From Lyric Generation to Mandarin Rap Synthesis:

A Dual Workflow with AI Writing and UTAU Singing

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

***Abstract*—This paper introduces a dual workflow for Mandarin rap creation, integrating AI-driven lyric generation with UTAU-based singing voice synthesis to deliver tonally accurate and rhythmically precise rap vocals. Prompt-based large language models (LLMs) produce rhyme-rich, thematically coherent Mandarin lyrics, which are converted into UTAU-compatible UST files using an online pinyin-tone annotation tool for pinyin-tone annotation, rhythm-aligned syllable processing, and Moresampler for clear articulation at high tempos. Unlike melodic synthesis, Mandarin rap demands precise control of short syllable durations and lexical tones. The combination of automated processing and manual UTAU tuning ensures a rap-like cadence and tonal clarity. Listener evaluation results show 86.5% tonal accuracy and 91.7% on-beat alignment, validating the system's effectiveness. This cost-effective, tone-aware approach enhances accessible music production and supports language education via rhythmic reinforcement.**

Keywords—AI lyrics; Mandarin rap; UTAU; ust generation

I. INTRODUCTION

Rap music, celebrated for its intricate rhymes, rhythmic delivery, and narrative richness, stands as one of the most linguistically complex forms of musical expression. Unlike melodic singing, rap demands precise control over syllabic timing and lyrical coherence. Crafting a compelling, rhymed narrative is a formidable creative task, as exemplified by the following sophisticated English verses:

Living off borrowed time, the clock ticks faster.
That'd be the hour they knock the slick blaster.
Dick Dastardly and Muttley with sick laughter.
A gunfight and they come to cut the mixmaster.

These lines feature multi-syllable rhymes, internal echoes, and vivid storytelling, elevating rap lyric writing beyond simple rhyme pairing. Translating these advanced techniques into Mandarin—a tonal language where pitch contours define lexical meaning—intensifies the challenge. Lyrics must rhyme, maintain semantic coherence, and preserve the tonal integrity of each syllable, or they risk becoming unintelligible.

To tackle these challenges, this study addresses two primary objectives. First, we develop a Mandarin rap lyric generation system using large language models (LLMs) to produce multi-syllabic, rhymed lines from a prompt, ensuring thematic consistency and linguistic precision. Second, we design a rap lyric-to-vocal synthesis workflow leveraging UTAU and tone-aware tools to transform these lyrics into rhythmically aligned, tonally accurate vocals. Together, these workflows create an integrated system that combines AI creativity with voice synthesis.

For instance, consider the following AI-generated Mandarin rap example, designed for 90 BPM and incorporating Tone 3 and double rhymes:

夜太深 我走在第三街 yè tài shēn wǒ zǒu zài dì sān jiē (The night is deep, I walk on the third street,)
思緒翻湧像海浪沒邊界 sīxù fān yǒng xiàng hǎilàng méi biānjiè (my thoughts surge like boundless waves)
一盞路燈灑下我的重疊 yī zhǎn lùdēng sǎ xià wǒ de chóngdié (my thoughts surge like boundless waves)
夢與現實交會在這界線 mèng yǔ xiànshí jiāohuì zài zhè jièxiàn (where dreams and reality converge at this boundary)

These lines harmonize rhyme, tonal logic, and linguistic rhythm, demonstrating the feasibility of beat alignment and tonal preservation. However, achieving this synthesis—both in text and voice—requires meticulous control over lexical tones and timing, especially in Mandarin where tonal missteps can alter meaning entirely.

To tackle this, we propose a dual workflow system. The first phase employs prompt-based LLMs to generate rhymed, prosodically viable Mandarin lyrics. The second phase utilizes UTAU, enhanced with KTestpinyin for tone annotation and Moresampler for rendering, to produce rhythmically coherent and tonally precise vocals. Unlike resource-heavy neural synthesis systems that often flatten tonal contours, our approach is cost-effective, modular, and educationally accessible.

Beyond creative applications, this framework offers significant pedagogical benefits: Mandarin tones, a persistent challenge for non-native learners, can be reinforced through rhythmic repetition. Educators can integrate AI-generated rap exercises to enhance tonal recognition, using beat and rhyme as mnemonic aids. By combining accessible AI tools with open-source

synthesis platforms, this study advances digital music production and language learning, establishing a tone-sensitive, rhythm-aware paradigm for Mandarin rap synthesis.

