Review

Machine Learning for Mandarin Tone Recognition: A Systematic Review with Applications to Neurotypical and Clinical Populations

Aaron Zou , Xinran Han , Yena Koo , Xu Yan , Yang Zhang *

Posted Date: 31 October 2025

doi: 10.20944/preprints202510.2478.v1

Keywords: Mandarin tone recognition; machine learning; deep learning; computer assisted language learning; speech-language therapy

Review

Machine Learning for Mandarin Tone Recognition: A Systematic Review with Applications to Neurotypical and Clinical Populations

Aaron Zou [1], Xinran Han [2], Yena Koo [3], Xu Yan [4] and Yang Zhang [2,*]

[1] Wayzata High School, Plymouth, MN
[2] Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN
[3] Department of Information and Decision Sciences, University of Minnesota, Minneapolis, MN
[4] Department of Computer Science, University of California, Los Angeles, CA
* Correspondence: zhang470@umn.edu

Abstract

Purpose: Machine learning (ML) is increasingly applied in natural language processing, speech evaluation, and computer-assisted language learning (CALL). This systematic review synthesizes ML approaches to Mandarin tone recognition to assess best-performing models and discuss challenges and opportunities for translational and clinical applications. **Method:** Following PRISMA guidelines, we searched five databases (IEEE Xplore, PubMed, Google Scholar, Web of Science, and Scopus) and analyzed 61 articles for model architecture, input features, datasets, evaluation, and validation metrics. **Results:** Deep learning models outperform traditional approaches in Mandarin tone classification (mean accuracy 88.8% vs. 83.1%). Convolutional Neural Networks (CNNs) achieve up to 99.16% accuracy for isolated syllables, while Bidirectional Long Short-Term Memory (BiLSTM or BLSTM) and attention-based models improve continuous speech recognition by capturing temporal dependencies with 7.03% error rate. Performance is affected by Tone 3 variability, neutral tones, and challenging conditions like background noise and disordered speech. Emerging areas of application include CALL and brain-computer interfaces with growing emphasis on robustness and inclusivity. **Conclusion:** While deep learning models represents the state of the art, several gaps limit practical deployment, including the lack of diverse datasets, weak prosody and dialect modeling, and insufficient validation rigor. Multimodal special-purpose corpora are needed to develop efficient lightweight models to improve real-time assessment and feedback for user-centered applications with strong pedagogical and clinical impact.

Keywords: Mandarin tone recognition; machine learning; deep learning; computer assisted language learning; speech-language therapy

Introduction

Speech recognition technologies have advanced rapidly over the past decades, from consumer-level tools such as Siri, Alexa, and ChatGPT's voice interfaces to specialized, user-centered applications in clinical and healthcare fields (Johnson et al., 2014; Shour et al., 2025; M. Zhang et al., 2024). Yet most developments have focused on non-tonal languages like English, overlooking lexical tones which are central to tonal languages like Mandarin Chinese with over 1.3 billion speakers. In Mandarin, pitch contours determine word meaning as a single syllable like "ma" can represent mother (mā, Tone 1, high flat tone), hemp (má, Tone 2, rising tone), horse (mǎ, Tone 3, dip-rising tone) or scold (mà, Tone 4, falling tone). Mispronounced or misperceived tones can alter meaning and communicative intent.

In phonological theory, lexical tones and segmental features like consonants and vowels are represented on separate tiers, and tone-bearing units (e.g., syllables) participate in systematic and

 doi:10.20944/preprints202510.2478.v1

predictable phonological processes such as tone sandhi (e.g., Tone 3 changing to Tone 2 before another Tone 3) (Duanmu, 2007; J. Zhang, 2010). However, real-world speech introduces considerable variability. While tones in isolated syllables tend to show stable pitch contours, these contours become highly dynamic in continuous speech due to coarticulation, phonological context, emotional prosody, dialectal differences, and speaking rate (S. Chen et al., 2022; P. Tang et al., 2017; Xiao & Liu, 2024; Y. Xu, 2019; J. Zhang & Liu, 2011). For example, a rising contour may signal either Tone 2 or emotional excitement whereas a sharp fall in pitch can indicate Tone 4 or anger/frustration. In some cases, coarticulatory effects can blur tone distinctions or shift pitch contour into a different tonal category (J. Zhou et al., 2004). Moreover, speakers from different dialect regions (e.g., Beijing, Taiwan) may produce the same tone with noticeably different pitch ranges or contour shapes (Fon & Chuang, 2024).

Cognitively, tone processing is vulnerable under load. For example, dual-task conditions degrade categorical perception of tones with shifted tone boundaries and reduced discrimination (Feng et al., 2021). This matters because L2 learners and some clinical populations may constantly face cognitive load issues in real-world settings. For L2 learners of Mandarin Chinese, their perception and production improvements are tightly linked to how they adapted to the critical perceptual cue of pitch contour (Leung et al., 2025). Tonal errors are among the most frequent and socially impactful barriers to fluency; unlike segmental mistakes, tonal mispronunciations often render speech not merely accented but unintelligible (M. Cao et al., 2024; Pelzl et al., 2021). Machine learning (ML)–based tone recognition thus offers a promising route to objective, real-time feedback in computer-assisted language learning (CALL), potentially accelerating acquisition and preventing error fossilization (B. Cheng et al., 2025).

In clinical contexts, the stakes are even higher. Children with cochlear implants often struggle to perceive tonal contrasts due to degraded spectral resolution (Hong et al., 2019; H. Zhang, Ma, et al., 2022; H. Zhang et al., 2023). Adults with post-stroke aphasia may produce tones inconsistently or omit them entirely (Gandour et al., 1988). Mandarin speakers with Alzheimer's disease show significant difficulties repeating words and non-words with notable impairments in producing correct tones, alongside additional effects of age and linguistic factors (Y.-T. Lin & Lai, 2024). Mandarin-speaking individuals with Parkinson's disease show impaired production of contrastive stress, especially in spontaneous speech, while their perception of these prosodic cues remained largely intact (X. Chen & Sidtis, 2024). Automated tone recognition could transform assessment, therapy, and assistive communication. For example, a feed-forward artificial neural network (ANN) was implemented to classify Mandarin tones produced by 61 normal-hearing children with an accuracy rate of 85%, outperforming adult listener performance (L. Xu et al., 2007). Zhou and Xu (2008) further developed and validated acoustic, ANN, and perceptual methods for assessing Mandarin tone production in children with cochlear implants, demonstrating that these measures were highly correlated and effective in evaluating tonal accuracy compared to normal-hearing peers. But few models have been validated on disordered speech, and even fewer integrate into real clinical workflows. Compounding these challenges is the gap between laboratory and real-world conditions. ML models trained and evaluated with speech recorded in controlled, studio-like environments often perform poorly when applied to more variable contexts that include background noise, emotional prosody, or dialectal diversity (M.-C. Lee et al., 2022; Q. Li et al., 2024; P. Zhang et al., 2023). Even for neurotypical native speakers, tone recognition accuracy drops markedly in these challenging listening conditions (H.-S. Chang et al., 2023; C.-Y. Lee et al., 2010; M. Liu & Chen, 2020; X. Wang & Xu, 2020).

Machine learning, especially deep learning, offers potential solutions (M. Chen et al., 2014; Song & Deng, 2025). The core task typically involves training computer models to recognize the distinct pitch contours by analyzing the fundamental frequency (F_0) and other acoustic features. After normalizing for speaker-related differences, these features are used to train classifiers on labeled examples of each tone. The trained model can then categorize new speech samples by mapping their acoustic patterns to the appropriate tone category. Different architectures have been explored for this

 doi:10.20944/preprints202510.2478.v1

task. Convolutional neural networks (CNNs) learn robust tone representations directly from mel-spectrograms and perform well on isolated syllables. Recurrent architectures such as bidirectional long short-term memory networks (BiLSTMs) and Transformers can model temporal dynamics of pitch contours in continuous speech. Attention mechanisms help models focus on the most informative parts of a signal. More recently, self-supervised approaches have been developed to encode tone information from unlabeled speech, allowing for more flexible and scalable recognition systems.

Despite these technical advances, research in this area remains fragmented. There has been no systematic comparison of which model types perform best under specific conditions or for particular speaker populations. A major obstacle is the lack of standardization. Existing datasets differ widely in recording conditions and speech style, and evaluation metrics are inconsistent. Crucially, few datasets include speech from L2 learners or clinical populations, or account for common variables like tone sandhi, emotional prosody, or dialectal variation.

To address these gaps, this systematic review aims to provide a comprehensive, evidence-based overview of machine learning methods for Mandarin tone recognition, with a focus on clinical applications, L2 learning, and real-world deployment. We compare traditional statistical models such as Hidden Markov Models (HMMs) (J. Cheng et al., 2000; M. Cheng et al., 2003; W.-J. Yang et al., 1988) and Support Vector Machines (SVMs) (Y. Chen & Xu, 2021) with modern deep learning architectures (CNNs, BiLSTMs, Transformers) across both isolated syllables and continuous speech. The review identifies best practices in feature engineering, dataset selection, and validation strategies that enhance model generalizability and robustness. Beyond accuracy, we assess models in terms of clinical and pedagogical utility, considering resilience to noise, emotional prosody, dialectal variation, as well as individual differences. We also highlight critical gaps, including overreliance on clean studio data, underrepresentation of disordered or learner speech, insufficient modelling of tone sandhi and dialectal variation, and the ongoing disconnect between technical benchmarks and meaningful communication or learning outcomes. The review concludes with actionable recommendations for researchers, clinicians, and developers to guide model selection, dataset design and practical application.

