[Pulse: Chinese Speech2IPA with High-Quality Annotation]

Authors: [Qingming Li, Youran Wang, Ruiyan Sun]
Student IDs: [224040228, 224040259, 224040284]
Email: [224040228@link.cuhk.edu.cn]

Introduction

In this paper, we present **a novel Speech2IPA model——Pulse**, designed for Chinese. *The model integrates **semi - automatic annotations** with a tone - optimized architecture.*

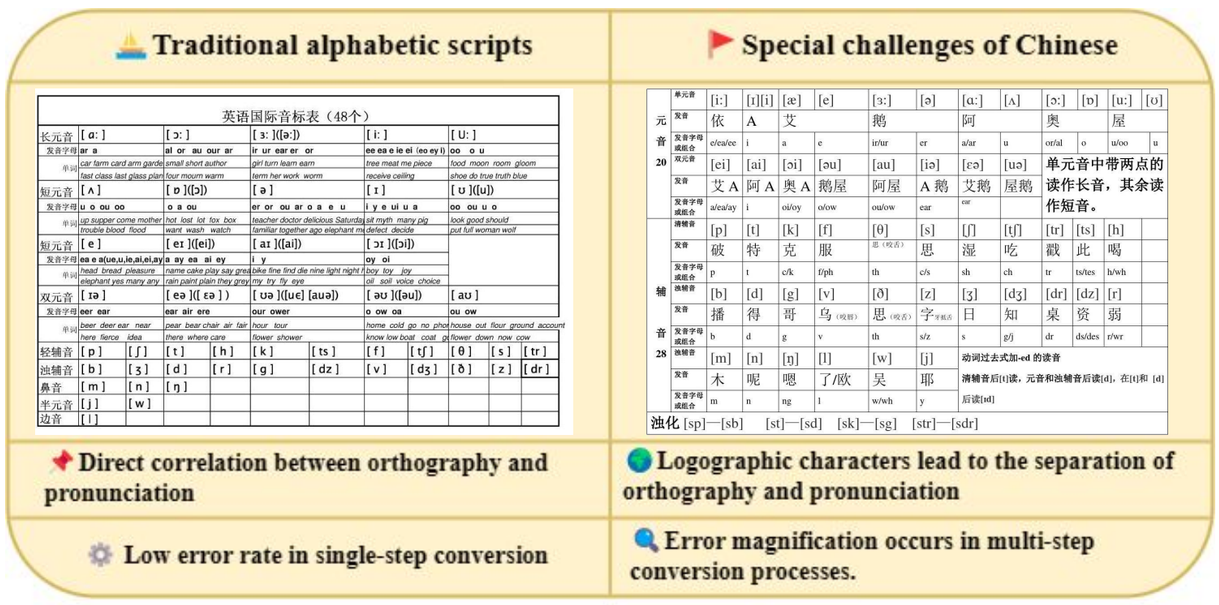
Existing automatic **Speech2IPA** systems, despite their remarkable progress, still fall short in handling the tonal complexity and phonological nuances of Chinese. For instance, Wav2Vec2Phoneme relies on automated G2P tools that often cause systematic errors in tone marking and contextual variation processing.

*Based on the **wav2vec 2.0** framework, we introduce **joint tone - phoneme modeling** via a self - attention mechanism.* This enables the explicit learning of interactions between segmental phonemes and suprasegmental tones.

Experiments show that our Chinese - specific Speech2IPA model, with high - quality data labeling and tone - sensitive design, breaks through the performance limitations of generic models. Our model with 2k training samples (15.6% PFER) performed the best among the tested models. In particular, it outperformed both of Allosaurus (20.9% PFER) and Wav2Vec2Phoneme (18.3% PFER). See our Github in https://github.com/leeqingming-BOL/mychinesemultipa.

Motivation

I. Core Motivation of the Project



II. Technical Motivation

🔬 *Data Efficiency Comparison:* Training Effect with Small Datasets (e.g. 1k vs. 50k in traditional schemes)

🔍 *Evaluation Method Upgrades:* Traditional CER (Character Error Rate) vs. PER (Phone Error Rate) vs. PFER (Phone Feature Error Rate).This method not only captures the overall phonetic accuracy but also provides a more detailed and linguistically informed evaluation.

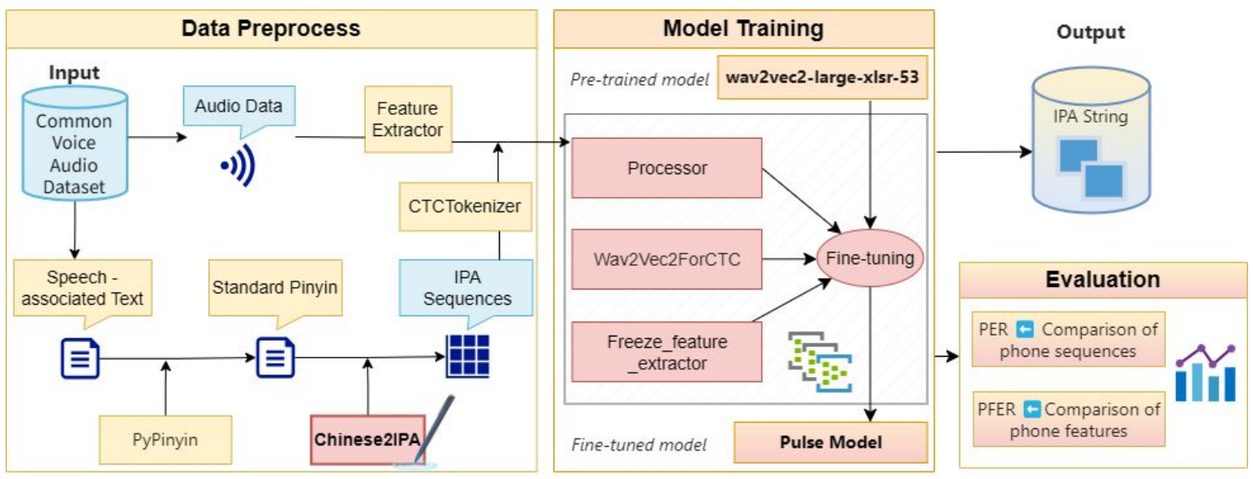
III. Value Extension and Application Scenarios

