ProWhistress: An Enhanced Dual-Stream Transcription Architecture for Prosody-Aware Sentence Stress Detection

Hujian Gu[1], *Li Tao*[1,**], *Fei Jiang*[2], *Ying Wang*[1], *Jingwei Qu*[1], *Zhaofang Yang*[1]

[1] Southwest University, China
[2] Chongqing Academy of Science and Technology, China

ghj097836@email.swu.edu.cn,lti@swu.edu.cn

Abstract

Sentence stress is crucial for conveying speaker intent and meaning. Yet recent alignment-free stress detection models struggle to capture fine-grained prosody due to dominant semantic representations. In this work, we introduce ProWhistress, an enhanced dual-stream transcription architecture that integrates explicit acoustic modeling to effectively preserve fine-grained acoustic details and significantly improve stress detection robustness. To mitigate the scarcity of Mandarin sentence stress data, we introduce a customized data generation pipeline specifically tailored for Mandarin, yielding the 12-hour synthetic SinoStress-Syn and the 3-hour human-recorded SinoStress-Real benchmark. Evaluations across five English and Mandarin datasets show that ProWhistress significantly surpasses competitive baselines, demonstrating superior accuracy and impressive zero-shot generalization on real-world speech. Code page: `https://github.com/Guhujian/ProWhistress.git`.

Index Terms: Sentence stress detection, automatic speech recognition, computational paralinguistics, Mandarin sentence stress dataset

1. Introduction

Sentence stress plays a central role in spoken communication by signaling information focus and speaker intent [1]. Linguistic theories describe stress both as a rule-governed phonological pattern [2] and as a discourse-level mechanism reflecting communicative intention [3]. Cross-linguistic studies show that stress is primarily realized through variations in pitch and duration [4, 5]. In Mandarin Chinese, stress interacts with lexical tones but is similarly expressed through pitch range expansion and syllable lengthening [6]. Furthermore, its distribution is intricately linked to syntactic structure and information focus [7], a complexity recently highlighted in Mandarin prosody benchmarks [8]. These findings suggest that despite language-specific prosodic constraints, the underlying cross-lingual acoustic regularities provide a strong foundation for computational modeling.

Accurate stress detection benefits downstream applications such as rich transcription in automatic speech recognition (ASR), automated prosodic feedback in computer-assisted language learning (CALL) [9, 10, 11], and expressive speech synthesis in text-to-speech (TTS) [12, 13, 14]. Driven by the demands of these applications, early computational stress detection approaches relied on handcrafted acoustic features, including pitch, duration, and intensity [4, 15, 16], as well as contour and wavelet-based representations [17, 18]. These features were typically embedded in multi-stage pipelines requiring forced alignment and external ASR systems [19, 20], which suffer from error propagation and lack end-to-end optimization [21].

To bypass the dependency on forced alignment, alternative approaches predict prosodic prominence directly from text. While models like BERT capture useful syntactic cues [22, 23], they inherently struggle with complex phenomena such as contrastive focus [24]. To address this, recent methods like ProMode [25] and SupraDoRAL [26] explicitly condition on both acoustic and textual inputs. This underscores that textual representations alone are insufficient, reinforcing the need for multimodal prosodic modeling.

Recent work has therefore shifted toward end-to-end ASR-aided frameworks leveraging self-supervised speech representations. Whistress [27], built upon Whisper [28], introduces an alignment-free architecture for joint transcription and token-level stress prediction. However, it remains constrained by a *semantic–prosodic trade-off*, where deeper layers favor semantic abstraction while attenuating fine-grained acoustic cues [29]. This limitation motivates explicit prosodic modeling within self-supervised speech backbones.

Our contributions are summarized as follows:

- We propose ProWhistress, a dual-stream transcription architecture that integrates explicit acoustic modeling with implicit semantic representations, effectively addressing the semantic–prosodic trade-off in ASR-based sentence stress detection.
- We construct two novel datasets, SinoStress-Syn and SinoStress-Real, filling the data gap for Mandarin sentence stress, and develop a synthetic data generation pipeline guided by a hierarchical annotation protocol tailored for Mandarin.
- Experimental results demonstrate that ProWhistress achieves leading performance across English and Mandarin benchmarks, validating its multilingual applicability and robustness to complex tonal prosody.