Methods

Search Strategy

This systematic review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (PRISMA) (Haddaway et al., 2022; Page et al., 2021). We searched five databases (IEEE Xplore, PubMed, Google Scholar, Web of Science, Scopus) using the following keywords: “Chinese tones”, “Mandarin tones”, “machine learning”, “deep learning”, “neural network”, “algorithm”, “classification”, “recognition.” The boolean operators used were: (“Chinese tones” OR “Mandarin tones”) AND (“machine learning” OR “deep learning” OR “neural network” OR “algorithm”) AND (classification OR recognition OR speech processing). These boolean operators were chosen to ensure that any ML articles related to the Chinese language or standard Mandarin were covered without excluding studies that examine Chinese dialects. The second section of the operators were specifically chosen in order to query relevant articles on machine learning methods which may use a variety of terminology. Finally, the last part of the search query was designed to focus specifically on tone recognition rather than studies that merely identify the presence of tones.

Our search yielded 245 articles in total. After removing duplicates, 158 articles remained. Two authors (AZ and YK) screened the remaining articles with the following inclusion criteria:

a) The study must focus on applying or improving machine learning techniques for Mandarin tone classification, including work on accented Mandarin due to dialect influence, L2 learner accents, dialectal tone identification to support Mandarin recognition, or clinical applications. The training and test materials could be mono-syllabic, multi-syllabic, or continuous speech.

 doi:10.20944/preprints202510.2478.v1

b) The study must be published in peer-reviewed journals or conference proceedings. Reviews were not included.
c) The study must report empirical results or datasets and provide performance or evaluation metrics to allow for comparative analysis.
d) The study must use a sufficiently large dataset appropriate for the classification task or otherwise provide justification.
e) The article must be written in English, publicly accessible, and published before March 2025.

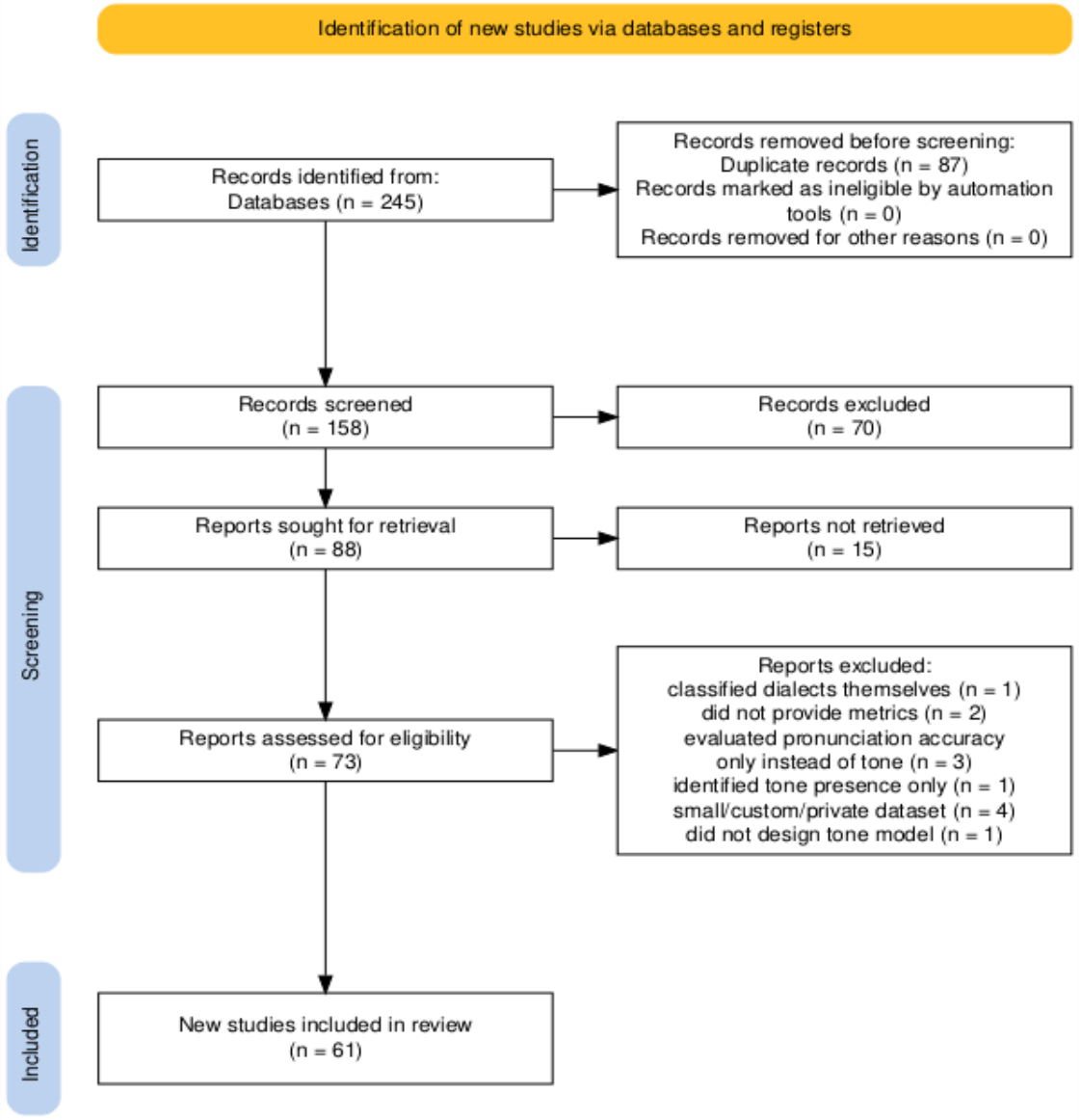

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for identifying 61 studies included in this review.

After applying the inclusion criteria and verifying accessibility, 73 articles remained. Studies were excluded if they (a) focused on dialect classification rather than tone recognition, (b) did not report performance metrics or cross-validation procedures, (c) examined pronunciation accuracy or production comparisons without tone classification, (d) primarily analyzed brain data such as EEG instead of acoustic measures for modeling tones, (e) relied on small, private, or non–peer-reviewed datasets, or (f) addressed tone recognition only as part of broader recognition systems without explicitly modeling tones. Based on these criteria, twelve articles were excluded, resulting in a final pool of 61 studies included in the systematic review (Figure 1).

 doi:10.20944/preprints202510.2478.v1

Quality Assessment

Included studies were assessed for experimental rigor and quality. Beyond just a sufficient dataset size, study datasets were evaluated for multiple speakers and dataset balance. In addition, studies were made sure to have proper, standard validation methods that could be compared to other studies. When an experiment was conducted, such as determining which model combination produced the best results, reproducibility of the experiment was assessed, as well as dataset size and proper validation methods. The Cochrane Risk of Bias tool was used to ensure randomized experiments within studies, if present, were of high quality. Two reviewers rated the reports independently, and disagreements were resolved by discussion to reach consensus.

Data Extraction

The following data were extracted from each included study if present: (a) study characteristics (e.g., first author, year of publication, source of publication), (b) the type of machine learning model (traditional or deep), (c) specific architecture of the model (e.g., HMM, CNN, joint-model), (d) database characteristics used for training and testing the proposed model as well as training characteristics (e.g., isolated vs continuous, database size, train/test splits, number of speakers, speaker population), (e) results of interest (e.g., validation methods, evaluation metrics, average accuracies, relative comparisons to other models).

Some studies provided accuracies for multiple models to provide empirical comparisons between their proposed approach and baseline systems in experiments. When this occurred, accuracies were calculated using the results of the intended recognition task with the proposed model reported by the authors. Studies that did not provide absolute performance metrics (such as only providing relative error reduction rates to a baseline), were not included in the computation of our aggregate statistics. When confusion matrices were provided, the mean across the recognition accuracies of the four tones was calculated, or the accuracy of the speech recognition system if the study focused on incorporating tone into ASR.

Results

Overview of Model Distribution

A wide range of machine learning techniques have been applied to Mandarin tone classification, ranging from traditional statistical models to deep neural architectures and hybrid systems. Among the 61 studies included in this review, 26 used traditional models such as Hidden Markov Models (HMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Gaussian Mixture Models (GMM), Decision Trees, Random Forests, and shallow Multi-Layer Perceptrons (MLPs). In contrast, 35 studies employed deep learning models, including CNN, BiLSTM, Deep Neural Networks (DNN), Transformers, Autoencoders, and End-to-End architectures. This distribution illustrates a clear shift toward deep learning methods, driven by their ability to automatically learn high-level tone features and model long-range temporal dependencies that traditional approaches cannot easily represent.

Model Performance and Characteristics

Early Mandarin tone recognition systems relied on traditional machine learning models (Table 1), particularly HMMs and SVMs, using pitch features like log-F_0 and its derivatives. These achieved moderate success on isolated syllables but struggled with continuous speech due to limited capacity to model non-linear temporal dynamics such as Tone 3 sandhi and the pitch-ambiguous neutral tone (an unstressed syllable whose pitch contour depends on the preceding syllable's tone). Recognition for connected speech remained poor unless models incorporated strong linguistic priors (J. Liu & Yu, 2000). Decision-tree-enhanced HMMs (Y. Cao et al., 2000) and tone-nucleus-guided variants (J. Zhang

 doi:10.20944/preprints202510.2478.v1

& Hirose, 2004) improved by encoding syllable position and sandhi rules, yet still faltered on Tone 3 and neutral tones.

Similarly, SVMs and shallow multilayer perceptrons (MLPs) succeeded in constrained settings: Chang (1990) reported 93.8% accuracy with a single-hidden-layer MLP (P.-C. Chang et al., 1990), and Wang et al. (2009) achieved 93.52% using binary-class SVMs (S. Wang et al., 2009). However, their fixed-window assumptions failed to capture bidirectional coarticulation. Using continuous F_0 cues, SVM models reached 88.3% accuracy for tone recognition (Y. Chen & Xu, 2021). Combined approaches for continuous speech achieved a recognition rate of 86.72% (S.-H. Chen & Wang, 1995). Even prosody-aware hybrids, such as the SRNN-MLP model (Y. R. Wang & Chen, 1994), which raised accuracy from 91.38% to 93.10% by modeling intonation, were constrained by manually crafted input features and limited temporal context.

