www.sciengine.com/CSD
ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2026.0038.zh
文献 DOI：
10.11922/11-6035.csd.2026.0038.zh
数据 DOI：
10.57760/sciencedb.j00001.01617

文献分类：计算机科学技术

收稿日期：2026-03-15
修回日期：2026-04-28
录用日期：2026-06-02
发表日期：2026-06-11

基于超声舌成像的汉语普通话多模态发音生理数据集

张金溪[1]，于洪志[2,3*]，张晶[1]

1. 兰州财经大学，商务传媒学院，兰州 730020
2. 西北民族大学，语言与文化计算教育部重点实验室，兰州 730030
3. 西北民族大学，甘肃省民族语言文化智能信息处理重点实验室，兰州 730030

摘要： 发音生理数据是汉语普通话语音研究、语音工程的核心基础。当前，汉语普通话的多模态发音生理数据集存在覆盖不全面、模态单一且不同步等问题，难以满足精准化研究需求。针对这一问题，本研究构建了一套基于超声舌成像的汉语普通话多模态发音生理数据集，弥补现有数据集在多维度发音生理信息融合上的不足。数据采集遵循语音学实验规范，选取一名汉语普通话发音标准的女研究生为发音人，采用 B 超设备动态记录发音时的舌头运动，同步采集语音音频和嘴唇视频数据。本数据集涵盖声韵母搭配组合及四个声调条件下的常用合法发音单元，形成了 1024 个完整发音单元；多模态数据包含文本语料、语音音频、嘴唇视频、超声舌成像 4 部分，可全面反映发音过程中的生理运动特征与声学表现。数据质控阶段采用人工校验与机器筛选相结合的方式，剔除发音不标准、图像模糊、音频失真等无效数据，最终得到高质量的数据集。本数据集不仅为汉语普通话发音生理机制、声调变化规律及二语习得等基础研究提供数据支撑，还可应用于语音合成与识别、言语障碍诊断与康复、发音机制建模及人工智能语音模型训练等实际场景，同时为跨语言发音生理对比研究提供参考。

关键词： 超声舌成像；汉语普通话；多模态数据集；发音生理；人工智能；语音工程

数据库（集）基本信息简介

数据库（集）名称	基于超声舌成像的汉语普通话多模态发音生理数据集
数据通信作者	张金溪（582361484@qq.com）
数据作者	张金溪、于洪志、张晶
数据时间范围	2026年
地理区域	甘肃省，兰州市
数据量	41.33 GB（解压后290 GB）
数据格式	*.avi, *.xml, *.wav, *.pdf
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01617
基金项目	国家自然科学基金（62366045、62566057）；国家社会科学基金（22BYY165）；甘肃省中央引导地方科技发展资金项目（25ZYJA034）；陇原青年英才项目“甘肃省内多民族语言共演化特征研究”；兰州市科技计划项目（2025-2-42）。
数据库（集）组成	发音人的数据文件包括：（1）1. text文本语料。包含5类不同发音特征的PDF格式文件，分别是：1-318第一声（阴平）文本语料.pdf，319-564第二声（阳平）文本语料.pdf，565-866第三声（上声）文

* 论文通信作者
于洪志：yhz@xbmu.cn

数据库（集）组成	本语料.pdf，867-1200第四声（去声）文本语料.pdf，1201-1204特殊单韵母文本语料.pdf。数据量共计1.54 MB。（2）audio语音音频。包含5类不同发音特征的RAR压缩文件，压缩文件中是WAV格式音频，分别是：1-318第一声（阴平）语音音频.rar，319-564第二声（阳平）语音音频.rar，565-866第三声（上声）语音音频.rar，867-1200第四声（去声）语音音频.rar，1201-1204特殊单韵母语音音频.rar。数据量共计180 MB。（3）video嘴唇视频。包含5类不同发音特征的RAR压缩文件，压缩文件中是AVI格式视频，分别是：1-318第一声（阴平）嘴唇视频.rar，319-564第二声（阳平）嘴唇视频.rar，565-866第三声（上声）嘴唇视频.rar，867-1200第四声（去声）嘴唇视频.rar，1201-1204特殊单韵母嘴唇视频.rar。数据量共计639 MB。（4）ultrasound超声舌成像。包含5类不同发音特征的RAR压缩文件，压缩文件中是AVI格式超声和XML格式文件，分别是：1-318第一声（阴平）超声舌成像.rar，319-564第二声（阳平）超声舌成像.rar，565-866第三声（上声）超声舌成像.rar，867-1200第四声（去声）超声舌成像.rar，1201-1204特殊单韵母超声舌成像.rar。数据量共计40.6GB（解压后289 GB）。