2. Mandarin Corpus Construction

The proposed corpora consist of a synthetic training set, SinoStress-Syn, and a human-recorded benchmark, SinoStress-Real. Motivated by the synthetic process of TinyStress-15k [27], SinoStress-Syn is also generated with four key steps: (i) transcription selection, (ii) stress annotation, (iii) speech synthesis, (iv) quality control.

Transcription Selection. To obtain coherent and diverse everyday language, we utilize the large-scale Mandarin AISHELL-3 corpus [30]. Sentences of 10-20 characters are selected as samples. To improve semantic coherence, we restore missing punctuation using the DeepSeek-v3 model.

** indicates the corresponding author.

Stress Annotation. For SinoStress-Syn, we adopt a hierarchical annotation strategy. First, at the sentence level, the primary stress words are identified by instructing DeepSeek-v3 Model to provide 2 or 3 options that reflect the semantic focus of each utterance. Second, at the lexical level, character-level annotation is determined using prosodic rules derived from phonetic studies [6]. Specifically, for modifier-head or verb-complement structures, stress is assigned to the first syllable (left-headed rule), while for general noun phrases and rhythmic units, prominence is assigned to the final syllable (right-headed rule).

Speech Synthesis. Same to [27], speech is synthesized using the Google Text-to-Speech (TTS) API by a randomly chosen voice from a pool of 4 possible voices (2 females and 2 males). For the prosodic features adjustment, stress-related parameters are sampled from Gaussian distributions: pitch elevation from $\mathcal{N}(1.5, 0.3)$ semitones, volume amplification from $\mathcal{N}(3, 0.5)$ dB (clipped to [1.5, 5] dB), and speaking rate from $\mathcal{N}(55, 3)\%$ of the original speed (clipped to [50, 70]%).

Quality Control. To ensure speech-label alignment, we apply Whisper-based filtering, discarding synthesized samples where the transcribed core Chinese character count (excluding punctuation) mismatches the ground truth.

Finally, SinoStress-Syn contains 14,315 samples, totaling approximately 12 hours, and serves as the primary training resource for Mandarin sentence stress detection in this work.

For SinoStress-Real, we recruit 7 graduates (2 females and 5 males) to record texts derived from the AISHELL-3 corpus. The recordings are guided by a character-level acting protocol strictly consistent with our hierarchical annotations. The resulting dataset contains 2,614 samples, totaling approximately 3 hours.

3. Method

To effectively address the semantic-prosodic trade-off in alignment-free stress detection, this section presents the technical framework of ProWhistress.

3.1. Architecture of ProWhistress

As shown in Figure 1, ProWhistress adopts a dual-stream transcription architecture that augments the Whisper backbone with explicit acoustic modeling. It integrates a Whisper implicit stream and a dedicated Acoustic Encoder via a Stress Detection Head with gated residual fusion.

Whisper Backbone (Implicit Stream). Following Whistress [27], we adopt the pre-trained Whisper model as a frozen backbone. It generates transcriptions while providing semantic and prosodic representations for stress modeling.

Acoustic Encoder (Explicit Stream). To model fine-grained prosody, we introduce a trainable Acoustic Encoder that distills acoustic representations from intermediate layers of the frozen Whisper backbone encoder. The module consists of stacked Transformer encoder layers that capture contextual prosodic patterns independent of the decoding process.

Stress Detection Head. The detection head is responsible for integrating the implicit semantic representations with the explicit acoustic features. This dual-stream fusion is executed through four core components:

- **Additional Decoder Block (Implicit Alignment):** Following Whistress, an additional decoder block aligns the final Whisper decoder states with encoder outputs via cross-attention.
- **Bottleneck Cross-Attention (Explicit Injection):** Explicit acoustic features are injected into the linguistic stream via bottleneck cross-attention. Prior to alignment, input features are projected into a low-dimensional latent space ($d_{ctx} = 256$), encouraging discriminative interaction while reducing computational cost. The fused representations are then projected back to the original feature dimension.
- **Gated Residual Fusion:** Let $\mathbf{H}_{imp}$ denote the aligned implicit features from the Additional Decoder Block and $\mathbf{H}_{exp}$ the aligned explicit features from the Bottleneck Cross-Attention. The fused representation is computed as