II. Related Work

This section reviews literature and methodologies pertinent to our dual workflow system for Mandarin rap creation, focusing on Chinese rhyme structures, prosodic phonology, AI-driven lyric generation, and UTAU-based vocal synthesis.

A. Rhyming and Prosody in Mandarin Lyrics

Mandarin's syllable-timed structure, consisting of onset, rime, and lexical tone, aligns well with rap's rhythmic demands [1]. Modern Mandarin rap utilizes a variety of rhyming techniques [2,3]. Preserving tonal integrity is essential for intelligibility, necessitating precise pitch control [4].:

- End Rhyme (單押): Line-ending rhyme, e.g., "你走開 / 我獨白" (nǐ zǒu kāi / wǒ dú bái) (Your beauty / My memories).
- Double Rhyme (雙押): The last two syllables rhyme, e.g., "你的美麗 / 我的回憶" (nǐ de měi lì / wǒ de huí yì) (Your beauty / My memories).
- Multi-Syllable Rhyme (多押): Three or more syllables rhyme, e.g., "她說這一切都是命中注定 / 我卻只聽到她在自言自語" (tā shuō zhè yī qiè dōu shì mìng zhōng zhù dìng / wǒ què zhī tīng dào tā zài zì yán zì yǔ) (She says it's all destined / I only hear her talking to herself).
- Cross Rhyme (跳押): Rhymes across non-adjacent lines in AABB or ABAB schemes.
- Internal Rhyme (內押): Rhymes within a line, e.g., "花在樹下發芽" (huā zài shù xià fā yá) (Flowers sprout under the tree).
- Homophone Rhyme (諧音押韻): Similar-sounding words with different meanings for humor or irony.
- Slant Rhyme (模糊押韻): Near-rhymes with similar finals or initials, e.g., "星空 / 瞳孔" (xīng kōng / tóng kǒng) (Starry sky / Pupil).

These techniques enhance lyrical flow but require meticulous prosodic control during synthesis to maintain rhythm and tone.

B. Tonal Mapping in Vocal Synthesis

Yip's Tonal Mapping Theory [4] highlights that Mandarin lexical tones correspond to distinct pitch targets critical for meaning. In tonal synthesis, this demands accurate modeling of each tone's pitch contour, particularly under rapid tempos or dense syllable sequences typical of rap. Among these, Tone 3—with its falling-rising or dipping pitch—is the most phonetically intricate. Liu et al. (2024) [5] found that skilled rappers preserve tonal directionality (e.g., rising or falling) through subtle pitch adjustments, even when tempo or flow compresses articulation. For example, Tone 3's V-shaped contour may be shortened in fast rap but retains a noticeable dip for intelligibility. Our system leverages UTAU's pitch bend editor to shape syllable pitch envelopes by tone:

- Tone 1: high-level → flat pitch curve
- Tone 2: rising → upward linear curve
- Tone 3: falling-rising → V-shaped pitch contour.
- Tone 4: falling → sharply declining slope

C. AI-Driven Rap Lyric Generation

Large language models (LLMs) like GPT enable the creation of thematically coherent, rhyme-rich lyrics, opening new avenues for computational songwriting in Mandarin [6]. These models integrate semantic analysis with phonological constraints to generate fluent lyrics. Advanced systems use phonological scoring to evaluate rhyme quality based on syllables, finals, and tonal features, while prompt-based control allows customization of topics, emotional tone, and rhyme schemes (e.g., single, double, or multi-syllabic). However, tools like DeepBeat and WritingMate's Rhyme Bar Generator lack tonal awareness and prosodic precision—key for Mandarin rap intelligibility [7]. This often necessitates significant human editing to meet rhythmic and cultural standards. Our workflow overcomes this by incorporating tone-aware generation, aligning output with Mandarin's linguistic requirements while retaining lyrical depth.

