Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 2, 2025

Personalized Learning of Tibetan Academic Mandarin Pronunciation Integrating Mask-GCT Voice CloningTechnology

Zhenye Gan [1, *], Wenhao Wei [2]

[1] School of Educational Technology, Northwest Normal University, Lanzhou 730070, China
[2] College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

Abstract: This study proposes a personalized Mandarin pronunciation learning framework for Tibetan learners that integrates Mask-GCT voice-cloning technology as a back-end data-augmentation module. By leveraging deep neural networks, the voice-cloning component reconstructs key speaker characteristics—timbre, intonation, and prosody—from limited samples, generating high-fidelity, individualized speech data. These synthetic samples not only alleviate labeled-data scarcity but also introduce diverse pronunciation scenarios, particularly modeling the tonal, vowel, and consonant errors characteristic of Tibetan students' Mandarin production. Adjustable cloning parameters enable simulation of multiple error patterns and subtle phonetic variations, thereby enriching training data and enhancing the model's capacity to detect, adapt to, and correct a wide range of pronunciation deviations. Experimental results demonstrate that our approach significantly improves error recognition accuracy and model generalization compared to baseline systems lacking voice-cloning augmentation. The flexible, controllable synthesis process provides empirical support for targeted pronunciation remediation, offering a scalable methodology for assisting Tibetan learners in mastering academic Mandarin pronunciation.

Keywords: Deep Neural Networks; Pronunciation Error Detection; Mandarin Pronunciation.

1. Introduction

Accurate pronunciation is a cornerstone of second-language proficiency, directly influencing intelligibility, listener perception, and overall communicative success. For native Tibetan learners of Mandarin, achieving accurate production of Mandarin's tonal contours, vowel (rhyme) qualities, and consonant articulations is especially challenging due to interference from the phonological properties of the Tibetan language. Existing computer-assisted pronunciation training (CAPT) systems typically depend on large, manually labeled speech corpora. However, collecting and annotating extensive databases of Tibetan-accented Mandarin is both time-consuming and resource-intensive, often resulting in models that lack sufficient exposure to the full spectrum of learner-specific pronunciation deviations.

To overcome the limitations imposed by scarce labeled data, we propose a novel framework that incorporates Mask-GCT voice-cloning technology as a back-end data-augmentation module for personalized pronunciation training. Voice cloning employs advanced deep neural networks to reconstruct speaker-specific traits—such as timbre, intonation, and prosody—from only a handful of audio examples. By synthesizing high-fidelity, individualized speech that embeds targeted pronunciation errors, our system can generate a rich variety of realistic training scenarios without the need for extensive manual recording or annotation.

Our approach specifically targets the characteristic pronunciation error patterns of Tibetan learners, including tonal misclassification, vowel quality shifts, and consonantal substitutions. The voice-cloning module's adjustable parameters enable controlled simulation of these error types at varying intensity levels, thus providing the downstream error-detection and feedback models with comprehensive, diverse examples of learner productions. We hypothesize that training on this augmented dataset will significantly improve the system's ability to detect, classify, and correct pronunciation deviations, leading to more effective, personalized feedback for Tibetan students.

2. Research Status of Voice Cloning

Voice cloning is a subtask of speech synthesis; therefore, to understand the current state of voice-cloning research, one must first review the development of speech synthesis. Speech synthesis (Text-to-Speech, TTS), also called text-to-speech conversion, is a technology that transforms arbitrary input text into matching speech. Early voice-cloning methods relied primarily on two approaches: concatenative synthesis and parametric synthesis.