$$\mathbf{H}_{fused} = \mathbf{H}_{imp} + \sigma(\mathbf{W}_g[\mathbf{H}_{imp}; \mathbf{H}_{exp}] + \mathbf{b}_g) \odot \mathbf{H}_{exp}, \quad (1)$$

  where σ denotes the sigmoid function, $\odot$ represents element-wise multiplication, and $\mathbf{W}_g$, $\mathbf{b}_g$ are learnable parameters. The gate modulates the contribution of the explicit stream before residual addition. The bias $\mathbf{b}_g$ is initialized to a negative value to stabilize early training.
- **FCNN Classifier:** The final prediction is produced by a two-layer Fully Connected Neural Network (FCNN) binary classifier. It takes the fused representations $\mathbf{H}_{fused}$ as input and outputs a stress label for each token.

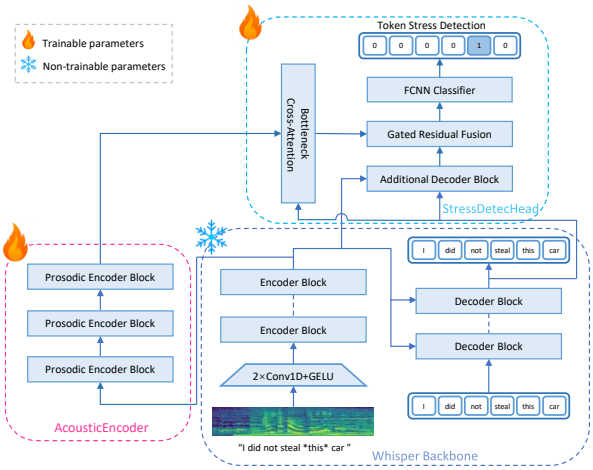


Figure 1: *Architecture of ProWhistress.*

3.2. Training and Inference

Training. We follow the training protocol of Whistress [27]. Length filtering is applied to ensure transcription–label consistency, excluding punctuation from length statistics. We optimize the model using weighted cross-entropy loss to address class imbalance.

Inference. ProWhistress retains the alignment-free property of the baseline. Given raw Mel-spectrograms, the Whisper backbone generates transcriptions autoregressively while the detection head predicts token-level stress probabilities synchronously.

4. Experimental Setup

In this section, we detail the datasets, implementation specifics, and evaluation metrics employed to assess the performance of our proposed model.

4.1. Datasets

TinyStress-15k. A large-scale synthetic dataset for English sentence stress detection, introduced in Whistress [27]. It contains 16,000 speech samples (15,000 for training and 1,000 for testing), with a total duration of approximately 15 hours. In this work, TinyStress-15k serves as the primary benchmark for English experiments.

Expresso. An expressive speech corpus featuring approximately 47 hours of recordings from four speakers (two female, two male) [31], encompassing 7 read speech styles and 25 conversational styles. To ensure a fair comparison with Whistress, we strictly adhere to their filtering strategy: we select samples solely from speakers `ex01` and `ex02`, retaining only those marked with emphasis labels for the test set.

EmphAssess. A synthetically generated dataset where each sentence contains at least one stressed word, accompanied by detailed annotations of transcriptions and stress indices [14]. It comprises 3,652 speech samples derived from 913 unique transcripts. Each transcript is rendered using four distinct voices, with varying stress placements across the voices. These four voices correspond directly to the Expresso speakers, namely `ex01`, `ex02`, `ex03`, and `ex04`.

SinoStress-Syn & SinoStress-Real. As detailed in Section 2, SinoStress-Syn is partitioned into 12,883 training and 1,432 testing samples. For the human-recorded SinoStress-Real benchmark, speech samples from two specific speakers (ex01 and ex02) are strictly reserved for evaluation to ensure speaker independence.

4.2. Baselines

To evaluate the performance of ProWhistress, we compare our results against competitive baselines reported in the Whistress study [27] and other relevant benchmarks.

