ICCSNT2025

Dialect-Adaptive Conformer Model: Application of Dynamic Parameter Adjustment in Multidialect Speech Recognition

Ying Zhang，Boyi Duan，Jiahao Hui，Xuan Tong

Introduction

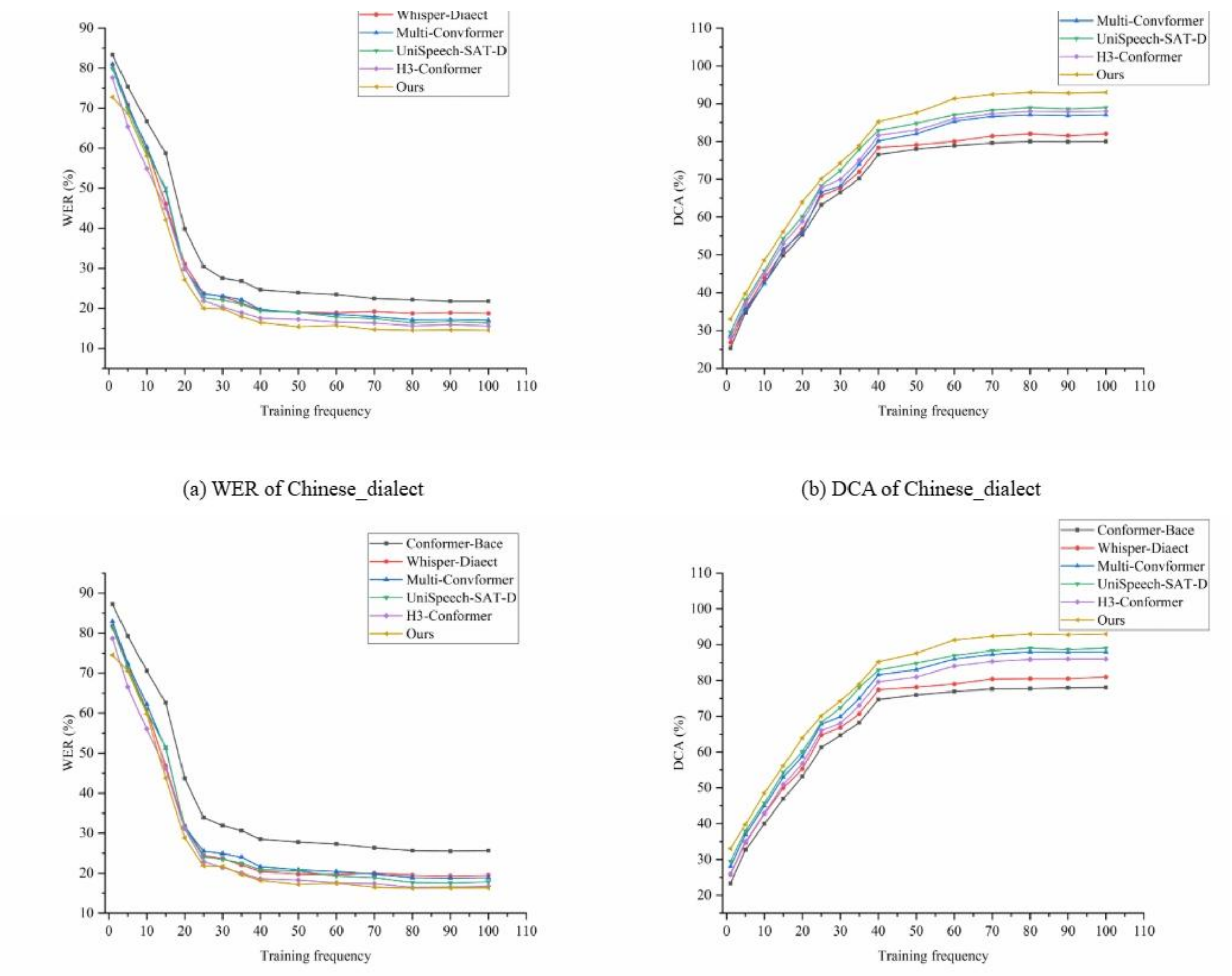
Speech recognition technology has made significant progress in the past decade, particularly in standard languages such as Mandarin and English, where performance has approached human-level accuracy. However, in the task of Multi-Dialect Speech Recognition (MD-ASR), existing systems still face considerable challenges. Acoustic differences across dialects, such as tonal systems, phoneme inventories, and prosodic patterns, as well as the scarcity of data for low-resource dialects, result in a significant increase in the word error rate (WER) of general speech recognition models in dialectal scenarios.The Dynamic Conformer Dialect Recognition Model (Dialect-Adaptive Conformer, DA-Conformer) proposed in this paper is designed based on an end-to-end framework. Through dynamic parameter generation and collaborative optimization of multi-granularity feature interaction, it achieves dialect-adaptive acoustic modeling and decoding.

Method

Our method proposes a novel multi-dialect speech recognition framework based on dynamic Conformer (DA-Conformer). The key innovations of DA-Conformer include:Dynamic Convolution Kernel Generation Module: This module maps dialect embedding vectors to convolutional kernel parameters using a lightweight multi-layer perceptron (MLP), enabling dialect-adaptive local feature extraction and significantly improving the model’s ability to capture dialect-sensitive features such as tone and plosives.

Dialect-Conditioned Biased Attention Mechanism: This mechanism injects dialect-related bias terms into the self-attention computation, dynamically adjusting the attention weight distribution to strengthen the modeling of global dependencies in critical acoustic regions.

Results



(a) WER of Chinese_dialect

(b) DCA of Chinese_dialect

TABLE II
Comparison Results of WER and DCA Across Different Models

Model	**Parameter (M)**	**chinese_dialect**		**ZDialect**	
		WER (%)	**DCA (%)**	**WER (%)**	**DCA (%)**
Conformer-Base	270	21.7	80	25.6	78
Whisper-Dialect	1500	18.7	82	19.5	81
UniSpeech-SAT-D	317	16.3	88	17.8	86
H3-Conformer	240	15.6	89	16.7	89
Multi-Convformer	287	17.0	87	18.9	88
Ours	253	**14.5**	**93**	**16.3**	**93**

Conclusion

1.Proposal of a recognition model for multiple dialects is proposed, built upon the Conformer model.

2.Dialect embedding vectors are utilized to guide the generation of frequency-domain sensitive convolution kernels, enhancing the ability to extract dialect-specific local features such as tone and plosive sounds.

3.A learnable bias matrix is introduced to correct attention weight distributions, strengthening the modeling of tonal boundaries and continuous pitch variation regions.