引　言

超声舌成像技术具有无创、实时、便携等优势，它能够以较高时间分辨率动态记录发音过程中舌头的运动轨迹，现在已经成为探究发音生理机制的重要手段和方法[1]。自 20 世纪 60 年代起，超声成像技术就开始应用于语音科学研究[2]，它为揭示元音舌位、辅音成阻方式及协同发音现象等提供了直接的观察依据。汉语（以普通话为代表）是我国使用人数最多的语言，也是世界上使用人数最多的语言。汉语普通话的声调系统和音节结构较为复杂，在发音生理研究上存在一些困难和挑战[3]。如果能够基于多模态发音生理数据集深入解析汉语普通话发音过程中舌运动与声学输出之间的映射关系，那么这项研究不仅有助于完善言语产生理论，也可为二语语音习得、言语障碍康复、基于生理信号的语音合成与识别以及人工智能语音模型训练提供关键支撑[4]。

发音生理数据库的建设是推动该领域发展的基础性工作。国际上，Haskins Laboratories 较早开展了基于电磁发音仪（EMA）的英语发音数据库建设[5]，为后续研究提供了重要参照。随着成像技术的进步，USC-TIMIT 数据库[6]结合实时磁共振与声学数据，实现了对声道动态的高分辨率记录，但设备昂贵且普及性有限。在超声技术应用方面，主流的超声数据库 UltraSuite 数据库[7]、TAL 数据库[8]、SSR7000 数据库[9]等多聚焦英语语系。国内相关研究起步相对较晚，早期主要采用电子腭位仪（EPG）、EMA 等技术构建汉语普通话发音数据库[10]，为后续研究奠定了基础。近年来，基于超声的汉语普通话研究逐步展开，学者们采集了汉语普通话单元音的超声数据，对元音发音的舌体运动和舌位特征进行了探究[11]，也有学者构建了汉语普通话的超声和音频数据集 AUSpeech[12]，但无嘴唇视频。本课题组前期也构建了藏语拉萨话元音的超声舌成像数据集[13]，为跨语言发音对比研究积累了经验。

然而，综合分析现有数据库可以发现，当前面向汉语普通话的发音生理数据集仍存在三个主要局限：其一，音系覆盖范围有限，多数研究聚焦于元音或部分辅音，尚缺少面向汉语普通话全部声母类别、主要韵母类别及四个声调条件下常用合法发音单元的系统性多模态数据集；其二，模态集成度不足，已有数据集多为单一或双模态（如超声+音频），缺少对唇部运动的同步记录，难以全面反映发音器官的协同运动；其三，数据同步性与质量控制标准不一，部分数据未实现多模态信号的精确对齐，且缺乏系统性质控流程。上述问题导致现有数据难以满足精准化、系统化的研究需求，

