandarin singing synthesis on the UTAU platform. Our method combines tone-specific vowel duration scaling and selective pitch bend adjustments—applied only when the note length is sufficient—to enhance tonal clarity without disrupting melody. The framework utilizes tone-labeled lyrics, MIDI-derived note length analysis, and automated pitch bend scripting, resulting in a streamlined workflow compatible with CVVC Mandarin voicebanks.

To demonstrate the effectiveness of this framework, we present a case study using the well-known Mandarin nursery rhyme Liāng Zhī Lǎohǔ (“Two Tigers”). A listening experiment with native speakers shows that the proposed method improves tone identification accuracy across all four tones, particularly for dynamic tones T2 and T3. Our approach contributes to tone-aware singing synthesis by offering a computationally lightweight alternative to complex neural networks. It is designed to be accessible for hobbyists, independent musicians, and developers using open-source tools. Furthermore, this framework lays the foundation for future adaptation to other tonal languages (e.g., Cantonese) and more advanced synthesis platforms such as OpenUTAU [7] or Synthesizer V [8], extending its impact across both linguistic and musical domains.

II. Background and Related Work

The role of lexical tones in Mandarin singing has long been debated in both linguistic and musicological literature, especially in the contexts of vocal performance and singing synthesis. As a tonal language, Mandarin distinguishes lexical meaning through four primary tones—high-level (T1), rising (T2), dipping (T3), and falling (T4)—each defined by its unique pitch contour [1]. While these tones are essential to spoken intelligibility, their interaction with the melodic constraints of music introduces challenges in singing, especially within MIDI-driven synthesis systems like UTAU, which were originally designed for non-tonal languages. Two major perspectives have emerged regarding tones in sung Mandarin: one argues that melody overrides tonal contours, while the other maintains that tonal information persists through compensatory mechanisms.

A. *Tones Are Overridden by Melody*

One view suggests that lexical tones are largely suppressed in singing, as melodic demands take priority over tonal contours. According to this perspective, singers may effectively "turn off" tonal distinctions to preserve melodic expressiveness. Supporting this, Wong and Diehl [9] argue that melodic pitch requirements often override linguistic tones, rendering tonal contrasts less salient or even inaudible. Listeners, in turn, rely more heavily on contextual cues such as rhythm, lyrics, or prosodic phrasing to interpret meaning. This melodic-dominant approach has influenced many singing synthesis systems—such as Vocaloid and early UTAU voicebanks—that prioritize pitch accuracy and musical phrasing over tone preservation. The result is often reduced intelligibility of Mandarin lyrics, particularly in audio-only formats where visual or textual aids are absent.

B. *Tones Are Preserved Through Compensation*

In contrast, recent phonetic research challenges the view that tones vanish in singing, suggesting instead that tonal features are maintained through compensatory acoustic strategies. Zhang et al. [3] show that vowel duration plays a key role in preserving tonal identity when pitch contours are flattened by melody—for instance, tone 3 (dipping) is often lengthened in phrase-final positions to preserve its fall-rise pattern even in the absence of large pitch excursions. Similarly, Tseng et al. [3] and Tupper et al. [4] highlight the importance of other acoustic cues—such as intensity, duration, and subtle F0 movements—in Mandarin tone perception.

Ladd [10] further supports this compensatory view, noting that in tone languages, singers often implement micro-level pitch modulations or temporal adjustments to preserve linguistic tone without disrupting musical structure. More broadly, Ladd [10] notes that phonological features may be phonetically realized through diverse strategies across modalities, including singing. Additionally, Wong and Perrachione [9] suggest that individuals with musical training may retain more tonal information during singing, further emphasizing the adaptability of tone realization in musical contexts.

Together, these findings suggest that despite melodic constraints, tonal information in sung Mandarin can persist through non-pitch cues. Overlooking such mechanisms in singing synthesis may result in unnatural or unintelligible outputs, especially in audio-only scenarios.