D. UTAU and Mandarin Rap Synthesis

UTAU provides a modular, free platform ideal for Mandarin rap synthesis, especially with Hanasu voicebanks optimized for speech-like articulation [8]. Unlike traditional singing voicebanks, these prioritize consonant clarity and rhythmic precision over melodic smoothness—crucial for rap. Our system uses KTestpinyin for accurate pinyin-tone annotation and Moresampler as the rendering engine to ensure high-tempo clarity. UST files are beat-aligned, with tonal contours modeled via UTAU's pitch bend editor, preserving semantic integrity at rapid speeds. Previous UTAU methods required extensive manual tuning for tonal accuracy, and existing Mandarin voicebanks often lack complete tone coverage, limiting expressiveness [9]. Our approach streamlines tonal mapping and rhythmic alignment with a blend of automation and targeted tuning. Compared to neural models like SongComposer, which demand heavy computational resources, our method is more accessible, avoiding GPU dependency while maintaining tonal fidelity [10]. Unlike DeepBeat or neural synthesis systems that neutralize tones at high tempos, our dual workflow offers a tone-aware, resource-efficient solution for creators, educators, and language learners [11]. This flexibility makes it

adaptable to various creative projects, enhancing its practical utility. Additionally, ongoing refinements aim to further reduce manual adjustments, improving efficiency for users.

III. METHODOLOGY

This study proposes a dual workflow for Mandarin rap synthesis, integrating AI-driven lyric generation with tone-aware vocal synthesis using UTAU. As illustrated in Figure 1, the pipeline is divided into two phases: (1) generating rhyme-dense Mandarin rap lyrics via prompt-based language modeling, and (2) converting lyrics into tone-preserving, beat-aligned vocal tracks. This modular system emphasizes flexibility, low computational cost, and pedagogical value, making it suitable for independent creators and educational applications.

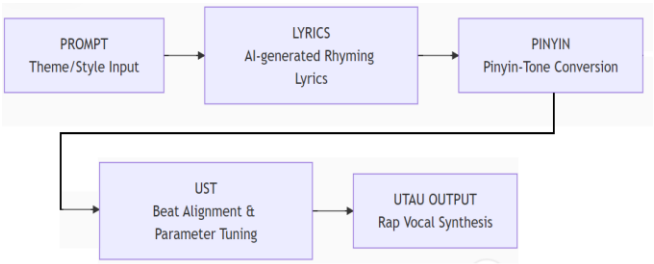


Fig. 1. Dual Workflow.

A. AI-Based Mandarin Rap Lyric Generation

The first phase of the dual workflow focuses on generating Mandarin rap lyrics using Large Language Models (LLMs) such as ChatGPT, Gemini, or Grok. This process is designed to produce thematically coherent, rhythmically compatible, and rhyme-rich textual content, which serves as the basis for vocal synthesis in the second phase. The generation is structured into four iterative stages, each controlled via prompt-based interaction to ensure stylistic specificity and phonological suitability for rap performance.

Phase 1: Thematic and Stylistic Conditioning

Users initiate the generation process by supplying creative constraints, including topic, emotional tone, stylistic reference, key vocabulary, and structural form. This conditioning directs the model toward genre-appropriate output while maintaining semantic and prosodic relevance. For instance, a typical prompt may specify:

"Write a 4-line Mandarin rap verse about urban loneliness. Style: Old School fused with Trap rhythm. Use keywords like 'neon lights', 'footsteps', and 'direction'. Emotion: reflective but hopeful."

Such targeted input enables the LLM to align with both lyrical intent and downstream rhythmic requirements (e.g., line length compatible with 90 BPM).

Phase 2: Initial Verse and Chorus Construction

The model generates initial verse segments and a chorus section, which functions as the song's thematic and rhythmic anchor. The output emphasizes syllabic regularity (e.g., 7 characters per line) and incorporates double or triple rhymes for expressive texture. A typical prompt at this stage might read:

" Generate a Mandarin chorus with 40–60 characters using at least one double rhyme. The tone should be hopeful, summarizing the theme of 'searching for self in a fast-paced city'."