制约了汉语普通话发音生理研究的深入发展。因此，构建一套音系覆盖完整、多模态同步采集、质量控制严格的高质量发音生理数据集，已成为推进该领域研究的迫切需求。

基于上述问题，在数据集的设计和构建上突出以下优势与特点。一是依据《汉语拼音方案》和汉语普通话音节结构规则，系统设计了覆盖汉语普通话声母类别、韵母类别及声调类型的发音文本语料。语料首先整理出 393 个基础声韵组合，包括三拼音节韵母（如 ia、uan、iong 等），并结合四个声调形成候选声韵调单元；随后依据本研究筛选规则进行筛选，最终形成 1024 个有效发音单元，这些发音文本语料为研究声韵调协同发音机制提供了数据支撑。二是实现了多模态数据信息的同步，采用了硬件触发与软件校准相结合的方式同步采集超声舌成像（探头频率 5.0 MHz）、语音音频（采样率 44.1 kHz）以及嘴唇视频（帧率 30 fps）数据，这些多模态数据能够完整记录舌头运动、嘴唇口型变化与声学输出之间的动态对应关系。三是建立了系统化的质量控制体系，在数据采集阶段严格筛选发音标准、无言语障碍的健康发音人；在数据处理阶段剔除了发音不标准、图像模糊、音频失真等无效样本，并对有问题的数据及时补录。与课题组构建的藏语拉萨话元音超声舌成像数据集[13]相比，本数据集增加嘴唇视频、声调、声母等信息，实现了多模态信息的深度融合，同时在模态集成度、音系覆盖面和质控标准化方面均有了显著提升，这些不仅可为汉语普通话发音机制研究提供基础资源，也可为跨语言对比、语音工程技术开发及人工智能语音模型训练提供高质量数据支撑[14]。

1 数据采集和处理方法

1.1 发音人选取

本数据集严格按照汉语普通话音系规范设计，选取一名以汉语普通话为母语的女性在读硕士研究生作为发音人，年龄 24 岁，无口腔、听力相关疾病史，汉语普通话水平达一级乙等标准，发音清晰标准无方言口音，符合语音学实验发音人选取要求。发音人对实验目的和流程了解后签署知情同意书，实验方案符合伦理规范要求。发音人保持固定坐姿确保发音过程中身体与头部无大幅晃动。

此处选择一名发音人采集数据主要是基于以下考虑。首先，在语音学实验研究中，采用一名标准发音人进行系统数据采集并进行个案分析是一种常用和可以接受的方法。其次，发音人的选取要综合考量，既要有语言学或实验语音学背景，又要对实验设备较为熟悉，同时汉语普通话水平等级要高，这样采集到的数据质量较为可靠。再次，同步采集语音音频、嘴唇视频、超声舌成像的工作量大，耗时长，对发音人要求较高。发音人前期需要投入时间和精力熟练掌握文本语料和超声探头的使用，采集数据时要长时间保持坐姿和头部稳定，这些对发音人构成了挑战。最后，尤为重要的是，超声舌成像受个体差异因素（身高、体重、年龄、性别、母语或方言、受教育程度等）影响很大，舌位前后、高低位置存在很大变数，对特定语音问题进行的超声舌成像个案分析是较为妥当的。

1.2 语料设计

语料设计基于《汉语拼音方案》系统梳理了所有声母与韵母的合法搭配组合。根据现代汉语音节结构规则，共形成 393 个有实际意义的声韵母搭配组合（即基础音节），如图 1 所示。每个基础音节与四个声调（阴平、阳平、上声、去声）组合，可以形成 1572 个独立发音单元，剔除了一些无汉字对应音节单元，最终筛选出了 1024 个发音单元。