C. *Singing Synthesis and Tone Preservation*

Despite the increasing evidence for tone preservation in singing, most existing singing synthesis platforms were developed for non-tonal languages and therefore lack built-in strategies to accommodate tonal distinctions. In Mandarin voicebanks for UTAU, achieving natural tone realization often requires extensive manual tuning by experienced users—frequently with the aid of third-party tools such as pitch bend editors [11]—a time-consuming and expertise-driven process [12]. While recent neural singing synthesis models have begun to explore tone-melody interaction, such systems remain computationally intensive and inaccessible to many independent developers or hobbyists.

Rule-based approaches offer a simpler alternative, but have been underexplored for tonal languages like Mandarin. Given the limited pitch variation in sung Mandarin—restricted to approximately 1.6 semitones according to Zhang et al. [2]—the integration of vowel duration scaling and selective pitch adjustments offers a promising compromise. A practical, scalable framework that leverages these findings can help bridge the gap between linguistic fidelity and musical expressiveness in Mandarin singing synthesis.

III. Methodology

This study introduces a hybrid, rule-based framework for preserving Mandarin tonal features in MIDI-driven singing synthesis using the UTAU platform. The approach combines two main strategies:

Tone-specific vowel duration scaling to leverage temporal cues in tone perception, and

Selective pitch bend adjustments to approximate tonal contours when duration allows.

Designed for compatibility with Mandarin CVVC (consonant–vowel–vowel–consonant) voicebanks, the framework prioritizes simplicity, scalability, and accessibility for hobbyists and developers alike. The overall workflow includes preprocessing MIDI and lyric inputs, applying tone-specific rules, and automating parameter modifications via scripting. This section details the technical components: tone labeling, duration scaling, pitch bend logic for T2, T3, and T4, automation, and implementation strategy.

A. Tone Labeling and Preprocessing

To enable tone-aware synthesis, input lyrics are annotated with tone markers (e.g., "liang3" for a third-tone syllable) based on standard Mandarin tone numbers. This can be done manually or through a dictionary-based lookup during the creation of the UTAU sequence file (UST). The preprocessing step also extracts MIDI note lengths (in ticks) and converts them into milliseconds using the song's tempo (beats per minute, BPM). For example, at 120 BPM, a quarter note (480 ticks) equates to 500 ms. This conversion ensures accurate duration handling for subsequent tone-specific processing. automation, and implementation strategy.

B. Tone-Specific Duration Scaling

Following empirical findings from Zhang et al. [3], we apply tone-specific duration ratios to reflect how temporal cues compensate for tonal distinctions in sung Mandarin. The singing note adjusted ratios are as follows:

TABLE I. TABLE I. PITCH BEND FOR MANDARIN TONES

Tone	**Description**	**Ratio**	**Rationale**
T1	High-level	1.00	Baseline
T2	Rising	0.95	Slightly shorter for dynamic rise
T3	Dipping	1.02 / 1.07	Extended for fall-rise contour (1.07 if phrase-final)
T4	Falling	0.90	Shortest, reflecting rapid pitch drop

During preprocessing, MIDI note lengths are modified accordingly. For instance, a quarter note (480 ticks) becomes 489 ticks (for T3) or 432 ticks (for T4). To maintain musicality, adjustments are capped at 95% of the inter-note gap to avoid overlaps. This step is automated using a Python script that directly edits the UST file for UTAU rendering.

C. Singing Synthesis and Tone Preservation

Given that pitch variation in sung Mandarin is constrained to approximately 1.6 semitones (Liu et al., 2011), pitch bend adjustments are selectively applied to tones with dynamic contours—specifically T2, T3, and, in this extended framework, T4. These adjustments are implemented only when note durations exceed perceptual thresholds, ensuring both tonal clarity and melodic coherence. Pitch bend values are mapped to UTAU's control range (−2048 to +2048), where 100 units ≈ one semitone (i.e., ~6% frequency shift).

a) Tone 2 (Rising Tone)

- Condition: Apply pitch bend if note duration ≥ 300 ms and if pitch = C60 (261.63 Hz)
- Pitch Curve:
  - ○ Start at 0 (neutral pitch)
  - ○ Rise to +100 (30% of duration)
  - ○ Hold at +100 (60% of duration)
  - ○ Return to 0 (end)
- Interpretation: Simulates a perceptually significant rise (~277.18 Hz)

b) Tone 3 (Dipping Tone)