- **Acoustic-feature-based Baselines:** Following [27], we include Bidirectional LSTM (BLSTM) models evaluated under two conditions: ideal conditions using ground-truth timestamps, denoted as Baseline (+GT alignment), and pipeline settings using timestamps predicted by MFA [32], denoted as Baseline (+MFA). These serve as a reference for traditional pipeline approaches.
- **EmphaClass:** A robust prosodic prominence detection model based on pre-trained XLS-R (wav2vec 2.0) [14]. We cite its reported performance on the Expresso and EmphAssess datasets as the primary benchmark for expressive speech tasks.
- **Whistress:** Serving as the leading alignment-free baseline and the direct predecessor of our work, we compare against the original Whistress architecture configured with the 9th layer of the Whisper decoder, as recommended in [27]. This direct comparison allows us to isolate and demonstrate the efficacy of our proposed explicit prosodic stream.

4.3. Implementation Details

We instantiate the Whisper backbone with `whisper-small` for Mandarin and `whisper-small.en` for English.[1] The Additional Decoder Block utilizes the 12th layer of the Whisper encoder and the 9th layer of the Whisper decoder as inputs. Meanwhile, our 3-layer Acoustic Encoder extracts from the 9th layer of the Whisper encoder. Additionally, the bias of the gating network is initialized to -3.0 to implement the cold-start strategy. Models are trained using AdamW with positive loss weight $w_{pos} = 2.33$ for 2 epochs on large-scale synthetic datasets, extended to 4 epochs on the smaller real datasets to ensure convergence. Metrics are averaged over five random seeds (42–46) on a single RTX 3090 GPU.

[1] Both are official pre-trained checkpoints introduced in [28].

Table 1: *Word-level sentence stress detection results on English benchmarks compared to baselines.*

Evaluated Dataset	Model	Prec	Rec	F1
TS-15K	BLSTM (+GT alignment)	0.862	0.853	0.858
	BLSTM (+MFA)	0.776	0.859	0.815
	Whistress	0.912	0.906	0.909
	ProWhistress	**0.953±0.014**	**0.966±0.011**	**0.959±0.007**
Expresso	EmphaClass	0.569	0.769	0.654
	BLSTM (+MFA) [0-shot]	0.404	0.599	0.482
	Whistress [0-shot]	0.573	0.863	0.689
	ProWhistress [0-shot]	**0.737±0.034**	**0.924±0.012**	**0.820±0.018**
EmphAssess	EmphaClass	0.938	0.94	0.938
	Whistress	0.945	0.942	0.943
	ProWhistress	**0.978±0.010**	**0.984±0.009**	**0.981±0.002**
	BLSTM (+MFA) [0-shot]	0.566	0.609	0.587
	Whistress [0-shot]	0.672	0.98	0.797
	ProWhistress [0-shot]	**0.758±0.033**	**0.974±0.009**	**0.852±0.018**

Note: "[0-shot]" denotes zero-shot evaluation (trained on TinyStress-15k); unmarked models are supervised on the target dataset.

5. Results

5.1. Main Results on English Benchmarks

Table 1 reports word-level stress detection results on English benchmarks.

Supervised Performance. Under in-domain supervised training, ProWhistress achieves the best performance on TinyStress-15k and EmphAssess, consistently outperforming Whistress and BLSTM-based baselines. EmphaClass, which leverages a closed-source dataset closely matching the Expresso distribution, serves as a strong in-domain reference. These results confirm the effectiveness of explicit acoustic modeling when training and testing distributions are aligned.

Zero-shot Generalization. In the zero-shot setting (trained on TinyStress-15k), ProWhistress demonstrates strong generalization to both Expresso and EmphAssess. On Expresso, it substantially surpasses Whistress despite domain shift. Consistent with [27], we do not report supervised results on Expresso, as surpassing zero-shot performance would require significantly more diverse annotated human-recorded speech, which is currently unavailable.

5.2. Cross-lingual Generalization on Mandarin Benchmarks

Table 2 reports the results on Mandarin benchmarks.

Supervised Performance. Under supervised training, ProWhistress consistently surpasses Whistress and BLSTM (+MFA) in overall F1 across both datasets. While BLSTM (+MFA) achieves a slightly higher precision (0.839) on SinoStress-Real, its noticeably lower recall suggests it struggles to capture the more subtle and diverse prosodic variations present in real-world speech. Conversely, ProWhistress maintains a superior precision-recall balance, yielding a robust F1 (0.870) and proving the efficacy of our explicit acoustic modeling for Mandarin.