(1) Concatenative synthesis

Concatenative synthesis generates target speech by stitching together basic units—such as phonemes or syllables—selected from a prerecorded speech database. Its core challenges lie in unit selection and waveform smoothing. Hunt and Black proposed a unit-selection method based on a large-scale speech database [1]; by carefully choosing prerecorded speech units, their concatenative TTS system achieved high naturalness. Dutoit systematically described the fundamental principles and key techniques of TTS systems—including front-end text analysis, phoneme conversion, and waveform generation—in his seminal work [2]. For example, the PSOLA algorithm (Pitch-Synchronous Overlap and Add) modifies speech by adjusting pitch and duration, but its naturalness is limited by spectral distortion [3]. Addressing the characteristics of Chinese, Tao Jianhua et al. introduced an improved segmental pitch-adjustment strategy that effectively reduces distortion in Mandarin tone synthesis. However, these methods require constructing large speech corpora—for instance, the English ARCTIC corpus—

and Chinese TTS typically relies on Tsinghua University's TH-CoSS dataset [4, 5], which comprises 50 hours of standard Mandarin covering all four tone combinations, thereby enabling high-quality Chinese speech synthesis.

(2) Parametric synthesis

Parametric synthesis uses statistical models to generate acoustic parameters. A typical example is the hidden Markov model (HMM) [6]. HMM-based TTS models phonetic acoustic features—such as Mel-frequency cepstral coefficients (MFCCs) and fundamental frequency (F0)—via state-transition probabilities, then passes the generated parameters through a vocoder to produce a waveform [7]. To improve naturalness for Mandarin, researchers proposed a hierarchical HMM architecture that models tones separately from phonemes [8], significantly enhancing synthesis quality. Nevertheless, the Markov assumption in HMMs leads to over-smoothed spectra and a mechanical "buzz," yielding a mean opinion score (MOS) of only around 3.0 out of 5.0.

(3) Traditional TTS

Traditional TTS methods have three main limitations:

They depend heavily on manual annotation and feature engineering—HMM training, for example, requires precise boundary labels for thousands of utterances, and Chinese segmentation complexity further increases annotation difficulty [9];

Concatenative synthesis naturalness is fundamentally limited—its MOS ceiling is about 3.8, and it struggles to simulate emotional and stress variations in real speech [10];

Cross-lingual adaptation demands manual rule redesign—for instance, prosodic rules for Chinese sentence-final particles like "a" or "ne" must be hand-crafted. Nonetheless, these traditional techniques laid the groundwork for modern TTS: PSOLA's real-time optimization inspired lightweight neural-network designs; HMM state-modeling influenced duration-prediction models such as the attention mechanism in Tacotron; and the WORLD vocoder remains a benchmark for evaluating neural vocoders. Zhang Wei et al. developed a WaveNet-based Mandarin vocoder that, while real-time, raised MOS to 4.08.

(4) Deep learning TTS

Deep learning has revolutionized speech synthesis. In 2016, DeepMind introduced WaveNet, a generative deep-neural-network framework that models raw audio waveforms' temporal dependencies, dramatically improving naturalness and fluidity compared to concatenative or parametric methods. Tacotron followed, achieving end-to-end TTS by mapping text directly to spectrograms and eliminating multiple traditional pipeline stages. As naturalness rose, researchers sought more efficient architectures: Transformer-based self-attention enabled global context modeling and parallel processing, boosting both speed and quality; diffusion models like DiffWave generate high-quality audio via a reverse-diffusion process, overcoming the slow generation of autoregressive models while preserving naturalness [11]. In 2021, VITS combined variational autoencoders with adversarial training for fully end-to-end waveform generation; its joint-training strategy simplified the modular design and yielded significant gains in naturalness and smoothness [12]. In the large-model era, Meta's Vall E leveraged massive speech corpora and meta-learning to clone a target speaker's timbre and prosody from just three seconds of audio, marking a breakthrough in zero-shot voice cloning [13]. This advance demonstrates deep learning's potential for few-shot learning and enables flexible, efficient personalized, virtual-assistant, and cross-lingual TTS applications. Mask-GCT (Masked Generative Codec Transformer) is a fully non-autoregressive TTS model that requires no explicit alignment between text and speech and no phoneme-level duration prediction; it outperforms state-of-the-art zero-shot TTS systems in quality, similarity, and intelligibility [14]. Overall, from WaveNet to Tacotron, Transformer, DiffWave, VITS, Vall E, and MASK-GCT, these developments showcase deep learning's tremendous advances in naturalness, generation efficiency, and end-to-end training—and point toward a future in which TTS increasingly relies on large-scale data and models to achieve more realistic, finely detailed, and highly personalized speech synthesis.