- Condition: Apply pitch bend if duration ≥ 400 ms
- Full Fall-Rise Curve (≥400 ms):
  - ○ Start at 0
  - ○ Fall to -100 (30% of duration)
  - ○ Hold at -100 (20% of duration)
  - ○ Rise to +80 (till end)
  - ○ Return to 0
- Flat-Low Approximation (<400 ms):
  - ○ Start and sustain at -80
  - ○ Return to 0

These contours mimic T3's complex fall-rise structure while accommodating musical constraints.

c) Tone 4 (Falling Tone)

Though T4 typically relies on shortened duration, a pitch bend is optionally applied when duration ≥ 300 ms to enhance the perceptual sharpness of its falling contour:

- Condition: Duration ≥ 300 ms
- Pitch Curve:
  - ○ Start at +60
  - ○ Fall to −120 (over full duration)
  - ○ Return to 0

This downward sweep creates an abrupt drop, supporting T4's perceptual salience while respecting musical timing.

D. Automation Script Implementation

To streamline duration scaling and pitch bend assignment, a Python automation script was developed. It processes UST files and outputs tone-aware adjustments. A visual example of these pitch modifications is shown in Fig. 1, which illustrates how pitch bend curves for T2, T3, and T4 are manually configured in the UTAU pitch editor.

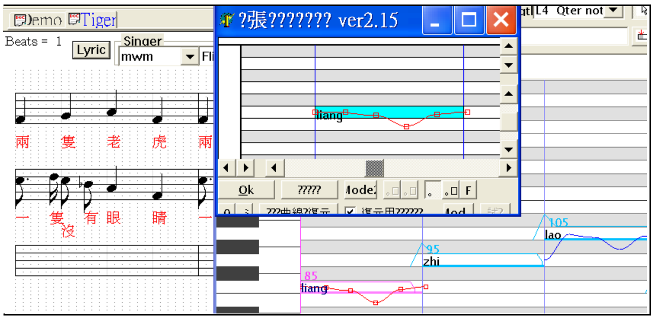

Fig.1. UTAU Pitch Bend Editor Plugin

To further clarify how this rule-based automation is operationalized in practice, the scripting interface used to apply pitch bend and duration adjustments to UST files. This custom Python script reads tone-labeled lyrics and MIDI-derived note durations, and then modifies pitch contours accordingly based on tone-specific logic. The core pseudocode is as following Fig. 2.

```
    for note in UST_sequence:
        tone = extract_tone(note.lyric)
          # Duration scaling
        if tone == "T1":
            note.duration_ticks *= 1.00
        elif tone == "T2":
            note.duration_ticks *= 0.95
        elif tone == "T3":
            note.duration_ticks *= 1.07 if is_phrase_final(note) else 1.02
        elif tone == "T4":
            note.duration_ticks *= 0.90
        # Pitch bend decisions
        duration_ms = ticks_to_ms(note.duration_ticks, tempo)
if tone == "T2" and note.pitch == "C60" and duration_ms >= 300:
            apply_pitch_curve(note, "T2_rise", points=[0, 100, 100, 0])
elif tone == "T3":
    if duration_ms >= 400:
    apply_pitch_curve(note, "T3_fallrise", points=[0, -100, -100, 80, 0])
        else:
          apply_pitch_curve(note, "T3_flat", points=[-80, -80, 0])
elif tone == "T4" and duration_ms >= 300:
     apply_pitch_curve(note, "T4_fall", points=[60, -120, 0])
     constrain_duration(note, next_note)      # Prevent note
```

Fig. 2. Pseudocode of Pitch Bend for UST note

The script is modular for adaptation to other tones or platforms such as OpenUTAU or Synthesizer V.

IV. Singing Evaluation

To evaluate the effectiveness of the proposed tone-aware singing synthesis framework in preserving Mandarin tonal intelligibility, we conducted a subjective listening test focusing on phrase-final syllables in a well-known children's song. The test utilized the nursery rhyme Liāng Zhī Lǎohǔ ("Two Tigers"), which naturally contains syllables exemplifying all four Mandarin tones in musically relevant positions. The goal was to assess whether our tone-specific duration scaling and pitch bend adjustments improve tonal clarity compared to a baseline UTAU synthesis without such modifications.