Zero-shot Generalization. In the zero-shot setting, ProWhistress surpasses all baselines on SinoStress-Real by a large margin. Crucially, while BLSTM (+MFA) degrades significantly when transferring to real data (F1: 0.835 $\rightarrow$ 0.762),

Table 2: *char-level stress detection results on Mandarin benchmarks.*

Evaluated Dataset	Model	Prec	Rec	F1
SinoStress-Syn	BLSTM (+MFA)	0.875±0.008	0.830±0.008	0.852±0.004
	Whistress	0.771±0.044	0.826±0.028	0.797±0.033
	ProWhistress	**0.950±0.009**	**0.965±0.009**	**0.958±0.009**
SinoStress-Real	BLSTM (+MFA)	**0.839±0.015**	0.830±0.022	0.835±0.016
	Whistress	0.734±0.034	0.767±0.046	0.749±0.027
	ProWhistress	0.821±0.030	**0.926±0.031**	**0.870±0.014**
	BLSTM (+MFA) [0-shot]	0.749±0.032	0.776±0.021	0.762±0.022
	Whistress [0-shot]	0.749±0.022	0.806±0.040	0.775±0.016
	ProWhistress [0-shot]	**0.832±0.019**	**0.912±0.033**	**0.869±0.015**

Note: "[0-shot]" rows denote zero-shot evaluation on real data (trained on synthetic SinoStress-Syn); other rows are supervised on the target dataset.

ProWhistress maintains near-identical supervised and zero-shot performance (0.870 vs. 0.869). This contrast proves our architecture learns domain-invariant prosodic representations, ensuring robust Sim-to-Real generalization without overfitting to synthetic acoustics.

5.3. Ablation Studies

To validate the contribution of each component and verify our design choices, we conducted ablation studies on the TinyStress-15k test set. The results are summarized in Table 3.

Table 3: *Ablation studies on the TinyStress-15k test set.*

Configuration	Prec	Rec	F1
Panel A: Contribution of Streams			
Implicit Stream Only (Whistress-Enc12)	0.928±0.008	0.938±0.004	0.933±0.004
Explicit Stream Only (AcousticEncoder)	0.878±0.022	0.900±0.011	0.889±0.015
ProWhistress (Dual Stream)	**0.953±0.014**	**0.966±0.011**	**0.959±0.007**
Panel B: Input Layer for Acoustic Encoder			
Layer 3	0.932±0.032	0.949±0.015	0.940±0.022
Layer 6	0.941±0.018	0.951±0.029	0.946±0.023
Layer 9 (Default)	**0.953±0.014**	**0.966±0.011**	**0.959±0.007**
Layer 12	0.927±0.027	0.925±0.024	0.926±0.023
Panel C: Fusion Mechanism			
Standard Cross-Attention	0.946±0.020	0.954±0.019	0.950±0.018
Bottleneck Cross-Attention (Ours)	**0.953±0.014**	**0.966±0.011**	**0.959±0.007**

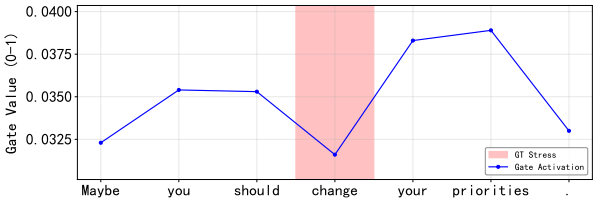
Dual-Stream Architecture (Panel A). We start by analyzing the impact of layer selection in the implicit stream. Using Whisper Encoder Layer 12 (paired with Decoder Layer 9) yields a stronger baseline (F1=0.933) than the original configuration (Encoder and Decoder both at Layer 9). This suggests that Layer 12's superior semantic alignment compensates for its acoustic loss. While the "Explicit Stream Only" model achieves 0.889, ProWhistress combines both to reach peak performance (0.959). This confirms that the streams are complementary: the implicit stream provides robust alignment anchors, while the explicit stream supplements missing fine-grained prosody.

Acoustic Source Selection (Panel B). Having established the dual-stream framework, we identify the optimal explicit source. Performance peaks at Encoder Layer 9, whereas using Layer 12 drops accuracy (0.959 $\rightarrow$ 0.926). This verifies the *semantic-prosodic trade-off*: Layer 12 lacks the acoustic fidelity required for explicit features, necessitating intermediate layers.