The limitations of these earlier approaches call for more powerful and adaptive models. With greater computational resources and larger datasets, deep learning architectures have delivered substantial gains in both isolated and continuous tone recognition (Table 2). Convolutional Neural Networks (CNNs) excel at extracting spatial patterns from spectral inputs (e.g., MFCCs, mel-spectrograms). ToneNet (Q. Gao et al., 2019), a CNN processing RGB mel-spectrograms in the 50–350 Hz range, achieved near-perfect performance on monosyllables (99.16% accuracy, F1 = 99.11%), demonstrating robustness to noise and speaker gender by focusing on low-frequency tone contours. Similarly, CNNs with denoising autoencoders reached 95.53% accuracy on children's speech without explicit F_0 input (C. Chen et al., 2016b). CNNs incorporating explicit F_0 features can reduce error rate by 4.3% -7.1% (X. Hu, Saiko, et al., 2014). As isolated-syllable performance plateaued near ceiling, research shifted toward the harder challenge of continuous speech. Models such as BiLSTMs, attention mechanisms, Transformers, and hierarchical architectures have yielded modest improvements by capturing non-fixed temporal windows and inter-syllabic dependencies (e.g., vowel quality, speaking rate) (Y. Gao et al., 2020; L. Yang et al., 2018a, 2018b).

Beyond native speakers, ML models are increasingly being adapted to L2 learners and clinical populations, where tone production is atypical. Li et al. (2018) used soft-target labels with BiLSTMs to model intermediate pronunciations, reducing mispronunciation detection error to 5.13% (W. Li et al., 2018). The Goodness of Tone (GOT) framework (Tong et al., 2015) provides diagnostic, category-specific confidence scores for personalized CALL feedback. In clinical applications, lightweight ANNs achieved 63.35% accuracy on post-stroke dysarthric speech (Mou et al., 2020), while EEG-based CNNs enabled 51.7% decoding of tones from neural activity (X. Wang et al., 2023), suggesting potential for brain–computer interfaces in nonverbal patients. These efforts reflect a paradigm shift from maximizing accuracy on clean data toward robustness, interpretability, and real-world utility.

Table 1. Summary of traditional models with performance accuracies in selected studies.

Study (Author, Year)	*Traditional Model Architecture*	*Recognition Task*	*Average Four-Tone Accuracy (%)*
Cao et al., 2000	Decision Tree + HMM	Continuous (speech recognition)	70.1
Chang et al., 1990	Shallow MLP	Isolated	93.8
Chen & Wang, 1995	Shallow MLP and basic RNN	Continuous	86.72
Cheng et al., 2003	HMM	Continuous	81.8
Cheng et al., 2020	Wavelet transform + HMM	Isolated	94.17

 doi:10.20944/preprints202510.2478.v1

Garg et al., 2018	Hierarchical SVMs	Isolated	97.93
Hirose & Zhang, 1999	HMM	Continuous	82.5
Hirose & Zhang, 2000	HMM	Continuous	85.5
Jian & Tiecheng, 2000	GMTM	Continuous	60.0
Lin, 2004	HMM + clustering	Broadcast news with noise, continuous	85.0
Liu & Tao, 2013	Fuzzy Algorithm	Continuous	89.5
Liu et al., 1999	GMTM	Continuous	61
Liu et al., 2007	GMM	Conversational telephone speech, continuous	42.8
Liu et al., 2013	SVM + pitch smoothing	Isolated	97.62
Mou et al., 2020	ANN	Isolated, clinical (post-stroke dysarthria)	63.35
Wang et al., 1994	SRNN + MLP	Continuous	93.1
Wang et al., 2008	MLP + tone nucleus modeling	Continuous	80.9
Wang et al., 2009	SVM	Continuous	93.52
Xu et al., 2006	ANN (cochlear)	Isolated, clinical children	90.0
Yan et al., 2023	Random Forest with Feature Fusion	Isolated (+low data)	95.0
Yang et al., 2002	HMM	Isolated	96.53
Zhang & Hirose, 2004	HMM + tone nucleus modeling	Continuous	83.1
Zhang & Kawanami, 1999	HMM	Continuous	86.4

Note: Only studies that reported non-error results were included in the table.

Table 2. Summary of deep learning models with performance accuracies in selected studies.

Study (Author, Year)	*Deep Learning Model Architecture*	*Recognition Task*	*Average Four-Tone Accuracy (%)**
Gao et al., 2019	CNN (ToneNet)	Isolated	99.16

 doi:10.20944/preprints202510.2478.v1

Chen at al., 2016	CNN + dAE	Isolated (children)	95.53
Yang et al., 2018a	CNN-DBLSTM-RNN	Continuous	92.97
Lin et al., 2018	DNN	Continuous	91.25
Peng et al., 2021	CNN-BiGRU	Continuous	89.49
Chen et al., 2014	Deep Maxout Network	Continuous	78.21
Yang et al., 2018b	CNN-BLSTM-RNN + Attention	Continuous	91.55
Tang & Li., 2021	End-to-end Res-NN	Continuous	92.6
Gao et al., 2020	BLSTM + Target Approx. Model	Continuous	87.56
Huang et al., 2021	Encoder-Decoder, DNN-bi-RNN with gating	Continuous	87.60
Hu et al., 2014	CNN	Broadcast/conversational, continuous	95.7
Li et al., 2023	Transformer, tone relabeling	Continuous	87.2
Ryant et al., 2014	DNN (no pitch features)	Broadcast news, continuous	84.44
Wu et al., 2013	NN (accented)	Continuous accented (Shanghai accent)	70.2
*Wang et al., 2023	CNN (using EEG)	Neurotypical spoken tones, EEG imagery-based classification	51.70

*Note: Only studies that reported non-error results were included in the table. *Study excluded in calculation of aggregate statistics.*

As seen in Tables 1 and 2, deep learning models generally achieve higher and more consistent accuracies across both isolated and continuous tasks (mean accuracy 88.82% with a standard deviation of 7.21% vs. 83.06% with a standard deviation of 14.04%). Nevertheless, some traditional models remain competitive, often due to careful integration of Mandarin-specific linguistic knowledge. For example, Garg et al. (2018) reached 97.93% with hierarchical SVMs (Garg et al., 2018), and Liu et al. (2013) achieved 97.62% using pitch-smoothed SVMs on isolated tones (Q. Liu et al., 2013). Traditional models also offer computational efficiency, enabling deployment on mobile devices (Yan et al., 2023). While traditional models can exceed 95% accuracy on isolated syllables, their performance on continuous speech typically falls below 90% with notable exceptions (W.-Y. Lin, 2004; S. Wang et al., 2009). In contrast, deep learning models consistently surpass 85% in continuous tasks, with several exceeding 92% (J. Tang & Li, 2021; Yang et al., 2018a). Even in noisy or accented

 doi:10.20944/preprints202510.2478.v1

conditions (e.g., broadcast speech, Shanghai-accented Mandarin), deep models maintain relatively high performance (X. Hu, Lu, et al., 2014; Wu et al., 2013).

Collectively, these advances trace a clear evolution: traditional models established foundational methods but were bottlenecked by feature engineering and limited context modeling, whereas deep learning enables end-to-end, context-aware, and robust tone recognition, setting the stage for architecture- and population-specific innovations in real-world applications.

Types of Input Features

A broad spectrum of input features has been explored in Mandarin tone recognition, ranging from conventional acoustic features to neurophysiological signals. Pitch-related features such as fundamental frequency (F_0) contours and their derivatives (e.g., velocity, acceleration) remain the most widely used, often smoothed or processed via algorithms like YAAPT, which outperforms alternatives by identifying pitch-informative windows (H. Hu, Zahorian, et al., 2014; H. Huang et al., 2021; H. C.-H. Huang & Seide, 2000; Wong & Siu, 2002). Mel-Frequency Cepstral Coefficients (MFCCs) are also common, capturing spectral envelope characteristics, though they add computational cost and yield minimal accuracy gains when combined with pitch features (C. Chen et al., 2016b; Lei et al., 2005; Lei & Ostendorf, 2007; Yan et al., 2023).

Mel-spectrograms have gained prominence in CNN-based systems, where they serve as image-like inputs for spatial feature learning as in Gao et al.'s (2019) ToneNet, which achieved 99.16% accuracy on isolated tones using RGB-encoded spectrograms (50–350 Hz). These learned representations are typically fed into dense classifiers (e.g., MLPs) for final prediction. Hybrid approaches further enhance performance: Lin et al. (2018) fused MFCCs and F_0 with DNN-derived articulatory posterior probabilities, reducing tone error rate by 52% (to 8.75%) compared to F_0-only baselines, demonstrating the value of combining data-driven and linguistically informed features (J. Lin et al., 2018).

Notably, deep models can internalize pitch information even without explicit F_0 input. Chen et al. (2016a) found that adding handcrafted F_0 or pooled raw MFCCs to a denoising autoencoder–CNN (dAE-CNN) actually reduced accuracy, suggesting redundancy and risk of overfitting. Similarly, pitch-agnostic features like Spectral Amplitude Cepstral Distances (SACD) can outperform F_0 in noise (Ryant et al., 2014), while Fundamental Frequency Variation (FFV) improves robustness in unvoiced regions (H. Hu, Zahorian, et al., 2014). Both sparse and denoising autoencoders yield comparable gains when paired with spectral inputs, underscoring that unsupervised pretraining enhances but does not replace the need for thoughtful input design.

More recently, neural signals have emerged as an alternative modality. Wang et al. (2023) trained an end-to-end CNN on raw EEG to classify imagined Mandarin tones and vowels across subjects, achieving 62.4% accuracy despite challenges like low SNR and inter-subject variability (X. Wang et al., 2023). Other ML studies reported EEG-based recognition rate of 67.7% for the visual only condition and 80.1% for the audio-visual condition (X. Zhang et al., 2020). While modest, this highlights EEG's potential in hybrid acoustic-neural systems, particularly for brain–computer interfaces aiding nonverbal or speech-impaired individuals.