3. The Role of Voice Cloning in Second-Language Acquisition

(1) Emotional Engagement and Motivational Arousal

When learners perceive speech that closely resembles their own voice, they are more receptive to external suggestions and feedback (MacIntyre et al.) [15]. In traditional second-language pronunciation training, learners must imitate an idealized "correct" pronunciation—whether produced by a robotic synthesizer or by an instructor—that often differs markedly from their habitual voice quality. This discrepancy can create a conflict in the learner's self-identity, leading to a sense of detachment that undermines their motivation to learn.

Voice cloning, however, generates personalized speech with acoustic characteristics that mirror the learner's own voice. When second-language learners hear synthesized samples that sound highly similar to themselves, they experience a sense of familiarity and emotional resonance. This personalized resonance reduces the resistance and reluctance typically encountered in conventional pronunciation training, lowers psychological barriers to adopting the correct pronunciation, and boosts learners' confidence. As a result, learners become more willing to accept corrective feedback and to adjust their pronunciation accordingly.

(2) The Psychological Proximity Effect of Perceived Similarity

Montoya et al. proposed that even when actual pronunciations differ, voice-cloning technology can psychologically establish a proximity effect through perceived similarity [16]. Tibetan and Mandarin differ significantly in phonological structure—particularly in consonant and syllable structure, as well as rhyme systems. Mandarin's consonant inventory is relatively simple, whereas Tibetan features far more complex consonant clusters and aspiration distinctions. In Tibetan, multiple consonants are often combined in a single syllable, with careful attention to aspiration; its variable syllable structures also lead to different pronunciation habits. These differences impose additional difficulty on learners attempting to imitate target Mandarin pronunciations.

By using voice-cloning technology to map a Tibetan learner's original vocal characteristics onto target speech, one can effectively mask the phonological disparities between Tibetan and Mandarin. When learners hear feedback that closely resembles their own voice, they unconsciously experience a sense of familiarity and closeness, thereby creating psychological proximity. This proximity mitigates the cognitive conflict arising from cross-linguistic

phonological differences and makes Tibetan speakers more receptive to pronunciation correction suggestions.

In summary, personalized pronunciation samples generated by voice cloning—by eliciting perceived similarity and emotional identification—reduce learners' sense of alienation and build psychological proximity. Even though Tibetan and Mandarin differ markedly in actual pronunciation, this novel method, which combines deep-learning and psychological principles, achieves highly personalized phonetic similarity, increases Tibetan learners' acceptance of corrective feedback, and stimulates their motivation to learn.

4. Pretrained MASK-GCT Model

Traditional TTS systems are typically divided into two categories: autoregressive and non-autoregressive. Autoregressive models implicitly model speech duration, yielding coherent generation but suffering from limited robustness and duration control; non-autoregressive models rely on explicit text-to-speech alignment and prediction of linguistic unit (e.g., phoneme) durations during training, which can compromise the naturalness of the generated speech. To address these issues, the Mask-GCT model introduces a novel design that completely eliminates the need for explicit alignment between text and speech and avoids phoneme-level duration prediction. Its model architecture is shown in Figure 1.