🌍 *Academic Research:* Dialect recording → IPA transcription → database construction. Conducting dialect recordings is the first step, which then undergo IPA transcription to accurately capture the phonetic nuances of various dialects.

🚀 *Technological Innovation:* We have achieved the direct end-to-end conversion from Chinese to IPA and the construction of a Chinese IPA precise annotation tool. It has verified the crucial role of a high-quality precise annotation dataset in model training!

🏫 *Practical Application:* For language learners, IPA - annotated pronunciation examples serve as an invaluable learning aid, helping them understand the correct articulation of Chinese sounds.

Methodology



I. Multidimensional Data Construction

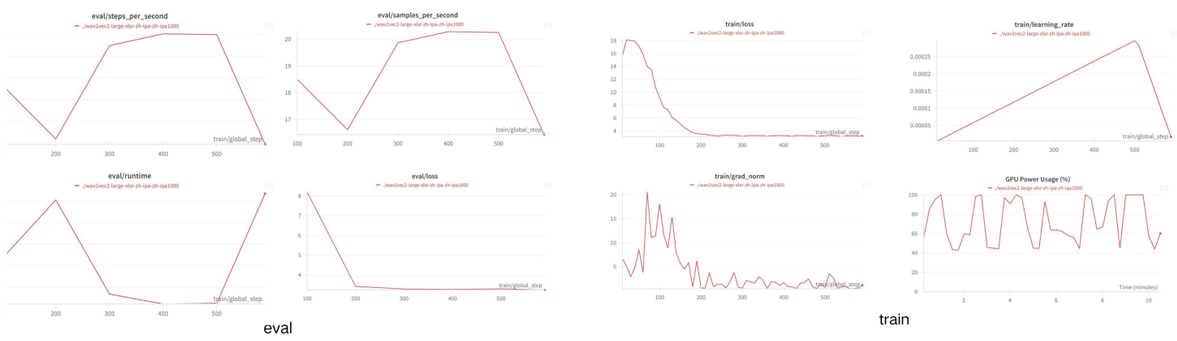
- *Source of Speech Data*
  Based on the method of obtaining speech data from the CommonVoice dataset, high-quality Mandarin Chinese speech segments are selected as the basic input.Through preprocessing steps such as removing low-quality audio, a speech dataset that meets the requirements of model training is constructed.
- *Hierarchical Generation of IPA Sequences*
  The **pypinyin** tool is used to convert the text corresponding to the speech in **CommonVoice** into standard pinyin, ensuring an accurate mapping from graphemes to phonemes. Subsequently, through the self-developed **Chinese2IPA** tool, the pinyin is further converted into IPA sequences, achieving a hierarchical conversion from text to phonetic symbols.

II. The Cross-Lingual Model Fine-Tuning Framework

In the model training stage, the wav2vec2CTC technical framework is proposed. Based on the pre-trained **wav2vec2-large-xlsr-53** model, the **Connectionist Temporal Classification (CTC)** loss function is adopted. Through the fine-tuning strategy, the output target of the model is shifted from multi-lingual phoneme prediction to Mandarin Chinese IPA sequence generation.

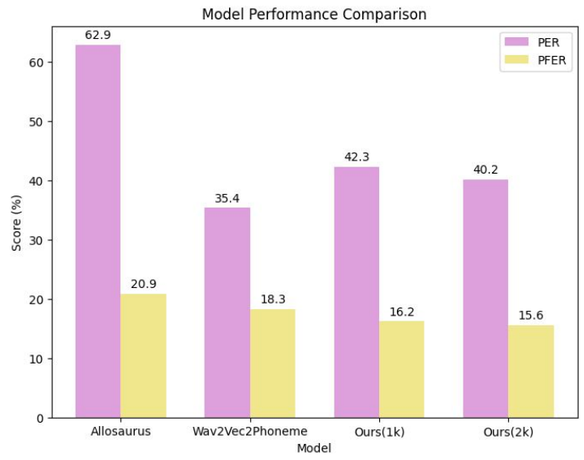
Results

I. Training Result



II. Evaluation Result

Metric	Model	Score
PER	Allosaurus	62.90%
	Wav2Vec2Phoneme	35.40%
	Ours(1k)	42.30%
	Ours(2k)	40.20%
PFER	Allosaurus	20.90%
	Wav2Vec2Phoneme	18.30%
	Ours(1k)	16.20%
	Ours(2k)	15.60%



We put more emphasis on the **PFER-based** comparison than PER in this study because PFER is more representative of the transcription accuracy.