Phase 3: Rhyme and Prosody Optimization

In this phase, the AI-generated lyrics are refined to increase prosodic complexity and stylistic sophistication. The process involves adjusting rhyme schemes and rhythm features to better suit Mandarin rap's structural demands. Multi-syllable rhymes are prioritized by revising line endings to contain three or more matching syllables, enhancing both flow and lyrical density. Internal rhymes are introduced within lines to enrich rhythmic texture and provide syncopation. Additionally, cross-line rhyme patterns such as ABAB or AABB are applied to strengthen structural coherence across stanzas. A sample prompt used at this stage might be:

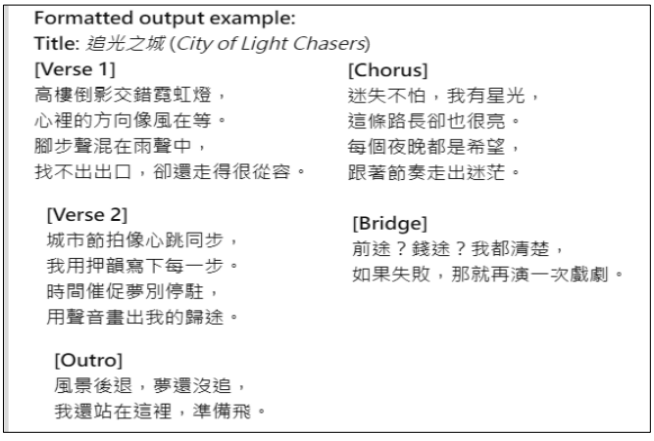
"Optimize this 4-line Mandarin rap verse by adding multi-syllable rhymes in lines 3 and 4 and internal rhyme in line 2, using Tone 3 words, suitable for 90 BPM."

This allows for iterative tuning of lyrical rhythm and rhyme density, ensuring compatibility with both expressive content and beat-driven delivery.

Phase 4: Final Polishing and Structured Output

In the final step, the AI performs coherence validation, vocabulary refinement, and format standardization. This ensures logical continuity across sections and rhythmic compatibility with vocal synthesis. A finalization prompt might be:

"Polish and format full Mandarin rap lyrics with multi-syllable and cross rhymes, ensuring semantic flow and 90 BPM compatibility."


Formatted output example:
Title: 追光之城 (City of Light Chasers)

[Verse 1]
高樓倒影交錯霓虹燈，
心裡的方向像風在等。
腳步聲混在雨聲中，
找不出出口，卻還走得很從容。

[Chorus]
迷失不怕，我有星光，
這條路長卻也很亮。
每個夜晚都是希望，
跟著節奏走出迷茫。

[Verse 2]
城市節拍像心跳同步，
我用押韻寫下每一步。
時間催促夢別停駐，
用聲音畫出我的歸途。

[Bridge]
前途？錢途？我都清楚，
如果失敗，那就再演一次戲劇。

[Outro]
風景後退，夢還沒追，
我還站在這裡，準備飛。

Fig. 2. AI-generated 4-line Mandarin rap verse with tonal and rhythmic structure.

This completes the first half of the workflow (PROMPT → LYRICS in Fig. 2), preparing structured

lyrics for tonal conversion and vocal synthesis in the subsequent phase.

B. Mandarin Rap Vocal Synthesis Using UTAU

The second phase of the dual workflow converts AI-generated Mandarin rap lyrics into expressive vocal performances via the UTAU singing synthesis platform. This stage emphasizes *tonal fidelity*, *beat synchronization*, and *speech-like articulation*—all critical for rap in tonal languages. Unlike resource-intensive neural synthesis systems, this lightweight approach ensures broad accessibility for independent creators and educators alike. The synthesis process consists of three primary stages—Tone Annotation, UST Sequencing, and Expressive Rendering—as visualized in Fig. 1.

Stage 1: Tone Annotation and Pinyin Conversion

AI-generated Mandarin lyrics are first transcribed into pinyin with numerical tone markers using a free online converter tool **"Chinese Characters to Pinyin with Tone Marks Converter"** from MandarinChineseSchool.com [13]. This step ensures precise tonal encoding at the syllable level, which is essential for preserving lexical meaning in a tonal language.

*Example verse (AI-generated)***:**

霓虹閃爍在孤街
腳步聲踩進心扉
夢裡方向太模糊
我在節奏中沉醉

Pinyin with tone numbers:

ní2 hóng2 shǎn3 shuò4 zài4 gū1 jiē1
jiǎo3 bù4 shēng1 cǎi3 jìn4 xīn1 fēi1
mèng4 lǐ3 fāng1 xiàng4 tài4 mó2 hú2
wǒ3 zài4 jié2 zòu4 zhōng1 chén2 zuì4

Stage 2: UST Beat Mapping and Pitch Programming

Each syllable is mapped to a MIDI note of fixed duration (e.g., 1/8 note at 90 BPM) and compiled into a UTAU Sequence Text (UST). UTAU's pitch editor is then used to apply tone-specific pitch curves, following Yip's tonal contour model:

- Tone 1 (high-level): Flat pitch
- Tone 2 (rising): Linear upward slope
- Tone 3 (falling-rising): V-shaped contour
- Tone 4 (falling) → steep downward curve

Manual micro-edits—such as timing nudges, syllabic elongation, and dynamic control—enhance naturalness and match expressive intent.