Model Validation and Performance Metrics

Evaluation strategies vary by task (See Supplemental Table S1). For isolated syllables, standard metrics including accuracy, precision, recall, F1-score, and confusion matrices are used. Confusion matrices consistently reveal Tone 3 as the most error-prone category. Given class imbalance (e.g., rare neutral tones), accuracy alone can be misleading; thus, 10-fold cross-validation is common to ensure robustness, especially with small or speaker-limited datasets.

For continuous speech, the Tone Error Rate (TER) analogous to Word Error Rate is preferred. TER computes the Levenshtein distance between predicted and reference tone sequences, accounting for substitutions, insertions, and deletions (J. Lin et al., 2018; Peng et al., 2021; L. Yang et al., 2018b). In mispronunciation detection, where outputs are not canonical tone labels, studies use False

 doi:10.20944/preprints202510.2478.v1

Acceptance Rate (FAR) and False Rejection Rate (FRR) instead (C. Huang et al., 2008; W. Li et al., 2016, 2018).

Across metrics, deep learning models consistently outperform traditional ones, with CNNs dominating isolated tasks (C. Chen et al., 2016b; Q. Gao et al., 2019) and RNNs, BiLSTMs, or attention-based architectures excelling in continuous speech (Y. R. Wang & Chen, 1994; L. Yang et al., 2018b), reflecting tone's inherently temporal nature.

Training and Testing Datasets

Research in this field relies heavily on a few core corpora such as AISHELL, and the Chinese National Hi-Tech Project 863 (See Supplemental Table S2). For continuous speech, Project 863 remains the most frequently used and comprehensive, appearing in 13 studies. It includes both read and spontaneous speech with diverse demographics across hundreds of speakers, covering approximately 110-150 k syllables (51 + hours). Its subset RASC863 adds dialectal variation from four regions, enhancing coverage of accented Mandarin.

Other major resources include HUB4 and GALE, which provide spontaneous, noisy broadcast speech, and Train04 (HKUST), featuring telephone-quality conversational speech, valuable for modeling natural and degraded acoustic conditions. Clean studio datasets such as AISHELL-3 and SCSC focus on controlled tone production, while COSPRO adds detailed prosodic annotation (intonation, stress, tempo).

However, critical gaps remain. Emotional or expressive speech (e.g., angry, whispered, shouted) is largely absent in the tone recognition models. L2 corpora like iCALL (142 hours, 305 learners) exist but receive less attention than native-speaker datasets (N. F. Chen et al., 2015). Clinical and pediatric data are scarce and often custom-collected. For example, datasets from children with cochlear implants (N. Zhou & Xu, 2008) and post-stroke dysarthria (PSD) (Mou et al., 2020) are generally small scale and lack demographic balance and generalizability with less than 3000 tokens in total for both training and testing, compared to Project 863's count of over 100-thousand.

Discussion

Handling Isolated Words vs. Connected Speech

While isolated tone accuracy averages above 95% for deep learning and traditional models, continuous speech lags by over 15%, suggesting that context-aware, adaptive architectures are essential for real-world performance. Isolated tone recognition typically uses monosyllabic utterances recorded under controlled conditions with minimal noise, limited speaker variability, and standardized equipment. These yield stable F_0 trajectories and clean signals that lead to near-perfect classification. Since Mandarin tones are encoded in the temporal evolution of fundamental frequency (F_0), models that jointly capture local spectral cues and global contour dynamics excel (C. Liu & Tao, 2013). Convolutional Neural Networks (CNNs) are particularly effective. For instance, ToneNet (Gao et al., 2019) applies convolutional filters to mel-spectrograms focused on the 50–350 Hz band (where tonal distinctions are most salient), achieving 99.16% accuracy and 99.11% F1-score on the Syllable Corpus of Standard Chinese (SCSC). This demonstrates that, in clean settings, data-driven feature extraction can nearly eliminate ambiguity in monosyllabic tones. In contrast, continuous speech introduces significant challenges due to tonal coarticulation, tone sandhi, variable syllable durations, and overlapping articulatory gestures. As the articulatory movements for consonants and vowels in the vocal tract do not instantaneously shift pitch, tones influence neighboring syllables bidirectionally though carryover effects often dominate (J. Liu et al., 1999; J.-S. Zhang & Kawanami, 1999). While this context can theoretically be argued to aid disambiguation, in practice it distorts canonical contours, making recognition harder. Moreover, syllables in natural speech lack fixed temporal boundaries, necessitating forced alignment for segmentation, a method that assumes uniform timing and often misaligns features under variable speaking rates.

 doi:10.20944/preprints202510.2478.v1

Feedforward models like DNNs and CNNs struggle here due to fixed input windows that miss long-range dependencies. For example, Li et al. (2018) showed DNNs fail to capture early discriminative cues in Tone 3 syllables (W. Li et al., 2018). BiLSTMs address this by modeling past and future context simultaneously, achieving a 7.03% Tone Error Rate (TER) on the Project 863 corpus (L. Yang et al., 2018a). To further refine temporal focus, attention mechanisms, inspired by tone nucleus theory (Zhang & Hirose, 2004), guide models to the most informative segments within each syllable. Yang et al. (2018b) showed attention-augmented CNN-BLSTMs outperform non-attentive versions, mimicking human-like syllable segmentation. This aligns with the anchoring hypothesis (Zhang & Hirose, 2000): the same F_0 contour may be perceived as different tones depending on neighboring pitch levels, showing the need for relative, not absolute, pitch modeling. Hu et al. (2014) corroborated this, showing optimal performance when attention adjusts window widths to emphasize lower-frequency pitch information.

Explicit modeling of coarticulation also helps. Tang and Li (2021) proposed a segment-based end-to-end framework using tri-tone windows (including half-frames from adjacent syllables) to directly encode bidirectional context while avoiding boundary ambiguities. Hierarchical models offer another strategy, especially under data or compute constraints. Garg et al. (2018) combined SVMs and shallow networks in a pipeline that first predicts speaker gender and vowel identity (both strongly influence pitch) before tone classification. By conditioning tone decisions on these factors via weighted voting, the system outperformed standalone CNNs, demonstrating that structured exploitation of interdependencies boosts accuracy even with simple architectures.

Recent work has begun bridging the gap between isolated and connected speech by demonstrating feasibility of learning from raw audio without explicit tone labels and revealing that lexical tone emerges naturally in their internal representations, even when trained on non-tonal languages (Shen et al., 2024). Shen et al. replicated human-like confusion patterns (e.g., T2–T3, T1–T4) and demonstrated that fine-tuning on tonal corpora enhanced tone encoding, whereas fine-tuning on non-tonal languages degraded it. This allows us to reconceptualize tone not as a classification target but as an emergent property shaped by context, prosody, and linguistic exposure. This shift is further advanced by another study (Schenck & Beguš, 2025). Schenck & Beguš introduced the first fully unsupervised generative model hat acquires tonal categories from unlabeled input and reproduces the phonological developmental trajectory observed in children (e.g., T1/T4 => T2 => T3). These advances suggest a paradigm shift: future tone modeling will not merely classify tones but model how they are learned and produced, linking computational systems to human phonological acquisition. With tailored priors or weighting schemes, such unsupervised frameworks could also simulate tonal processing differences in L2 learners or clinical populations.

Cross-Validation Strategies

To evaluate model performance and ensure generalizability across speakers and conditions, researchers have employed a range of cross-validation strategies tailored to dataset size, task complexity, and deployment goals. The most common approach is k-fold cross-validation, particularly 5-fold and 10-fold, which provides robust error estimates on moderately sized corpora. For instance, Chen et al. (2016) used 10-fold CV on children's speech to validate a CNN–denoising autoencoder (95.53% accuracy), while Yan et al. (2023) applied 5-fold CV on the 600-syllable SCSC, showing Random Forest with feature fusion maintains >95% accuracy even with limited data. Although CNN-based models can achieve near-perfect accuracy in isolated syllable conditions, they demand large datasets, high computational cost, and careful preprocessing, whereas lighter traditional models can deliver comparable accuracy with far greater efficiency.

When speaker independence is a primary concern, especially in clinical or real-world applications, leave-one-speaker-out (LOSO) validation is preferred. Historically, cross-speaker validation was underutilized. Xu et al. (2007), analyzing 61 children, were among the first to apply LOSO validation, achieving 84.7% accuracy on unseen speakers. Wang et al. (2008) used LOSO on HKU96 (20 speakers), achieving 80.9% accuracy with a tone nucleus–based MLP, 10% better than

 doi:10.20944/preprints202510.2478.v1

HMMs. Mou et al. (2020) employed a 10× half-split protocol on post-stroke dysarthric speech, simulating real-world generalization to new users.

In studies using large-scale corpora such as the Chinese National Hi-Tech Project 863 and AISHELL-3, most researchers adopted hold-out validation with fixed train/validation/test splits (typically 90/5/5 or 80/10/10). For instance, Yang et al. (2018a) applied a standard hold-out split on the 863 corpus to train a CNN-DBLSTM model, achieving a state-of-the-art 7.03% tone error rate (TER) on continuous speech. Similarly, Peng et al. (2021) divided the same corpus 9:1 for a multi-scale CNN–BiGRU model, while Tang and Li (2021) followed AISHELL-3's official partition for end-to-end tone classification. Although computationally efficient, fixed splits can introduce bias if speaker overlap or class imbalance is not carefully controlled. To counter overfitting, recent models employed early stopping based on validation loss. For example, Huang et al. (2021) terminated training when TER plateaued during joint pitch–tone modeling.

These approaches illustrate a methodological progression from rigorous k-fold or leave-one-speaker-out (LOSO) validation in small or clinical datasets to scalable hold-out protocols suited for deep learning on large corpora, all aiming to achieve robust, speaker-independent tone recognition in real-world conditions. It is important to point out that while many studies rely on high-dimensional acoustic representations, simply concatenating all available features often leads to redundancy and overfitting. Recent work shows that intelligent feature selection can reduce input dimensions to 12-15 critical features while maintaining over 97.5% accuracy (Yan et al., 2023), suggesting that feature quality outweighs feature quantity for robust cross-speaker generalization.

