

[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

-	Tradiu	onal alp	паре	uc	SCI	ıpı	5				эþ	eci	al cł	iane	nge	25 0	ı C	1111	ies	e	
	0000								单元音	[i:]	[1][i]	[æ]	[e]	[3:]	[e]	[a:]	[A]	[:c]	[\alpha]	[u:]	[ʊ]
		吾国际音标表(48			Torono sor			元	发音	依	A	艾		鹅		阿		奥		屋	
元音 [a:]	[ɔ:]	[3:]([ə:])	[i:]		[U:]			音	发音字母	e/ea/ee	i	a	e	ir/ur	er	a/ar	u	or/al	o	u/oo	u
发音字母 ar a	al or au our ar	ir ur ear er or	ee ea e ie			n room al		20	成组合 双元音	F - 23	F - 11	F - 23	F1	F1	F1 - 1	F	Francis I	24 -	: 77: rH	the are d	: 44
	garde small snort autnor s plan four mourn warm	term her work worm	receive ceil	- P	-	on room gi rue truth blu		20		[ei]	[ai]	[ic]	[əu]	[au]	[iə]	[e3]	[uə]	2.10		带两点	3.60(2)
豆元音 [^]	([c])[a]	[ə]	[1]		[ʊ]([u	u])			发音	艾Α	阿 A	奥 A	鹅屋	阿屋	A 鹅	艾鹅	屋鹅	读作	长音	,其分	读
发音字母 u o ou oo	o a ou	er or ou ar o a e u	i y e ui ı	ıa	00 OU I	u o			发音字母或组合	a/ea/ay	i	oi/oy	o/ow	ou/ow	ear	ear		作短	i音。		
ரு [ii] up supper come n	other hot lost lot fox box	teacher doctor delicious Sa familiar together ago elepha			look good				清精音	[n]	[t]	[k]	[f]	[θ]	[s]	[f]	[tʃ]	[tr]	[ts]	[h]	
記元音 [e]	[eɪ]([ei])	[ar]([ai])	c])[rc]		put ruit we	oman won			发音	[p]			~ ~	思(咬舌)		-0	_				
发音字母 ea e a(ue,u,ie,ai	2 - 2 - 27	i y	oy oi	.,	1					破	特	克	服		思	湿	吃	戳	此	喝	
作词 head bread pleas	ure name cake play say g	rea bike fine find die nine light r	ight r boy toy je						发音字母或组合	p	t	c/k	f/ph	th	c/s	sh	ch	tr	ts/tes	h/wh	
元音 [Iə]	[eə]([ɛə])	[və]([uɛ] [auə])	[əʊ]([a	-	[au]			辅	油辅音	[b]	[d]	[g]	[v]	[ð]	[z]	[3]	[d ₃]	[dr]	[dz]	[r]	
发音字母 eer ear	ear air ere	our ower	o ow oa	-1/	on ow				发音	播	得	哥	点 (成居)	思(咬舌)	字明語	Ε	知	桌	资	弱	
m in beer deer ear ne	ar pear bear chair air f	air hour tour				it flour gro	und account		发音字母	эщ	ার্য	нJ	一月 (咬唇)	2-167.65/		Н	6.00	200			
here fierce idea	there where care	flower shower	know low b		100000000000000000000000000000000000000	wn now c	100	音	或组合	b	d	g	v	th	s/z	S	g/j	dr	ds/des	r/wr	
至辅音 [p] [ʃ		[k] [ts]	[f]	[tʃ]	[0]		[tr]	28	28 ^{強縮音} [m] [n] [ŋ] [l] [w] [j] 动词过去式						式加-ed 的	加-ed 的读音					
性辅音 [b] [3		[g] [dz]	[v]	[dʒ]	[6]	[z]	[dr]		发音	木	呢	嗯	了/欧	吴	耶	清輔音后[t]读,元音和浊輔音后读[d],在[t]和 [d]			I (d)		
h音 [m] [n									发音字母		-76	-	1 / 15/4								
[≚] 元音 [j] [w	1					1			或组合	m	n	ng	1	w/wh	У	后读[ɪd]					
位音 [1]					_			油	化 [sp	—[sb] [s	t]—[s	d] [sk]-	-[sg] [str]—[sdr]					
♣ Direct correlation between orthography and pronunciation					Logographic characters lead to the separation of orthography and pronunciation																
Low error rate in single-step conversion					Error magnification occurs in multi-step conversion processes.																

II. Technical Motivation

Mata 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.

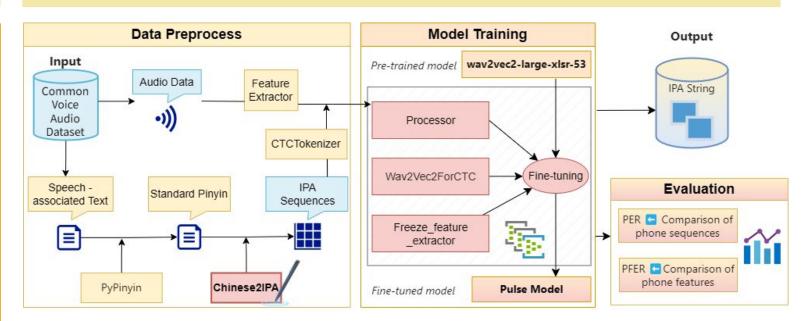
III. Value Extension and Application Scenarios

Academic Research: Dialect recording \rightarrow IPA transcription \rightarrow 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

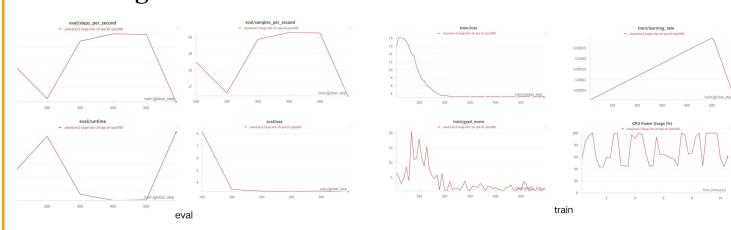
The **pypinyin** tool is used to convert the text corresponding to the speech in **CommonVoice** into standard pinyin, ensuring an accurate mapping from graphemes to phonemes. Subsequently, through the self-developed **Chinese2IPA** tool, the pinyin is further converted into IPA sequences, achieving 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.

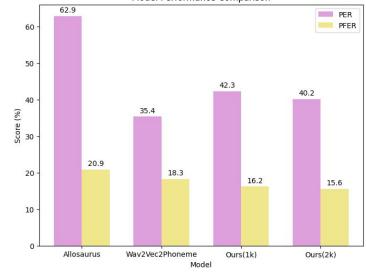
Results

I. Training Result



II. Evaluation Result

Metric	Model	Score		62.9	
Wetric	Wodei	Score	60 -		
PER	Allosaurus	62.90%			
	Wav2Vec2Phoneme	35.40%	50 -		
	Ours(1k)	42.30%	40 - 		
	Ours(2k)	40.20%	Score (%)		
PFER	Allosaurus	20.90%	20 -		2
	Wav2Vec2Phoneme	18.30%	20		
TTEK	Ours(1k)	16.20%	10 -		
	Ours(2k)	15.60%	∟₀	Allosa	ur



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