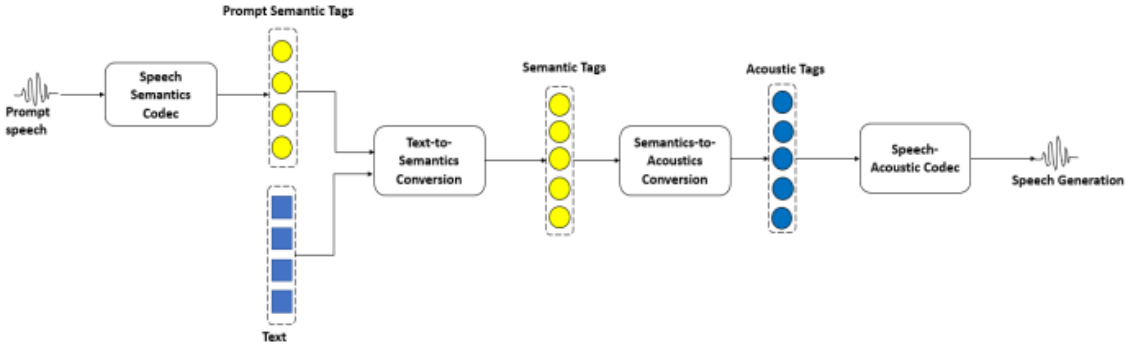

Figure 1. MASK-GCT Model Diagram

The MaskGCT model adopts a two-stage architecture, consisting of:

Text-to-Semantic (T2S) Stage

In this stage, the model predicts semantic tokens extracted from a speech self-supervised learning (SSL) model based on the input text. This step transforms discrete textual information into representations rich in semantic content, laying the foundation for subsequent acoustic modeling.

Semantic-to-Acoustic (S2A) Stage

Based on the semantic tokens obtained in the first stage, the S2A model further predicts corresponding acoustic tokens, which contain specific audio details such as timbre and prosody. This process realizes the transformation from abstract semantic information to concrete acoustic representations.

Additionally, MaskGCT follows a masked prediction learning paradigm. During training, the model learns to predict masked semantic or acoustic tokens given certain conditions and prompts, thereby enhancing its ability to capture critical information. During inference, the model is able to generate token sequences of preset lengths in parallel, significantly improving generation efficiency.

This study utilizes the pretrained MaskGCT model in combination with a self-constructed dataset of Tibetan-accented Mandarin pronunciation errors to further train and evaluate the model, aiming to validate its effectiveness in low-resource personalized speech synthesis, particularly in enhancing pronunciation correction for Tibetan students learning Mandarin.

5. Experimental Study

(1) Dataset Construction

The corpus was recorded using a single carbon microphone and smartphones under silent office conditions, with a sampling rate of 16000 Hz and a sample size of 16 bits. Sentences (text prompts in recordings) were selected from daily expressions to ensure coverage of all syllables. The recording subjects were Tibetan students with significant accent variations. To ensure accurate annotation, we invited graduate students majoring in phonetics for cross-annotation. In cases of inconsistency, final decisions were made by phonetics experts. This corpus was used as the test corpus in this article. The design of this corpus is detailed in Table 1.

Table 1. Intermediate Corpus

TEXT	1200 UTTERANCE
SPEAKER	5 FEMALES AND MALES
NUMBER UTTERANCE	2368
NUMBER OF PHONEMES	46654

(2) Speech Evaluation Metrics

In the development and optimization of speech synthesis systems, accurate evaluation metrics are crucial for assessing system performance and guiding model improvements.

Speech synthesis evaluation metrics aim to quantify the output quality of a speech system from various perspectives. Commonly used objective metrics such as PESQ (Perceptual Evaluation of Speech Quality), STOI (Short-Time Objective Intelligibility), LSD (Log-Spectral Distance), and MCD (Mel-Cepstral Distortion) provide fast and quantitative evaluation results. However, these objective metrics often fail to fully capture the human auditory system's perception of naturalness, emotional expression, and overall listening experience.

Therefore, in the context of speech synthesis systems—especially in research focused on Mandarin pronunciation training for Tibetan students—subjective evaluation becomes particularly important.