Main Challenges and Opportunities

Despite advances in deep learning and hybrid approaches, there are still several main areas of challenges. The first is acoustic and linguistic complexity (Koser et al., 2018; Y. Xu, 2019; J. Zhang, 2010). Tone 3 remains consistently the most difficult to classify due to heavy coarticulation and tonal sandhi effects in continuous speech, and yet these processes are rarely modeled explicitly. Embedding explicit features to reflect these articulatory patterns and linguistic rules within architectures could potentially improve tone recognition performance (C. Wang & Seneff, 1998; J.-S. Zhang & Hirose, 2000). For example, incorporating prosodic focus into tone recognition significantly improves accuracy, reducing error rates from 15.2% to 8.7% by training focus-conditioned SVMs that exploit the greater tonal clarity of focused syllables (Surendran et al., 2005). But many models omit the neutral tone (L. Tang & Yin, 2006; P. Tang et al., 2017; J. Zhang & Liu, 2011), limiting real-world applicability. In noisy environments (e.g., cafés, streets, telephone calls), F_0-based features degrade rapidly (Z. Liu et al., 2007; P. Zhang et al., 2023). Wang and Seneff (1998) found that telephone-quality digit strings yielded a 22.7% error rate with F_0-only models, while context-dependent HMMs that explicitly modeled sandhi markedly reduced errors. Recent work further shows that careful feature engineering such as fusing F_0 statistics, duration, and energy within ensemble frameworks like Random Forest can achieve over 98% accuracy even under variable acoustic conditions, demonstrating the importance of sound-source features over purely spectral ones. While F0 is dominant, human listeners also exploit secondary cues such as duration, intensity, voice quality, spectral tilt, visual articulatory and gestural cues, and top-down linguistic and contextual knowledge, particularly in noisy or ambiguous conditions (Ding et al., 2025; Farran & Morett, 2024; Garg et al., 2019; S. Liu & Samuel, 2004; J. Wang et al., 2013). Many current models underutilize these multimodal features, limiting robustness in real-world and clinical scenarios.

The second is emotional and prosodic variability (H.-S. Chang et al., 2023; Ding & Zhang, 2023; Xiao & Liu, 2024). Most ML studies optimize for accuracy rather than perceptual realism, creating a gap between computational performance and human-like tonal perception. Incorporating psycholinguistically inspired benchmarks could improve interpretability and relevance to human listeners. For instance, emotional or listener-adaptive speech (such as infant-directed speech) introduces major variability in pitch, intensity, and duration, which can distort F_0 contours that tone models rely on. Early research showed the benefit of incorporating prosodic features (e.g., pitch

 doi:10.20944/preprints202510.2478.v1

contour, syllable duration, energy) (L.-W. Cheng & Lee, 2008; Y. R. Wang & Chen, 1994). For example, incorporating prosodic features reduced character error rate by 17.8%, even without speaker normalization (L.-W. Cheng & Lee, 2008). However, large standardized emotional-speech corpora remain scarce and have not been widely used to test ML models for lexical tone recognition.

The third is dialectal, speaker- and accent-related variability. Most models are trained on young adult male speakers, leaving performance on female, child, elderly or less-represented individuals untested. Mandarin dialects and L2 learner speech also pose distinct challenges (B. Cheng et al., 2025; Fon & Chuang, 2024; Leung et al., 2025; W. Li et al., 2018; H. Zhang et al., 2021; J. Zhang, 2010). Dialectal accents alter tone realizations, while L2 speakers often produce tones that fall between canonical categories, especially T2–T3. Domain adaptation, soft-label learning, and probabilistic modeling (e.g., Li et al., 2018) improve robustness, but training data diversity remains limited. Promising directions include dialect identification as a preprocessing step and hierarchical models that adapt tone classifiers to specific dialectal or L2 patterns. Models can effectively identify dialects and provide a foundation for handling dialectal variation in Mandarin. For example, Zhang et al. (2022) used a gated neural spike P (GSNP) model to classify seven Shandong dialects with 99.7% accuracy (H. Zhang, Liu, et al., 2022), while Zhang et al. (2021) achieved 93% accuracy on Feixian dialect tones using a CNN on 1,000 words (H. Zhang et al., 2021). Identifying dialects enables training specialized submodels or hierarchical architectures to better account for accent and pronunciation effects.

Similarly, second-language (L2) Mandarin learners present a unique challenge for tone recognition systems: their productions often fall between canonical tone categories, exhibit irregular speaking rates, and suffer from incomplete sandhi realization, particularly for Tone 3 (S. Chen et al., 2022; Pelzl et al., 2021; Xi et al., 2021). Traditional hard-label classifiers, trained on native speech, misclassify these ambiguous utterances as errors rather than legitimate intermediate forms. To address this, probabilistic modeling has emerged as the gold standard. Li et al. (2018) demonstrated that replacing hard targets with soft-label tone probabilities (e.g., [0.2, 0.6, 0.1, 0.1] for a Tone 2–Tone 3 blend) reduced the Equal Error Rate (EER) from 5.77% to 5.13% in mispronunciation detection. Their CD-BLSTM-HMM model outperformed DNN baselines because BLSTMs explicitly capture long-range coarticulation and variable syllable durations, which are common in L2 speech where learners slow down on difficult tones like Tone 2 and Tone 3. Critically, forced alignment fails in spontaneous L2 speech due to variable speaking rates. Instead, it could be beneficial to use attention mechanisms (Yang et al., 2018b) or CTC-based models (Peng et al., 2021) that dynamically align frames without pre-segmentation. For real-time CALL applications, however, computational efficiency is paramount (Hussein et al., 2012; Su & Miao, 2006). Yan et al. (2023) showed that Random Forest with feature fusion achieves >95% accuracy on monosyllables with <0.003s inference time, making it ideal for on-device mobile apps. Crucially, it generalizes well to unbalanced data (e.g., 1:7:1:1 tone splits), a common scenario in L2 corpora where Tone 3 errors dominate. In addition, it outperforms ToneNet (Q. Gao et al., 2019) and CNN-dAE (denoising autoencoder) (C. Chen et al., 2016a) in low-data situations, while requiring both less computational power and training data. Thus a hybrid deployment strategy is feasible: use efficient lightweight models for initial screening, and cloud-based BLSTM systems (e.g., Li et al., 2018) for detailed diagnostic feedback.

The fourth is clinical utility and cross-linguistic scalability. Tone recognition for clinical populations (e.g., post-stroke dysarthria, cochlear implant users) faces dual constraints: small, non-public datasets and high variability in voice and speech quality. It has been shown that manipulating amplitude envelopes can enhance tone perception in cochlear implant users (Luo & Fu, 2004; Meng et al., 2016, 2017). For clinical speech conditions such as post-stroke dysarthria (PSD) or cognitive decline, models must prioritize sensitivity to subtle prosodic degradation over raw accuracy. Mou et al. (2020) reported that a simple ANN using F_0 inputs achieved 63.35% accuracy on PSD speech, comparable to human listeners (70.28%) while the CNN model performed poorly (34.71%), showing the limitations of spectrogram-based models on small, atypical datasets. Effective clinical models should employ prosody-rich features (F_0 slope, duration, energy), track longitudinal changes, and

 doi:10.20944/preprints202510.2478.v1

avoid forced alignment, which introduces errors in disfluent speech. However, speech corpora for clinical populations remain scarce. Expanding datasets for groups such as PSD, aphasia, autism , cochlear implant users is essential, though ethical considerations, especially regarding consent and data privacy for children must guide collection. Furthermore, existing methods are highly Mandarin-specific, limiting extension to under-resourced tonal languages with different tone inventories. Adapting feature engineering and model architectures for languages such as Thai, Yoruba or Hmong remains a significant challenge and opportunity. Transparent data practices and larger, representative clinical corpora are critical for advancing inclusive translational tone recognition research.

Practical Recommendations for Model Development and Deployment

To address the identified gaps, we make the following evidence-based recommendations that could guide the design, training, and deployment of Mandarin tone recognition systems across diverse contexts.

1. Integrate Mandarin-specific linguistic structure into model design. Rather than treating tone as a generic classification problem, architectures should explicitly account for phonological rules such as tone sandhi and prosodic phrasing. Effective strategies include implementing sandhi-aware output subclasses (Cao et al., 2000), attention mechanisms guided by tone nucleus theory (Hirose & Zhang, 1999; X.-D. Wang et al., 2008; J. Zhang & Hirose, 2004), or prosody-sensitive hierarchical models (W.-H. Li et al., 2023). These approaches improve robustness, particularly in low-resource or noisy settings, by aligning model inductive biases with the underlying structure of Mandarin tone.

2. Prioritize data-efficient and multimodal learning for underrepresented populations. Labeled data for L2 learners and clinical speakers remains scarce. To address this, leverage self-supervised pretraining on unlabeled speech (Tang & Li, 2021) and augment acoustic inputs with complementary signals such as articulatory features (Lin et al., 2018), visual cues (Garg et al., 2019), or neural data (Wang et al., 2023) to stabilize and enhance model performance when acoustic cues are degraded. Multi-stage architectures that combine deep feature extractors (e.g., CNNs) with interpretable classifiers (e.g., BLSTMs, HMMs, or SVMs) in hierarchical configurations, explicitly modeling gender, speaking rate, or sandhi context, offer a balanced trade-off between performance and transparency.

3. Collect and use more representative, naturalistic datasets. We need to move beyond read speech from standard Mandarin speakers in studio conditions and prioritize corpora that include L2 learners, clinical populations (e.g., cochlear implant users, individuals with autism, aphasia, or dysarthria), regional accents, emotional prosody, and spontaneous conversation recorded on consumer devices. Such data captures real-world phenomena like coarticulation, disfluencies, and turn-taking, enabling the development of accent-adaptive systems and dialect-aware front-ends that route input to specialized submodels.