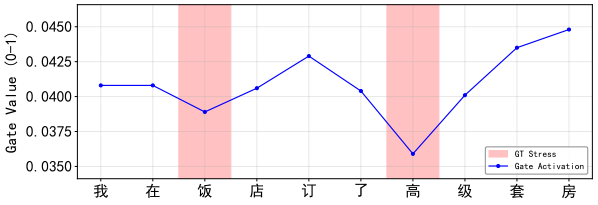
Bottleneck Fusion (Panel C). With the stream components optimized, we finally evaluate how to best integrate them. Replacing our Bottleneck Cross-Attention with standard full-dimension attention degrades performance (0.959 $\rightarrow$ 0.950). This indicates that the bottleneck design functions as an information filter, effectively distilling salient stress cues while discarding irrelevant noise from the explicit stream.

5.4. Visualization Analysis of the Gating Mechanism

To intuitively understand how ProWhistress integrates explicit acoustic cues, we visualized the activation values of the gating mechanism ($\sigma(\cdot)$). As shown in Figure 2, we observe a remarkably consistent Low-Amplitude Inverse Correlation Pattern across both English and Mandarin real-world datasets. Specifically, the gate activation consistently operates in a low-value regime ($\approx$ 0.04), yet exhibits distinct local minima at ground-truth stress positions.



(a) *An English sample from Expresso*



(b) *A Mandarin sample from SinoStress-Real*

Figure 2: *Visualization of gate activation (σ) on real-world datasets. Blue line: gate activation value; Red regions: ground-truth stress.*

This behavior suggests that ProWhistress treats the explicit acoustic stream as a subtle residual correction rather than a dominant signal. Since the frozen Whisper backbone (Layer 12) already provides strong semantic representations, the model does not need to inject explicit features with high magnitude. Instead, it uses small, precise modulations—suppressing injection at semantic peaks (stress) and enhancing it slightly in the background—to refine the decision boundary. The consistency of this "subtle correction" strategy across languages validates that our explicit modeling approach is both parameter-efficient and robust.

6. Conclusion

In this work, we presented ProWhistress, integrating a dual-stream architecture into Whisper via explicit acoustic modeling. This design significantly boosts stress detection performance over baselines while preserving ASR capabilities. To support Mandarin, we introduced SinoStress-Syn and SinoStress-Real, mitigating resource scarcity with a comprehensive benchmark using a linguistically grounded hierarchical protocol. Experiments confirm strong in-domain performance and robust zero-shot generalization to challenging real-world acoustic scenarios.

Beyond empirical gains, our findings underscore the necessity of explicit prosodic modeling to resolve the semantic–prosodic trade-off. Recovering fine-grained acoustic cues proves essential for capturing the interplay between linguistic content and paralinguistic expression, enabling the system to transcend mere transcription to decode communicative intent and enhance interaction-aware understanding. Future work will extend this framework to automated CALL feedback and expressive TTS, advancing prosody-aware speech modeling across diverse languages and paving the way for more natural, intent-sensitive human-computer interaction.

7. Acknowledgments

This work was supported by Chongqing Academy of Science and Technology Basic Research Funding (No. 2022024030029 and No. 2022025030009).

8. Generative AI Use Disclosure

We used generative AI tools only for language polishing and grammar checking. All authors take full responsibility for the content of the paper and have consented to its submission.

9. References

[1] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge University Press, 2008.

[2] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper & Row, 1968.

[3] D. Bolinger, "Accent is predictable (if you're a mind-reader)," *Language*, vol. 48, no. 3, pp. 633–644, 1972.

[4] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[5] V. J. van Heuven, "Acoustic correlates and perceptual cues of word and sentence stress: Towards a cross-linguistic perspective," in *The Study of Word Stress and Accent: Theories, Methods and Data*, R. Goedemans, J. Heinz, and H. van der Hulst, Eds. Cambridge: Cambridge University Press, 2018, pp. 15–59.

[6] Y. Wang, M. Chu, and L. He, "Location of sentence stresses within disyllabic words in mandarin," in *Proc. 15th ICPhS Barcelona*, 2003, pp. 1827–1830.