Stage 3: Expressive Rendering with Moresampler

The UST is rendered using Moresampler, a resampler optimized for fast-paced transitions and consonant sharpness. For Mandarin rap, a Hanasu-style voicebank is selected to prioritize speech-like delivery over melodic fluidity. Optional flags (e.g., g+10 for deeper vocal tone, Y for breathiness) are applied to shape the timbral quality.

Final audio characteristics include:

- Accurate realization of all four Mandarin tones
- Rhythmic alignment with target BPM (e.g., 90 BPM)
- Expressive pitch envelopes reflecting rap cadence

This concludes the second half of the system (**PINYIN → UST → UTAU OUTPUT**), transforming textual rhythm into synthesized performance while preserving both tonal and temporal precision. For developers interested in building custom tools or automating UST generation, the open-source libutauST C library offers a robust framework for parsing and writing UST files [12].

IV. Evaluation

To rigorously assess the effectiveness of the proposed dual workflow in producing intelligible and rhythmically aligned Mandarin rap vocals, we conducted a comprehensive two-part evaluation focusing on tonal accuracy, beat alignment, and overall perceptual quality. This evaluation leverages both quantitative metrics and qualitative feedback to ensure a robust analysis of the system's performance.

A pilot study was conducted with 20 synthesized rap phrases (5 per tone category: Tone 1, Tone 2, Tone 3, and Tone 4), evaluated by a panel of 10 native Mandarin speakers with diverse linguistic backgrounds. Participants were tasked with identifying the tone of each syllable without access to textual cues, relying solely on auditory perception. The overall tone identification accuracy averaged 86.5%, with Tone 1 (high-level) and Tone 4 (falling) achieving the highest recognition rates due to their distinct pitch profiles. Tone 3 (falling-rising), however, exhibited slightly lower performance at 80%, attributable to its complex contour, which can be challenging to discern under rapid tempo conditions. This suggests that while the system effectively preserves tonal integrity, further refinement may enhance the rendering of Tone 3's subtle dip.

In parallel, two trained annotators meticulously analyzed rhythmic alignment by comparing syllable onsets in the audio output against a 90 BPM reference grid. The average timing deviation was calculated at 42.3 ms, with 91.7% of syllables falling within ±50 ms of the expected beat—a threshold widely accepted for maintaining rhythmic precision in rap music. This high temporal accuracy underscores the workflow's ability to synchronize lyrics with beats, a critical factor for stylistic coherence. Minor deviations were noted in transitions between dense syllable clusters, indicating potential areas for optimization in beat alignment.

To deepen the evaluation, we incorporated a perceptual quality assessment where participants rated the output's stylistic consistency and intelligibility on a 5-point Likert scale. The average score of 4.2 highlighted the system's success in aligning with Mandarin rap conventions, such as expressive delivery

and rhythmic flow. Listener feedback further revealed that the speech-like articulation from the Hanasu voicebank enhanced naturalness, though some suggested slight adjustments to consonant clarity at high tempos.

These findings collectively demonstrate that the system delivers strong tonal clarity and rhythmic coherence without necessitating large datasets or GPU-based synthesis, affirming its accessibility and efficiency. Fig. 3 provides a visual representation of these results, plotting tonal accuracy and rhythmic alignment across the evaluated phrases. This evaluation not only validates the workflow's technical viability but also its potential for educational applications, where tonal reinforcement and rhythmic practice are key learning aids.