The MOS (Mean Opinion Score) is a subjective evaluation method based on human auditory perception, and its formula is shown in Equation 1-1.

$$MOS = \frac{\sum_{j=1}^{M}\left(\frac{\sum_{i=1}^{N} S_{ij}}{N}\right)}{M} \tag{1-1}$$

(3) Experimental Results

An example from the test set is presented to illustrate the experimental process. One speaker mispronounced the sentence “xian zai, he tao de jia ge xia die le” (“Now, the price of walnuts has dropped”) as “xian zai he tao de jia ge xia tie le” by substituting the phoneme /d/ in “die” with /t/, resulting in “tie.” Using the MASK-GCT model, we performed voice cloning to synthesize speech that retains the speaker’s unique vocal characteristics. The spectrogram of the mispronounced “tie” is shown in Figure 2.

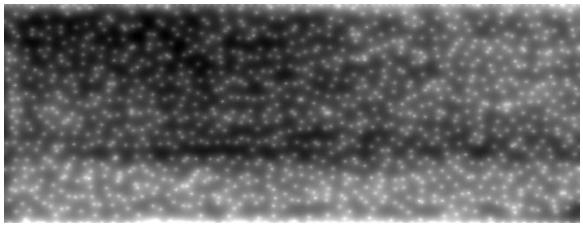

Figure 2. Spectrogram of the Mispronounced "tie"

The spectrogram of the synthesized correct pronunciation "die" is shown in Figure 3.

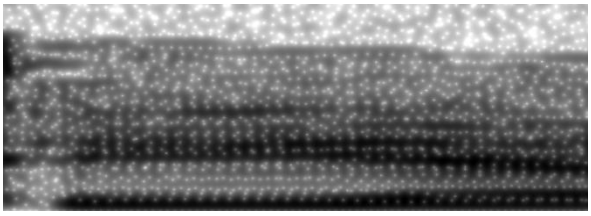

Figure 3. Spectrogram of the Correct Pronunciation "die"

Based on the above analysis, we can conclude that the student mispronounced the plosive /d/ as the plosive /t/. This error can be attributed to the lack of an aspirated-unaspirated contrast in the Ü-Tsang dialect of Tibetan, which leads the learner to interpret Mandarin /d/ as a weakly aspirated voiceless \[t]. Both /d/ and /t/ are alveolar sounds, with energy concentrated in the high-frequency range between 4000–8000 Hz. In the spectrogram of the mispronounced sound, there is a feather-like diffusion of energy and fragmented bright spots above 4000 Hz. Due to the aspiration in the learner’s /t/ pronunciation, energy is dispersed, and the overall brightness is low, indicating reduced vocal energy, possibly because the speaker was subconsciously aware of the error.

In contrast, the synthesized speech shows compact energy distribution. The energy is uniformly colored in the frequency domain, forming concentrated and evenly distributed short bursts in the spectrogram. This indicates that the synthesized audio exhibits a more focused energy distribution during the stop burst phase, and its clarity is significantly improved compared to the original pronunciation.

The MOS (Mean Opinion Score) results rated by ten students majoring in phonetics are presented in Table 2 below.

Table 2. MOS Score Table

Evaluator	MOS Score
Student 1	4.62
Student 2	3.91
Student 3	4.37
Student 4	4.05
Student 5	4.78
Student 6	3.85
Student 7	4.52
Student 8	4.19
Student 9	4.29
Student 10	4.45

The bar chart it generates provides a more intuitive display of the quality of the synthesis results, with the average value shown as a line on the chart. The resulting bar chart is shown in Figure 4.