4. Match model complexity to deployment constraints. High-accuracy models like ToneNet (99.16% on isolated tones) are unsuitable for on-device or clinical monitoring due to computational demands. For real-time applications, especially in CALL or mobile health, favor lightweight, interpretable methods (Wei et al., 2007). Yan et al. (2023) demonstrated that Random Forest with feature fusion achieves >95% accuracy with sub-3ms latency, even on tiny datasets. For scalable deployment, consider cloud-edge hybrids: use efficient on-device models for initial screening and reserve resource-intensive BLSTMs or Transformers for cloud-based refinement when needed.

5. Design for equity and linguistic diversity. We should avoid penalizing non-native or dialect speakers by replacing canonical tone labels with surface-tone representations that reflect actual pronunciation (Li et al., 2023), regularly audit model performance across speaker subgroups (e.g., by gender, L1 background, or dialect; Tong et al., 2015), and involve end-users such as L2 learners and speech language pathologists in dataset curation and evaluation. Additionally, we should explore cross-linguistic extensions(e.g., Cantonese, Thai) and code-switching scenarios (e.g., Mandarin-English), which can reveal generalizable tonal features and broaden real-world utility.

 doi:10.20944/preprints202510.2478.v1

6. Integrate tone modeling into end-to-end Mandarin ASR. Transformer-based end-to-end architectures offer a natural framework for jointly modelling lexical, prosodic, and tonal information within full ASR pipelines. Model selection should remain task- and context-aware, taking into account and balancing accuracy, efficiency, and inclusivity principles as operationalized in Table 3.

Table 3. Model selection recommendations for specific tasks/purposes.

Recognition Task	Resource Constraints	Recommended Approach	Rationale
Isolated monosyllables	High data availability with high computing resources	CNN with low-frequency mel-spectrogram input Ex: ToneNet	Maximizes accuracy (>99%) by focusing on tone-relevant spectral regions; robust to noise
Isolated monosyllables	Low data for training, limited computing resources/mobile applications	Random Forest with feature fusion	Achieves >95% accuracy on tiny datasets (<600 syllables), real-time inference (<0.003s), ideal for mobile CALL apps
Continuous read speech	High data availability with high computing resources	CNN-BLSTM with attention	Captures bidirectional coarticulation and sandhi; lowest reported TER (7.03%)
Continuous spontaneous/conversational speech	High data availability with high computing resources	Multi-scale CNN-BiGRU with prosodic features	Handles variable tempo, intonation, and phrase-level effects; improves robustness in real-world settings
L2 learner or clinical speech, usually monosyllable	Limited data availability	BLSTM with soft-target tone labels, probabilistic approaches	Models ambiguous, non-canonical productions more effectively than hard classification; reduces EER to 5.13%

Conclusions

This systematic literature review examined developments in Chinese tone recognition with machine learning from early models such as HMMs and SVMs to modern deep architectures and hybrid systems, including CNNs and BLSTMs. While the advances have yielded excellent performance on controlled, isolated-syllable tasks, significant challenges persist in handling the complexities of continuous speech, emotional prosody, dialectal variation, and robust cross-validation across diverse datasets. Real-world deployment remains limited by a lack of diverse, representative datasets, weak modeling of prosody and dialect variation, and inconsistent validation practices. Future research should prioritize multimodal, special-purpose corpora and lightweight, efficient models capable of real-time tone assessment. Such systems would enhance user-centered

 doi:10.20944/preprints202510.2478.v1

applications in computer-assisted language learning and clinical speech therapy to ensure that technical progress translates into meaningful pedagogical and communicative outcomes.

Supplementary Materials: The following supporting information can be downloaded at website of this paper posted on Preprints.org.

Data Availability Statement: The data that support the findings of this systematic review were derived from existing, published literature and are available within the article and its references. Due to copyright restrictions, the original, full-text articles from the primary sources cannot be openly shared, but are available from the original publishers/third parties as cited in the reference list.

Acknowledgements: This project received support by an award to YZ from the University of Minnesota's Grant-in-Aid of Research, Artistry, and Scholarship program.

Conflict of Interests: We have no conflicts of interest to disclose.

References

1. Cao, M., Pavlik, P. I., & Bidelman, G. M. (2024). Enhancing lexical tone learning for second language speakers: Effects of acoustic properties in Mandarin tone perception. Frontiers in Psychology, 15, 1403816. https://doi.org/10.3389/fpsyg.2024.1403816
2. Cao, Y., Deng, Y., Zhang, H., Huang, T., & Xu, B. (2000). Decision tree based Mandarin tone model and its application to speech recognition. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00ch37100)*, *3*, 1759–1762 vol.3. https://doi.org/10.1109/ICASSP.2000.862093
3. Chang, H.-S., Lee, C.-Y., Wang, X., Young, S.-T., Li, C.-H., & Chu, W.-C. (2023). Emotional tones of voice affect the acoustics and perception of Mandarin tones. *PLOS One*, *18*(4), e0283635. https://doi.org/10.1371/journal.pone.0283635
4. Chang, P.-C., Sun, S.-W., & Chen, S.-H. (1990). Mandarin tone recognition by multi-layer perceptron. *International Conference on Acoustics, Speech, and Signal Processing*, 517–520 vol.1. https://doi.org/10.1109/ICASSP.1990.115763
5. Chen, C., Bunescu, R., Xu, L., & Liu, C. (2016a). Mandarin tone recognition based on unsupervised feature learning from spectrograms. *Journal of the Acoustical Society of America*, *140*(4_Supplement), 3394–3394. https://doi.org/10.1121/1.4970872
6. Chen, C., Bunescu, R., Xu, L., & Liu, C. (2016b). Tone classification in Mandarin Chinese using convolutional neural networks. *Interspeech 2016*, 2150–2154. https://doi.org/10.21437/Interspeech.2016-528
7. Chen, M., Yang, Z., & Liu, W. (2014). Deep neural networks for Mandarin tone recognition. *2014 International Joint Conference on Neural Networks (IJCNN)*, 1154–1158. https://doi.org/10.1109/IJCNN.2014.6889515
8. Chen, N. F., Tong, R., Wee, D., Lee, P., Ma, B., & Li, H. (2015). iCALL corpus: Mandarin Chinese spoken by non-native speakers of european descent. *Interspeech 2015*, 324–328. https://doi.org/10.21437/Interspeech.2015-148
9. Chen, S., Li, B., He, Y., Chen, S., Yang, Y., & Zhou, F. (2022). The effects of perceptual training on speech production of Mandarin sandhi tones by tonal and non-tonal speakers. *Speech Communication*, *139*, 10–21. https://doi.org/10.1016/j.specom.2022.02.008
10. Chen, S.-H., & Wang, Y.-R. (1995). Tone recognition of continuous Mandarin speech based on neural networks. *IEEE Transactions on Speech and Audio Processing*, *3*(2), 146–150. https://doi.org/10.1109/89.366544
11. Chen, X., & Sidtis, D. (2024). Contrastive stress in persons with parkinson's disease who speak Mandarin: Task effect in production and preserved perception. *Journal of Neurolinguistics*, *69*, 101173. https://doi.org/10.1016/j.jneuroling.2023.101173
12. Chen, Y., & Xu, Y. (2021). Parallel recognition of Mandarin tones and focus from continuous F0. *1st International Conference on Tone and Intonation (TAI)*, 171–175. https://doi.org/10.21437/TAI.2021-35
13. Cheng, B., Liao, K., Xiang, Y., Zou, Y., Zhang, X., & Zhang, Y. (2025). Development and validation of an AI-enhanced multimodal training program: Evidence from non-native Mandarin tone learning. *Computer Assisted Language Learning*, *0*(0), 1–25. https://doi.org/10.1080/09588221.2025.2571696

14. Cheng, J., Yi, K., & Li, B. (2000). Mandarin tone recognition based on wavelet transform and hidden markov modeling. *Journal of Electronics (China)*, *17*(1), 1–8. https://doi.org/10.1007/s11767-000-0015-y
15. Cheng, L.-W., & Lee, L. (2008). Improved large vocabulary Mandarin speech recognition by selectively using tone information with a two-stage prosodic model. *Interspeech 2008*, 1137–1140. https://doi.org/10.21437/Interspeech.2008-346
16. Cheng, M., Cheng, X., & Zhao, L. (2003). HMM based recognition of Chinese tones in continuous speech. *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, *2*, 916-919 Vol.2. https://doi.org/10.1109/ICNNSP.2003.1280749
17. Ding, H., Zhang, J., Zhang, H., Chen, F., & Zhang, Y. (2025). Multimodal training using pitch gestures improves Mandarin tone recognition in noise for children with cochlear implants. *Journal of the Acoustical Society of America*, *158*(4), 2995–3005. https://doi.org/10.1121/10.0039561
18. Ding, H., & Zhang, Y. (2023). Speech prosody in mental disorders. *Annual Review of Linguistics*, *9*, 335–355. https://doi.org/10.1146/annurev-linguistics-030421-065139
19. Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University PressOxford. https://doi.org/10.1093/oso/9780199215782.001.0001
20. Farran, B. M., & Morett, L. M. (2024). Multimodal cues in L2 lexical tone acquisition: Current research and future directions. *Frontiers in Education*, *9*, 1410795. https://doi.org/10.3389/feduc.2024.1410795
21. Feng, G., Gan, Z., Llanos, F., Meng, D., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2021). A distributed dynamic brain network mediates linguistic tone representation and categorization. *NeuroImage*, *224*, 117410. https://doi.org/10.1016/j.neuroimage.2020.117410
22. Fon, J., & Chuang, Y.-Y. (2024). When a rise is not only a rise: An acoustic analysis of the impressionistic distinction between northern and central taiwan Mandarin using tone 1 as an example. *Journal of the International Phonetic Association*, *54*(2), 738–769. https://doi.org/10.1017/S0025100324000100
23. Gandour, J., Petty, S. H., & Dardarananda, R. (1988). Perception and production of tone in aphasia. *Brain and Language*, *35*(2), 201–240. https://doi.org/10.1016/0093-934X(88)90109-5
24. Gao, Q., Sun, S., & Yang, Y. (2019). ToneNet: A CNN model of tone classification of Mandarin Chinese. *Interspeech 2019*, 3367–3371. https://doi.org/10.21437/Interspeech.2019-1483
25. Gao, Y., Zhang, X., Xu, Y., Zhang, J., & Birkholz, P. (2020). An investigation of the target approximation model for tone modeling and recognition in continuous Mandarin speech. *Interspeech 2020*, 1913–1917. https://doi.org/10.21437/Interspeech.2020-2823
26. Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication*, *113*, 47–62. https://doi.org/10.1016/j.specom.2019.08.003
27. Garg, S., Hamarneh, G., Jongman, A., Sereno, J., & Wang, Y. (2018). Joint gender-, tone-, vowel-classification via novel hierarchical classification for annotation of monosyllabic Mandarin word tokens. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5744–5748. https://doi.org/10.1109/ICASSP.2018.8462587
28. Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230
29. Hirose, K., & Zhang, J. (1999). Tone recognition of Chinese continuous speech using tone critical segments. *6th European Conference on Speech Communication and Technology*, 879–882. https://doi.org/10.21437/Eurospeech.1999-227
30. Hong, T., Wang, J., Zhang, L., Zhang, Y., Shu, H., & Li, P. (2019). Age-sensitive associations of segmental and suprasegmental perception with sentence-level language skills in Mandarin-speaking children with cochlear implants. *Research in Developmental Disabilities*, *93*, 103453. https://doi.org/10.1016/j.ridd.2019.103453
31. Hu, H., Zahorian, S. A., Guzewich, P., & Wu, J. (2014). Acoustic features for robust classification of Mandarin tones. *Interspeech 2014*, 1352–1356. https://doi.org/10.21437/Interspeech.2014-334