韵母＼声母	b	p	m	f	d	t	n	l	g	k	h	j	q	x	zh	ch	sh	r	z	c	s	y	w
a	ba	pa	ma	fa	da	ta	na	la	ga	ka	ha				zha	cha	sha		za	ca	sa	ya	wa
o	bo	po	mo	fo				lo														yo	wo
e			me		de	te	ne	le	ge	ke	he				zhe	che	she	re	ze	ce	se	ye	
i	bi	pi	mi		di	ti	ni	li				ji	qi	xi	zhi	chi	shi	ri	zi	ci	si	yi	
u	bu	pu	mu	fu	du	tu	nu	lu	gu	ku	hu				zhu	chu	shu	ru	zu	cu	su		wu
ü							nü	lü				ju	qu	xu								yu	
ai	bai	pai	mai		dai	tai	nai	lai															wai
ei	bei	pei	mei	fei	dei		nei	lei	gei	kei	hei				zhei		shei		zei				wei
ui					dui	tui			gui	kui	hui				zhui	chui	shui	rui	zui	cui	sui		
ao	bao	pao	mao		dao	tao	nao	lao	gao	kao	hao				zhao	chao	shao	rao	zao	cao	sao	yao	
ou		pou	mou	fou	dou	tou	nou	lou	gou	kou	hou				zhou	chou	shou	rou	zou	cou	sou	you	
iu			miu		diu		niu	liu				jiu	qiu	xiu									
ie	bie	pie	mie		die	tie	nie	lie				jie	qie	xie								ye	
üe							nüe	lüe				jue	que	xue								yue	
er																							
an	ban	pan	man	fan	dan	tan	nan	lan	gan	kan	han				zhan	chan	shan	ran	zan	can	san	yan	wan
en	ben	pen	men	fen	den		nen		gen	ken	hen				zhen	chen	shen	ren	zen	cen	sen		wen
in	bin	pin	min				nin	lin				jin	qin	xin								yin	
un					dun	tun		lun	gun	kun	hun				zhun	chun	shun	run	zun	cun	sun		
ün												jun	qun	xun								yun	
ang	bang	pang	mang	fang	dang	tang	nang	lang	gang	kang	hang				zhang	chang	shang	rang	zang	cang	sang	yang	wang
eng	beng	peng	meng	feng	deng	teng	neng	leng	geng	keng	heng				zheng	cheng	sheng	reng	zeng	ceng	seng		weng
ing	bing	ping	ming		ding	ting	ning	ling				jing	qing	xing								ying	
ong					dong	tong	nong	long	gong	kong	hong	jiong	qiong	xiong	zhong	chong		rong	zong	cong	song	yong	
ia					dia			lia				jia	qia	xia									
ian	bian	pian	mian		dian	tian	nian	lian				jian	qian	xian									
iang							niang	liang				jiang	qiang	xiang									
iao	biao	piao	miao		diao	tiao	niao	liao				jiao	qiao	xiao									
iong												jiong	qiong	xiong									
ua									gua	kua	hua				zhua	chua	shua	rua					
uai									guai	kuai	huai				zhuai	chuai	shuai						
uan					duan	tuan	nuan	luan	guan	kuan	huan				zhuan	chuan	shuan	ruan	zuan	cuan	suan		
uang									guang	kuang	huang				zhuang	chuang	shuang						
uo					duo	tuo	nuo	luo	guo	kuo	huo				zhuo	chuo	shuo	ruo	zuo	cuo	suo		
üan												juan	quan	xuan								yuan	

图 1　汉语普通话声韵母搭配组合

Figure 1　Combination of initials and finals in Mandarin Chinese

筛选规则以中国社会科学院语言研究所编修的《新华字典（第 12 版）》为核心依据，剔除了无对应常用汉字的纯拼音音节。所有筛选均保留汉语普通话合法声韵调组合，不删减音系结构要素。最终筛选出 1024 个发音单元，其中第一声 318 个、第二声 246 个、第三声 302 个、第四声 334 个、特殊单韵母 4 个，23 个声母、24 个韵母全覆盖，含单韵母、复韵母、鼻韵母、特殊韵母全类别，声韵母合法搭配覆盖 100%，声调分布均衡，无某类音系要素缺失，完整覆盖汉语普通话核心音系分布。

部分发音文本语料如表 1 所示。所有发音单元均为汉语普通话中实际存在的合法音节，涵盖全部声母、韵母及声调组合，实现对汉语普通话音系发音特征的覆盖。文本语料以三号黑体字呈现于高清显示器软件中，比如“1- bā八”，此时发音人可以同时看到 ID 号、含声调的拼音以及常用汉字。文本语料记录每个发音单元的声母、韵母、声调标注信息，为每个发音单元分配唯一 ID，建立与其余三类模态数据的一一映射关系。其中 ID 为 1201-1204 的语料为特殊单韵母，分别是ér（儿）、ěr（耳）、èr（二）、ǎi（欸）。

表 1　汉语普通话发音文本语料

Table 1　Mandarin Chinese pronunciation text corpus

ID	声母	韵母	拼音	声调	拼音声调组合	常用汉字
1	b	a	ba	第一声（阴平）	bā	八
2	b	o	bo	第一声（阴平）	bō	波
3	b	i	bi	第一声（阴平）	bī	逼
4	b	ai	bai	第一声（阴平）	bāi	掰
5	b	ei	bei	第一声（阴平）	bēi	杯
6	b	ao	bao	第一声（阴平）	bāo	包
7	b	ie	bie	第一声（阴平）	biē	憋
8	b	an	ban	第一声（阴平）	bān	班
9	b	en	ben	第一声（阴平）	bēn	奔