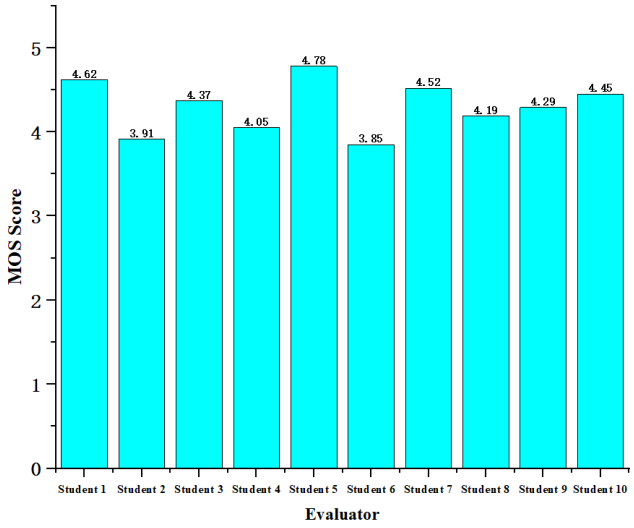

Figure 4. Intuitive Display of MOS Scores

6. Conclusion

This paper first provides a detailed introduction to the speech cloning model proposed in this paper—the MASK-GCT model, including the model's overall structure, parameter settings, and key training strategies. Subsequently, speech cloning experiments were conducted based on this model. By comparing multiple groups of speech sample generation results, the system's performance in terms of naturalness, clarity, and emotional expression was comprehensively evaluated. The experimental results show that the speech cloning system based on the MASK-GCT model achieved an average MOS score of 4.303 in subjective evaluation, demonstrating the model's advantages in speech restoration and cloning quality.

References

[1] Hunt A J, Black A W. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database [C]// Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP). Philadelphia, PA, 1996.

[2] Dutoit T. An Introduction to Text-to-Speech Synthesis [M]. Springer, 1997.

[3] Moulines E, Charpentier F. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones [J]. Speech Communication, 1990, 9(5-6): 453-467.

[4] Kominek J, Black A W. The CMU Arctic Speech Databases [C]// Fifth ISCA Workshop on Speech Synthesis. 2004.

[5] CAI Lian-hong, CUI Dan-dan, CAI Rui. TH-CoSS,a Mandarin Speech Corpus for TTS [J], Journal of Chinese information processing, 2007: 96-101.

[6] JING Xiao-yang, LUO Fei, WANG Ya-qi. Overview of the Chinese Voice Synthesis Technique [J], Computer Science, 2012

[7] Tokuda K, Yoshimura T, Masuko T, et al. Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis [C]// 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2000, 3: 1315-1318.

[8] Zen H, Tokuda K, Black A W. Statistical Parametric Speech Synthesis [J]. Speech Communication, 2009, 51(11): 1039-1064.

[9] Taylor P. Text-to-Speech Synthesis [M]. Cambridge University Press, 2009: 147-162.

[10] Black A W, Taylor P. Automatically Clustering Similar Units for Unit Selection in Speech Synthesis [C]// Eurospeech. 1997.

[11] Chen W, Zhang Y, Lei J, et al. DiffWave: A Versatile Diffusion Model for Audio Synthesis [C]// Proceedings of Interspeech 2020. Shanghai, China: ISCA, 2020: 3565-3569.

[12] Kim Y, Kong J, Son J, et al. VITS: Conditional Variational Inference with Adversarial Learning for End-to-End Text-to-Speech [EB/OL]. (2021-06-11) [2024-07-20]. https://arxiv.org/abs/2106.06103.

[13] Chen N, Zhang Y, Zheng H, et al. VALL-E: A Text-to-Speech System for Zero-Shot Voice Cloning [EB/OL]. (2022-12-15) [2024-07-20]. https://arxiv.org/abs/2212.08025.

[14] Yuancheng W, Haoyue Z, Liwei L, Ruihong Z, Haotian G, Jiachen Z, Qiang Z, Xueyao Z, Shunsi Z, Zhizheng W, et al. MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer [J]. CoRR, 2024, abs/2409.00750.

[15] MacIntyre P, Gregersen T. Emotions That Facilitate Language Learning: The Positive-Broadening Power of the Imagination [J]. Studies in Second Language Learning and Teaching, 2012, 2(2): 193-193.

[16] Montoya R M, Horton R S, Kirchner J. Is Actual Similarity Necessary for Attraction? A Meta-Analysis of Actual and Perceived Similarity [J]. Journal of Social and Personal Relationships, 2008, 25(6): 889-922.