 doi:10.20944/preprints202510.2478.v1

32. Hu, X., Lu, X., & Hori, C. (2014). Mandarin speech recognition using convolution neural network with augmented tone features. *The 9th International Symposium on Chinese Spoken Language Processing*, 15–18. https://doi.org/10.1109/ISCSLP.2014.6936674
33. Hu, X., Saiko, M., & Hori, C. (2014). Incorporating tone features to convolutional neural network to improve Mandarin/thai speech recognition. *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 1–5. https://doi.org/10.1109/APSIPA.2014.7041576
34. Huang, C., Zhang, F., Soong, F. K., & Chu, M. (2008). Mispronunciation detection for Mandarin Chinese. *Interspeech 2008*, 2655–2658. https://doi.org/10.21437/Interspeech.2008-658
35. Huang, H. C.-H., & Seide, F. (2000). Pitch tracking and tone features for Mandarin speech recognition. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00ch37100)*, *3*, 1523–1526 vol.3. https://doi.org/10.1109/ICASSP.2000.861942
36. Huang, H., Wang, K., Hu, Y., & Li, S. (2021). Encoder-decoder based pitch tracking and joint model training for Mandarin tone classification. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6943–6947. https://doi.org/10.1109/ICASSP39728.2021.9413888
37. Hussein, H., Mixdorff, H., & Hoffmann, R. (2012). Real-time tone recognition in a computer-assisted language learning system for german learners of Mandarin. In R. Mamidi & K. Prahallad (Eds.), *Proceedings of the Workshop on Speech and Language Processing Tools in Education* (pp. 37–42). The COLING 2012 Organizing Committee. https://aclanthology.org/W12-5805/
38. Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., & Dawson, L. (2014). A systematic review of speech recognition technology in health care. *BMC Medical Informatics and Decision Making*, *14*(1), 94. https://doi.org/10.1186/1472-6947-14-94
39. Koser, N., Oakden, C., & Jardine, A. (2018). Tone association and output locality in non-linear structures. *Proceedings of the Annual Meetings on Phonology*. https://doi.org/10.3765/amp.v7i0.4476
40. Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *Speech Communication*, *52*(11), 900–910. https://doi.org/10.1016/j.specom.2010.01.004
41. Lee, M.-C., Yeh, S.-C., Chang, J.-W., & Chen, Z.-Y. (2022). Research on Chinese speech emotion recognition based on deep neural network and acoustic features. *Sensors (Basel, Switzerland)*, *22*(13), 4744. https://doi.org/10.3390/s22134744
42. Lei, X., Hwang, M.-Y., & Ostendorf, M. (2005). Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR. *Interspeech 2005*, 2981–2984. https://doi.org/10.21437/Interspeech.2005-134
43. Lei, X., & Ostendorf, M. (2007). Word-level tone modeling for Mandarin speech recognition. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV-665-IV–668. https://doi.org/10.1109/ICASSP.2007.367000
44. Leung, K. K. W., Lu, Y.-A., & Wang, Y. (2025). Examining speech perception–production relationships through tone perception and production learning among indonesian learners of Mandarin. *Brain Sciences*, *15*(7), 671. https://doi.org/10.3390/brainsci15070671
45. Li, Q., Mai, Q., Wang, M., & Ma, M. (2024). Chinese dialect speech recognition: A comprehensive survey. *Artificial Intelligence Review*, *57*(2), 25. https://doi.org/10.1007/s10462-023-10668-0
46. Li, W., Chen, N. F., Siniscalchi, S. M., & Lee, C.-H. (2018). Improving Mandarin tone mispronunciation detection for non-native learners with soft-target tone labels and BLSTM-based deep models. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6249–6253. https://doi.org/10.1109/ICASSP.2018.8461629
47. Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C.-H. (2016). Using tone-based extended recognition network to detect non-native Mandarin tone mispronunciations. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–4. https://doi.org/10.1109/APSIPA.2016.7820701
48. Li, W.-H., Chiang, C.-Y., & Liu, T.-H. (2023). Tone labeling by deep learning-based tone recognizer for Mandarin speech. *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 873–880. https://doi.org/10.1109/APSIPAASC58517.2023.10317518

 doi:10.20944/preprints202510.2478.v1

49. Lin, J., Li, W., Gao, Y., Xie, Y., Chen, N. F., Siniscalchi, S. M., Zhang, J., & Lee, C.-H. (2018). Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks. *Journal of Signal Processing Systems*, *90*(7), 1077–1087. https://doi.org/10.1007/s11265-018-1334-2
50. Lin, W.-Y. (2004). Tone variation modeling for fluent Mandarin tone recognition based on clustering. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I–933. https://doi.org/10.1109/ICASSP.2004.1326140
51. Lin, Y.-T., & Lai, Y.-H. (2024). Phonological processing in Chinese word repetition: Dementia effect and age effect. *Clinical Linguistics & Phonetics*, *38*(10), 970–986. https://doi.org/10.1080/02699206.2023.2278421
52. Liu, C., & Tao, J. (2013). Mandarin tone recognition considering context information. *2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2013)*, 1–5. https://doi.org/10.1109/ICSPCC.2013.6663920
53. Liu, J., He, X., Mo, F., & Yu, T. (1999). Study on tone classification of Chinese continuous speech in speech recognition system. *6th European Conference on Speech Communication and Technology*, 891–894. https://doi.org/10.21437/Eurospeech.1999-230
54. Liu, J., & Yu, T. (2000). New tone recognition methods for Chinese continuous speech. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, vols. 1, 377-380–0. https://doi.org/10.21437/ICSLP.2000-93
55. Liu, M., & Chen, Y. (2020). The roles of segment and tone in Bi-dialectal auditory word recognition. *Speech Prosody 2020*, 640–644. https://doi.org/10.21437/SpeechProsody.2020-131
56. Liu, Q., Wang, J., Wang, M., Jiang, P., Yang, X., & Xu, J. (2013). A pitch smoothing method for Mandarin tone recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *6*(4), 245–254.
57. Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*, *47*(Pt 2), 109–138. https://doi.org/10.1177/00238309040470020101
58. Liu, Z., Shao, J., Zhang, P., Zhao, Q., Yan, Y., & Feng, J. (2007). Real context model for tone recognition in Mandarin conversational telephone speech. *Third International Conference on Natural Computation (ICNC 2007)*, *2*, 696–699. https://doi.org/10.1109/ICNC.2007.595
59. Luo, X., & Fu, Q.-J. (2004). Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. *Journal of the Acoustical Society of America*, *116*(6), 3659–3667. https://doi.org/10.1121/1.1783352
60. Meng, Q., Zheng, N., & Li, X. (2016). Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants. *Journal of the Acoustical Society of America*, *139*(1), 301–310. https://doi.org/10.1121/1.4939707
61. Meng, Q., Zheng, N., & Li, X. (2017). Loudness contour can influence Mandarin tone recognition: Vocoder simulation and cochlear implants. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(6), 641–649. https://doi.org/10.1109/TNSRE.2016.2593489
62. Mou, Z., Ye, W., Chang, C.-C., & Mao, Y. (2020). The application analysis of neural network techniques on lexical tone rehabilitation of Mandarin-speaking patients with post-stroke dysarthria. *IEEE Access*, *8*, 90709–90717. https://doi.org/10.1109/ACCESS.2020.2994069
63. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71
64. Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021). Even in the best-case scenario l2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioral and event-related potentials experiments. *Studies in Second Language Acquisition*, *43*(2), 268–296. https://doi.org/10.1017/S027226312000039X
65. Peng, L., Dai, W., Ke, D., & Zhang, J. (2021). Multi-scale model for Mandarin tone recognition. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362063