[7] S. Duanmu, *The Phonology of Standard Chinese*, 2nd ed. Oxford University Press, 2007.

[8] Z. Wang, X. Zhang, X. Jiang, K. Song, and J. Yu, "Can AI Understand Mandarin Speech Prosody? A Framework and Benchmark Showcase," in *Proc. Interspeech*, 2025, pp. 5378–5382.

[9] A. Rosenberg, J. Hirschberg, and K. Manis, "Perception of English prominence by native Mandarin Chinese speakers," in *Proc. Speech Prosody*, 2010, p. 982.

[10] G. Lee, H.-Y. Lee, J. Song, B. Kim, S.-N. Kang, J.-H. Lee, and H.-G. Hwang, "Automatic sentence stress feedback for non-native english learners," *Computer Speech & Language*, vol. 41, pp. 6–20, 2017.

[11] A. Mondal, M. Surtani, A. K. Vuppala, P. Krishnamurthy, and C. Yarra, "ExagTTS: An Approach Towards Controllable Word Stress Incorporated TTS for Exaggerated Synthesized Speech Aiding Second Language Learners," in *Proc. Interspeech*, 2025, pp. 449–453.

[12] Y. Mass, S. Shechtman, M. Mordechay, R. Hoory, O. Sar Shalom, G. Lev, and D. Konopnicki, "Word emphasis prediction for expressive text to speech," in *Proc. Interspeech*, 2018, pp. 2868–2872.

[13] H. Li, L. Qu, J. Hu, and T. Li, "EME-TTS: Unlocking the Emphasis and Emotion Link in Speech Synthesis," in *Proc. Interspeech*, 2025, pp. 4368–4372.

[14] M. de Seyssel, A. d'Avirro, A. Williams, and E. Dupoux, "EmphAssess: A prosodic benchmark on assessing emphasis transfer in speech-to-speech models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 10 186–10 190.

[15] R. Silipo and S. Greenberg, "Prosodic stress revisited: Reassessing the role of fundamental frequency," in *Proceedings of the NIST Speech Transcription Workshop*, 2000.

[16] T. Mishra, V. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Proc. Interspeech*, 2012, pp. 1864–1867.

[17] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[18] S. Kakouros and O. Räsänen, "3pro–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[19] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, 2002.

[20] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[21] S. Stehwien and N. T. Vu, "Prosodic event recognition using convolutional neural networks with context information," in *Proc. Interspeech*, 2017, pp. 2326–2330.

[22] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proc. Nordic Conf. Comput. Linguist. (NoDaLiDa)*, 2019, pp. 281–290. [Online]. Available: https://aclanthology.org/W19-6129/

[23] T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda, and H. Kawai, "GST-BERT-TTS: Prosody Prediction Without Accentual Labels For Multi-Speaker TTS Using BERT With Global Style Tokens," in *Proc. Interspeech*, 2025, pp. 444–448.

[24] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "BERT, can HE predict contrastive focus? predicting and controlling prominence in neural TTS using a language model," in *Proc. Interspeech*, 2022, pp. 3383–3387.

[25] E. Eren, Q. Liu, H. Kim, P. Garrido, and A. Alwan, "ProMode: A Speech Prosody Model Conditioned on Acoustic and Textual Inputs," in *Proc. Interspeech*, 2025, pp. 429–433.

[26] J. Mallela, U. V. Y. S., S. B. Rangavajjala, B. Bhatt, and C. Yarra, "SupraDoRAL: Automatic Word Prominence Detection Using Suprasegmental Dependencies of Representations with Acoustic and Linguistic Context," in *Proc. Interspeech*, 2025, pp. 5773–5777.

[27] I. Yosha, D. Shteyman, and Y. Adi, "WhiStress: Enriching Transcriptions with Sentence Stress Detection," in *Proc. Interspeech*, 2025, pp. 4718–4722.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 28 492–28 518.

[29] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[30] Y. Shi, H. Bu, X. Xu, S. Zhang, and K. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus and the baselines," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 5169–5173.

[31] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani *et al.*, "Expresso: A benchmark and analysis of discrete expressive speech resynthesis," in *Proc. Interspeech*, 2023, pp. 4773–4777.

[32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.