ID	声母	韵母	拼音	声调	拼音声调组合	常用汉字
10	b	in	bin	第一声（阴平）	bīn	宾
11	b	ang	bang	第一声（阴平）	bāng	帮
12	b	eng	beng	第一声（阴平）	bēng	崩
13	b	ing	bing	第一声（阴平）	bīng	冰
14	b	iɑn	bian	第一声（阴平）	biān	边
15	b	iɑo	biao	第一声（阴平）	biāo	标
……	……	……	……	……	……	……
327	p	a	pa	第二声（阳平）	pá	爬
328	p	o	po	第二声（阳平）	pó	婆
329	p	i	pi	第二声（阳平）	pí	皮
330	p	u	pu	第二声（阳平）	pú	仆
331	p	ai	pai	第二声（阳平）	pái	牌
332	p	ei	pei	第二声（阳平）	péi	陪
333	p	ao	pao	第二声（阳平）	páo	刨
334	p	an	pan	第二声（阳平）	pán	盘
335	p	en	pen	第二声（阳平）	pén	盆
336	p	in	pin	第二声（阳平）	pín	贫
337	p	ang	pang	第二声（阳平）	páng	庞
338	p	eng	peng	第二声（阳平）	péng	朋
339	p	ing	ping	第二声（阳平）	píng	平
340	p	iɑn	pian	第二声（阳平）	pián	便
341	p	iɑo	piao	第二声（阳平）	piáo	嫖
……	……	……	……	……	……	……
588	m	a	ma	第三声（上声）	mǎ	马
589	m	o	mo	第三声（上声）	mǒ	抹
590	m	i	mi	第三声（上声）	mǐ	米
591	m	u	mu	第三声（上声）	mǔ	母
592	m	ai	mai	第三声（上声）	mǎi	买
593	m	ei	mei	第三声（上声）	měi	美
594	m	ao	mao	第三声（上声）	mǎo	铆
595	m	ou	mou	第三声（上声）	mǒu	某
596	m	an	man	第三声（上声）	mǎn	满
597	m	in	min	第三声（上声）	mǐn	敏
598	m	ang	mang	第三声（上声）	mǎng	莽
599	m	eng	meng	第三声（上声）	měng	猛
600	m	iɑn	mian	第三声（上声）	miǎn	免
601	m	iɑo	miao	第三声（上声）	miǎo	秒
……	……	……	……	……	……	……
911	f	a	fa	第四声（去声）	fà	发
912	f	u	fu	第四声（去声）	fù	负
913	f	ei	fei	第四声（去声）	fèi	费
914	f	an	fan	第四声（去声）	fàn	饭
915	f	en	fen	第四声（去声）	fèn	份
916	f	ang	fang	第四声（去声）	fàng	放
917	f	eng	feng	第四声（去声）	fèng	凤
……	……	……	……	……	……	……

注：本表仅展示部分样本

1.3 采集环境、采集设备与系统配置

所有数据采集在专业隔音消音实验室内完成，室内背景噪声≤30 dB (A)，环境温度与湿度保持恒定，避免环境因素对采集结果的干扰。数据采集系统由超声成像设备、音频采集设备和视频采集设备三部分组成，三路信号通过外置同步触发器实现各采集设备的同步触发，同时在音频轨道嵌入同步脉冲信号，为后续多模态数据对齐提供统一时间基准，解决现有数据集同步性不足的核心问题。