66. Ryant, N., Yuan, J., & Liberman, M. (2014). Mandarin tone classification without pitch tracking. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4868–4872. https://doi.org/10.1109/ICASSP.2014.6854527
67. Schenck, K., & Beguš, G. (2025). *Unsupervised learning and representation of Mandarin tonal categories by a generative CNN* (No. arXiv:2509.17859). arXiv. https://doi.org/10.48550/arXiv.2509.17859
68. Shen, G., Watkins, M., Alishahi, A., Bisazza, A., & Chrupała, G. (2024). Encoding of lexical tone in self-supervised models of spoken language. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4250–4261. https://doi.org/10.18653/v1/2024.naacl-long.239
69. Shour, A. R., Anguzu, R., & Onitilo, A. A. (2025). Speech recognition technology and documentation efficiency. *JAMA Network Open*, *8*(3), e251526. https://doi.org/10.1001/jamanetworkopen.2025.1526
70. Song, W., & Deng, I. (2025). A hybrid architecture combining CNN, LSTM, and attention mechanisms for automatic speech recognition. *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*, 285–292. https://doi.org/10.1109/ICCAI66501.2025.00052
71. Su, W., & Miao, Z. (2006). Speech and tone recognition for a Mandarin e-learning system. *TENCON 2006 - 2006 IEEE Region 10 Conference*, 1–3. https://doi.org/10.1109/TENCON.2006.343765
72. Surendran, D., Levow, G.-A., & Xu, Y. (2005). Tone recognition in Mandarin using focus. *Interspeech 2005*, 3301–3304. https://doi.org/10.21437/Interspeech.2005-577
73. Tang, J., & Li, M. (2021). End-to-end Mandarin tone classification with short term context information. *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 878–883. https://ieeexplore.ieee.org/abstract/document/9689521
74. Tang, L., & Yin, J. (2006). Mandarin tone recognition based on pre-classification. *2006 6th World Congress on Intelligent Control and Automation*, *2*, 9468–9472. https://doi.org/10.1109/WCICA.2006.1713835
75. Tang, P., Xu Rattanasone, N., Yuen, I., & Demuth, K. (2017). Acoustic realization of Mandarin neutral tone and tone sandhi in infant-directed speech and lombard speech. *Journal of the Acoustical Society of America*, *142*(5), 2823–2835. https://doi.org/10.1121/1.5008372
76. Tong, R., Chen, N. F., Ma, B., & Li, H. (2015). Goodness of tone (GOT) for non-native Mandarin tone recognition. *Interspeech 2015*, 801–805. https://doi.org/10.21437/Interspeech.2015-254
77. Wang, C., & Seneff, S. (1998). A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition. *5th International Conference on Spoken Language Processing (ICSLP 1998)*, paper 535-0. https://doi.org/10.21437/ICSLP.1998-140
78. Wang, J., Shu, H., Zhang, L., Liu, Z., & Zhang, Y. (2013). The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. *Journal of the Acoustical Society of America*, *134*(1), EL91–EL97. https://doi.org/10.1121/1.4811159
79. Wang, S., Tang, Z., Zhao, Y., & Ji, S. (2009). Tone recognition of continuous Mandarin speech based on binary-class SVMs. *2009 First International Conference on Information Science and Engineering*, 710–713. https://doi.org/10.1109/ICISE.2009.1313
80. Wang, X., Li, M., Li, H., Pun, S. H., & Chen, F. (2023). Cross-subject classification of spoken Mandarin vowels and tones with EEG signals: A study of end-to-end CNN with fine-tuning. *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 535–539. https://doi.org/10.1109/APSIPAASC58517.2023.10317100
81. Wang, X., & Xu, L. (2020). Mandarin tone perception in multiple-talker babbles and speech-shaped noise. *Journal of the Acoustical Society of America*, *147*(4), EL307–EL313. https://doi.org/10.1121/10.0001002
82. Wang, X.-D., Hirose, K., Zhang, J.-S., & Minematsu, N. (2008). Tone recognition of continuous Mandarin speech based on tone nucleus model and neural network. *IEICE Transactions on Information and Systems*, *E91.D*(6), 1748–1755. https://doi.org/10.1093/ietisy/e91-d.6.1748
83. Wang, Y. R., & Chen, S. H. (1994). Tone recognition of continuous Mandarin speech assisted with prosodic information. *Journal of the Acoustical Society of America*, *96*(5 Pt 1), 2637–2645. https://doi.org/10.1121/1.411274

84. Wei, S., Wang, H.-K., Liu, Q.-S., & Wang, R.-H. (2007). CDF-matching for automatic tone error detection in Mandarin CALL system. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV-205-IV–208. https://doi.org/10.1109/ICASSP.2007.367199
85. Wong, P.-F., & Siu, M.-H. (2002). Integration of tone related feature for Chinese speech recognition. *6th International Conference on Signal Processing, 2002.*, *1*, 476–479 vol.1. https://doi.org/10.1109/ICOSP.2002.1181095
86. Wu, J., Zahorian, S. A., & Hu, H. (2013). Tone recognition for continuous accented Mandarin Chinese. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7180–7183. https://doi.org/10.1109/ICASSP.2013.6639056
87. Xi, J., Xu, H., Zhu, Y., Zhang, L., Shu, H., & Zhang, Y. (2021). Categorical perception of Chinese lexical tones by late second language learners with high proficiency: Behavioral and electrophysiological measures. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4695–4704. https://doi.org/10.1044/2021_JSLHR-20-00210
88. Xiao, C., & Liu, J. (2024). The perception of emotional prosody in Mandarin Chinese words and sentences. *Second Language Research*, 2676583241286748. https://doi.org/10.1177/02676583241286748
89. Xu, L., Chen, X., Zhou, N., Li, Y., Zhao, X., & Han, D. (2007). Recognition of lexical tone production of children with an artificial neural network. *Acta Oto-Laryngologica*, *127*(4), 365–369. https://doi.org/10.1080/00016480601011477
90. Xu, Y. (2019). Prosody, tone, and intonation. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 314–356). Routledge.
91. Yan, J., Meng, Q., Tian, L., Wang, X., Liu, J., Li, M., Zeng, M., & Xu, H. (2023). A Mandarin tone recognition algorithm based on random forest and feature fusion. *Mathematics*, *11*(8), 1879. https://doi.org/10.3390/math11081879
92. Yang, L., Xie, Y., & Zhang, J. (2018a). Applying deep bidirectional long short-term memory to Mandarin tone recognition. *2018 14th IEEE International Conference on Signal Processing (ICSP)*, 1124–1127. https://doi.org/10.1109/ICSP.2018.8652486
93. Yang, L., Xie, Y., & Zhang, J. (2018b). Improving Mandarin tone recognition using convolutional bidirectional long short-term memory with attention. *Interspeech 2018*, 352–356. https://doi.org/10.21437/Interspeech.2018-2561
94. Yang, W.-J., Lee, J.-C., Chang, Y.-C., & Wang, H.-C. (1988). Hidden markov model for Mandarin lexical tone recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(7), 988–992. https://doi.org/10.1109/29.1620
95. Zhang, H., Liu, X., & Shao, Y. (2022). Chinese dialect tone's recognition using gated spiking neural P systems. *Journal of Membrane Computing*, *4*(4), 284–292. https://doi.org/10.1007/s41965-022-00113-6
96. Zhang, H., Ma, W., Ding, H., Peng, G., & Zhang, Y. (2022). Phonological awareness and working memory in Mandarin-speaking preschool-aged children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *65*(11), 4485–4497. https://doi.org/10.1044/2022_JSLHR-22-00059
97. Zhang, H., Ma, W., Ding, H., & Zhang, Y. (2023). Sustainable benefits of high variability phonetic training in Mandarin-speaking kindergarteners with cochlear implants: Evidence from categorical perception of lexical tones. *Ear and Hearing*, *44*(5), 990–1006. https://doi.org/10.1097/AUD.0000000000001341
98. Zhang, H., Shao, Y., Yu, X., & Jin, Y. (2021). The method based on spectrogram's image classification for Chinese dialect's tone recognition. *2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET)*, 75–78. https://doi.org/10.1109/CCET52649.2021.9544275
99. Zhang, J. (2010). Issues in the analysis of Chinese tone. *Language and Linguistics Compass*, *4*(12), 1137–1153. https://doi.org/10.1111/j.1749-818X.2010.00259.x
100. Zhang, J., & Hirose, K. (2004). Tone nucleus modeling for Chinese lexical tone recognition. *Speech Communication*, *42*(3), 447–466. https://doi.org/10.1016/j.specom.2004.01.001
101. Zhang, J., & Liu, J. (2011). Tone sandhi and tonal coarticulation in tianjin Chinese. *Phonetica*, *68*(3), 161–191. https://doi.org/10.1159/000333387

 doi:10.20944/preprints202510.2478.v1

102. Zhang, J.-S., & Hirose, K. (2000). Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00ch37100)*, *3*, 1419–1422 vol.3. https://doi.org/10.1109/ICASSP.2000.861859
103. Zhang, J.-S., & Kawanami, H. (1999). Modeling carryover and anticipation effects for Chinese tone recognition. *6th European Conference on Speech Communication and Technology*, 747–750. https://doi.org/10.21437/Eurospeech.1999-194
104. Zhang, M., Tang, E., Ding, H., & Zhang, Y. (2024). Artificial intelligence and the future of communication sciences and disorders: A bibliometric and visualization analysis. *Journal of Speech, Language, and Hearing Research*, *67*(11), 4369–4390. https://doi.org/10.1044/2024_JSLHR-24-00157
105. Zhang, P., Huang, Y., Yang, C., & Jiang, W. (2023). Estimate the noise effect on automatic speech recognition accuracy for Mandarin by an approach associating articulation index. *Applied Acoustics*, *203*, 109217. https://doi.org/10.1016/j.apacoust.2023.109217
106. Zhang, X., Li, H., & Chen, F. (2020). EEG-based classification of imaginary Mandarin tones. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3889–3892. https://doi.org/10.1109/EMBC44109.2020.9176608
107. Zhou, J., Tian, Y., Shi, Y., Huang, C., & Chang, E. (2004). Tone articulation modeling for Mandarin spontaneous speech recognition. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I–997. https://doi.org/10.1109/ICASSP.2004.1326156
108. Zhou, N., & Xu, L. (2008). Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants. *Journal of the Acoustical Society of America*, *123*(3), 1653–1664. https://doi.org/10.1121/1.2832623