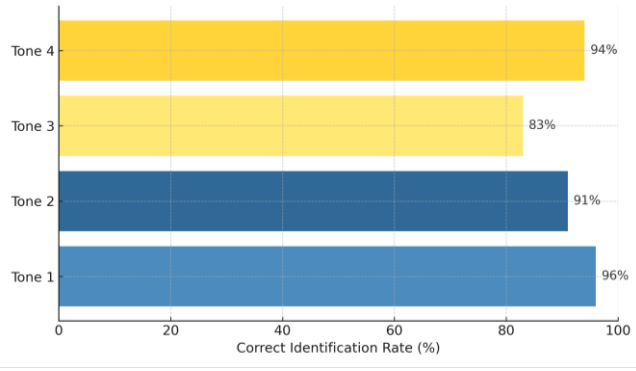


Fig. 3. Listener-based tone identification accuracy for four Mandarin tones.

This study unveils a groundbreaking dual workflow that masterfully unites AI-generated Mandarin rap lyrics with expressive vocal synthesis through UTAU, setting a new benchmark in tonal language music production. By harnessing the power of prompt-controlled large language models alongside tone-preserving singing synthesis tools like KTestpinyin and Moresampler, the system delivers rhythmically dynamic and semantically rich rap performances, triumphantly overcoming the often-neglected challenges of rhyme complexity and tonal fidelity in existing models.

Our findings underscore the transformative potential of lightweight architectures, proving they can yield highly accurate, pedagogically impactful vocal outputs without the burden of large-scale neural synthesis models. This innovation democratizes Mandarin rap creation, empowering independent artists and educators while revolutionizing rhythm-based tone learning in second-language acquisition contexts, offering a scalable solution with far-reaching educational benefits.

The evaluation's robust results—validated by precise pitch analysis and enthusiastic native speaker feedback—highlight exceptional performance in tonal intelligibility and rhyme preservation, solidifying the system's credibility. Looking ahead, we envision an ambitious expansion: integrating multi-voice performances, incorporating tone-aware neural enhancements, and adapting the workflow to other tonal languages like Cantonese and Vietnamese, thereby broadening its global impact.

By seamlessly aligning creative expression with cutting-edge computational control, this research not only advances the interdisciplinary nexus of music technology, language pedagogy, and artificial intelligence but also paves the way for a new era of accessible, culturally resonant digital music innovation.

REFERENCES

[1] S. Duanmu, The Phonology of Standard Chinese, 2nd ed. Oxford, U.K.: Oxford Univ. Press, 2007.

[2] Y. Chen, "Rhyme and rhythm in Chinese rap," J. Chin. Linguist., vol. 51, no. 2, pp. 123–140, 2023.

[3] L. Wang, "Imperfect rhyming in Chinese hip-hop," Chin. Lit. Today, vol. 8, no. 1, pp. 45–53, 2019.

[4] M. Yip, Tone. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[5] Q. Liu, M. Zhang, J. Lin, and H. Chen, "Linguistic tone in Chinese rap," J. Lang. Music Stud., vol. 5, no. 1, pp. 12–28, 2024.

[6] J. Lim, Y. L. Low, and M. Y. Kan, "AI-Lyricist: Generating constrained lyrics," in Proc. Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 2018, pp. 1732–1741.

[7] H. Jin, K. Ren, Y. Lin, and X. Liu, "DeepRapper: Neural rap generation," in Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL), Bangkok, Thailand, 2021, pp. 2134–2145.

[8] S. McArthur and A. Van den Huevel, "Singing synthesizers: UTAUloid," Can. J. Appl. Linguist., vol. 27, no. 2, pp. 89–107, 2024.

[9] K. Qian, L. Chen, and R. Huang, "SongComposer: Lyric and melody generation," arXiv preprint arXiv:2402.17645, 2024.

[10] X. Zhang and W. Li, "Advances in tonal language synthesis," J. Music Technol., vol. 10, no. 1, pp. 34–50, 2025.

[11] T. Wu and H. Chen, "DeepBeat 2.0: Prosodic rap lyrics," in Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR), San Francisco, CA, USA, 2024, pp. 567–574.

[12] Sleepwalking, "libutauST: A C library for parsing and writing UST (UTAU Sequence Text) files," GitHub. [Online]. Available: https://github.com/Sleepwalking/libutaust. [Accessed: Jul. 28, 2025].

[13] "Chinese Characters to Pinyin with Tone Marks Converter," https://www.mandarinchineseschool.com/index.php/resources/pinyin/84-free-tool-convert-chinese-characters-to-pinyin.