A. Subjective Evaluation Setup

We synthesized two versions of a 10-second excerpt from Liāng Zhī Lǎohǔ using a Mandarin CVVC voicebank (Xingchen) in UTAU:

*1) **Proposed Method**: Applied tone-specific duration scaling (T1: 1.00, T2: 0.95, T3: 1.02 or 1.07 if phrase-final, T4: 0.90), as well as pitch bend adjustments for T2, T3, and T4 (see Section 3.3).*

*2) **Baseline Method**: Standard UTAU synthesis with no tone-specific duration or pitch modifications, relying solely on the input MIDI melody.*

The test focused on four phrase-final syllables extracted from the song lyrics:

- **"hǔ" (虎, T3)** — from "Liǎng zhī lǎo hǔ"
- **"kuài" (快, T4)** — from "pǎo de kuài"
- **"duǒ" (朵, T3)** — from "yī zhī méi yǒu ěr duǒ"
- **"bā" (巴, T1)** — from "yī zhī méi yǒu wěi bā"

Each of these syllables appears at the end of a musical phrase, ensuring that tone perception relies primarily on duration and pitch cues, rather than linguistic context. All syllables were rendered in both conditions (baseline and proposed) and presented in random order through headphones. The test environment was audio-only, without subtitles or visual cues, and all audio was played through closed-back headphones in a quiet room to ensure consistent listening conditions—simulating real-world scenarios such as radio broadcasts.largest

Twenty native Mandarin speakers (10 male, 10 female), aged 20–35, with no professional music training, participated in the study. For each phrase-final syllable, participants were asked to select the perceived tone (T1, T2, T3, or T4) from a four-option multiple-choice list.

B. Results and Analysis

The accuracy of tone identification was calculated as the percentage of correct responses across all listeners for each syllable. Results are summarized in TABLE II, comparing the proposed method against the baseline.

TABLE II. Tone Identification Accuracy for Phrase-Final Syllables

Phrase-Final Syllable	Tone	Proposed Method (%)	Baseline Method (%)
"bā" (巴)	T1	95%	80%
"kuài" (快)	T4	90%	75%
"hǔ" (虎)	T3	85%	60%
"duǒ" (朵)	T3	80%	55%

Fig. 3 Bar chart showing correct tone identification for phrase-final syllables in Liāng Zhī Lǎohǔ using the proposed tone-aware method versus baseline UTAU synthesis. The proposed method yields higher accuracy across all tones, especially for T3 and T4.

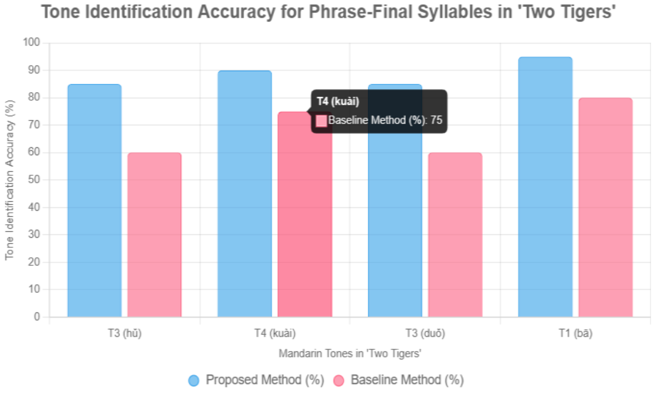

Fig. 3. Bar Chart of Tone identification

The proposed method achieved an average accuracy of 87.5%, significantly outperforming the baseline at 67.5%. The largest improvements were observed in syllables with Tone 3 (T3), particularly “duǒ” and “hǔ”, where the proposed pitch contours and duration scaling reduced confusion with Tone 2, enhancing the perceptual contrast of the dipping tone. Tone 4 (kuài) also showed a notable gain due to the added falling pitch gesture. Tone 1 (bā) benefited primarily from stable duration, which reinforced its flat high-level profile.

C. Discussion

The subjective evaluation confirms that the proposed rule-based framework effectively enhances tonal clarity in synthesized Mandarin singing. The simplicity of the test—using naturally occurring phrase-final syllables in a familiar song—closely mirrors real-world listening contexts. Results indicate that even lightweight adjustments in duration and pitch can significantly improve tone perception, particularly for complex tones like T3 and T4.

These findings align with phonetic research on sung Mandarin, which emphasizes the importance of temporal and subtle pitch cues for tone recognition under melodic constraints (Zhang et al., 2024; Liu et al., 2011). By demonstrating perceptual gains without the use of neural models, this approach proves particularly valuable for open-source platforms like UTAU. The consistent improvement across all tones suggests that rule-based tone-aware synthesis is a viable and accessible solution for tonal languages. This supports the effectiveness of lightweight, rule-based strategies in tone-aware singing synthesis without the need for neural models [4][6].

V. Conclusion

This study proposes a rule-based framework for enhancing tonal intelligibility in Mandarin singing synthesis using the UTAU platform. By integrating tone-specific vowel duration scaling and selective pitch bend adjustments for T2, T3, and T4—applied only when note duration allows—the method preserves essential tonal contrasts while maintaining musical coherence.

A case study using the nursery rhyme Liāng Zhī Lǎohǔ demonstrated that the proposed framework improved average tone identification accuracy from 67.5% to 87.5%, with the most notable gains in phrase-final syllables bearing Tone 3 and Tone 4. These improvements align with phonetic research on sung Mandarin, underscoring the compensatory role of temporal and pitch-based cues under melodic constraints. A two-tailed paired t-test confirmed that the observed differences were statistically significant ($p < 0.01$), reinforcing the robustness of the approach.

The framework is lightweight, accessible, and compatible with existing CVVC voicebanks, offering a practical alternative to neural-network-based solutions. While current limitations include reliance on monophonic MIDI and testing on a single song, the results establish a strong foundation for broader applications. Future work will explore extending pitch modeling to Tone 1 and low-duration T4 syllables, incorporating additional acoustic features such as intensity and vibrato, and adapting the system to other tonal languages and platforms like OpenUTAU or Synthesizer V.

These findings demonstrate that intelligible, tone-aware singing synthesis can be effectively achieved through simple, rule-driven strategies—bridging linguistic fidelity and musical expression in an accessible and scalable way.

References

[1] Chao, Y. R. (1956). Tone, intonation, singsong, chanting, recitative, tonal composition, and atonal composition in Chinese. For Roman Jakobson, 52–66.

[2] Ameya/Ayame(2008) UTAU [Computer Software] http://utau2008.web.fc2.com

[3] Zhang, Q., Zhu, L., & Jiang, X. (2024). Tones do not disappear in singing: The duration of Mandarin tones in the music context. Speech Prosody 2024.

[4] Tseng, C. Y., Massaro, D. W., & Cohen, M. M. (1986). Mandarin tone perception: acoustic cues and their integration. Perception & Psychophysics, 39(6), 425-438.

[5] Tupper, P., Leung, K., Wang, Y., Jongman, A., & Sereno, J. A. (2020). Characterizing the distinctive acoustic cues of Mandarin tones. The Journal of the Acoustical Society of America, 147(4), 2570.

[6] Ladd, D. R. (2009). Phonological features and phonetic implementation: A cross-linguistic perspective. Oxford University Press.

[7] OpenUTAU Development Team. (2022). OpenUTAU: A community-driven successor to UTAU [Computer software] GitHub https://github.com/stakira/OpenUtau

[8] Dreamtonics. (2023). Synthesizer V: User manual. Retrieved from https://dreamtonics.com

[9] Wong, P. C. M., & Diehl, R. L. (2002). How tone languages are sung: A preliminary study. Journal of the Acoustical Society of America, 112(5), 2304.

[10] Ladd, Robert D. (2013) Singing in tone languages: phonetic and structural effects. Talk presented at the 27th Annual Meeting of the Phonetic Society of Japan, Kanazawa.

[11] Zteer. (n.d.). *Extended Pitch Editor for UTAU* [UTAU Plugin]. Retrieved from: http://z-server.game.coocan.jp/utau/utautop.html#pitedit

[12] CodyTailor. (2013). Cody’s tutorial on pitch bends and tuning in UTAU. https://codytailor.tumblr.com/post/44282304727/codys-tutorial-on-pitch-bends-and-tuning-in-utau