（1）超声成像设备与探头固定装置。超声采用加拿大 Ultrasonix Medical 公司研制的 SonixOP 成像系统，如图 2 左图所示，配备 5.0 MHz 的 EC9-5 型凸阵探头。探头固定于发音人颏下正中位置，采用矢状面扫描方式记录发音过程中舌体正中矢状面的动态运动。超声成像参数设置为：采集深度 10.0 cm，扇区 100%，帧频 33 Hz，动态范围 80 dB，余辉 6，色度图指数 1，清晰度 High，柔化关，放大 100%，时间增益补偿根据个体情况进行优化以确保舌面轮廓清晰。超声视频通过 S-Video 端子输出至视频采集卡，采集帧率为 25 fps，单帧图像分辨率为 864×768 像素，以 AVI 格式保存。为减少头部运动对超声图像稳定性的干扰，采用定制式机械固定支架将超声探头固定于可调节悬臂系统，如图 2 右图所示。发音人头部倚靠于椅背头托，探头支架与头托相对固定，可随头部微小运动同步移动，最大程度确保探头与下颌的相对位置稳定。

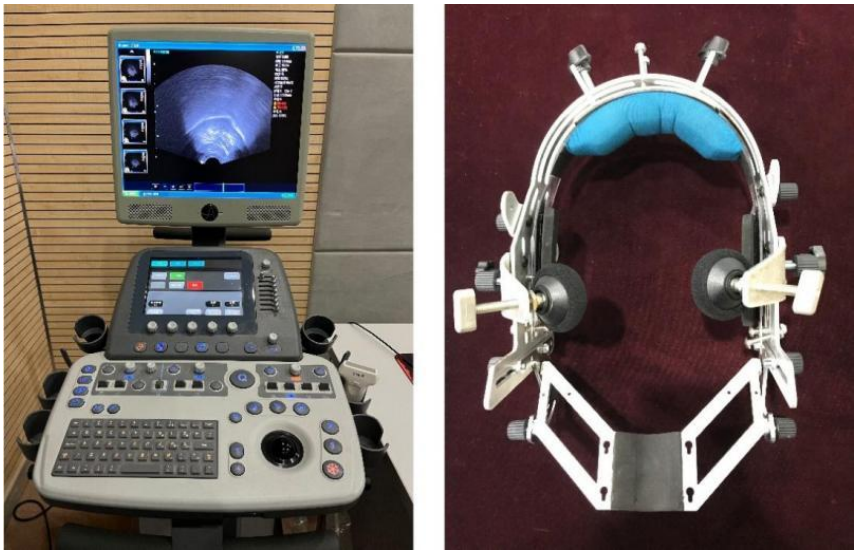


图 2　超声设备（左）、探头固定支架（右）

Figure 2　Ultrasound device (left), probe bracket (right)

（2）音频采集设备。采用音视频摄录机软件系统采集音频和视频（如图 3 所示），同时，在另外一台电脑上利用 Cool Edit Pro 软件对语音音频再次采集（如图 4 所示），为后续声学特征分析做好双重备份。采用 COMICA/科唛 V020 领夹式麦克风（如图 5 所示），频率响应范围 50 Hz–20 kHz。麦克风固定于发音人嘴唇一侧约 5 cm 处，保持与口部相对位置恒定以避免移动造成的音量变化。音频信号通过 Creative/创新 Blaster X-Fi Surround 5.1 Pro 外置声卡（如图 6 所示）和百灵达 Q502 调音台（如图 7 所示）接入计算机，采样率设为 44.1 kHz，量化精度 16 bit，单声道记录，以 WAV 格式保存。

（3）视频采集设备。采用 HIKVISION/海康威视 4K 高清摄像头和固定三脚架（如图 8 所示）放置于发音人正前方，并连接计算机中的语保摄录机软件系统，固定于发音人正前方约 50 cm 处，录制发音过程中唇部正面图像。视频参数设置为 1920×1080 分辨率，帧率 30 fps，以 AVI 格式保存。

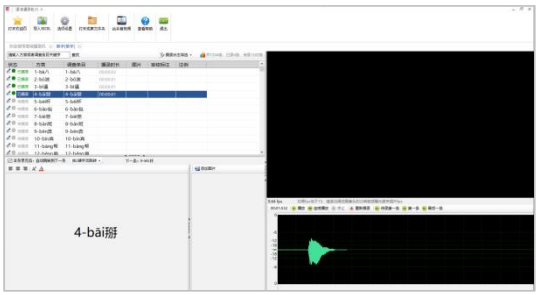


图 3　音视频摄录机软件系统

Figure 3　Audio-video recording system



图 4　Cool Edit Pro 软件

Figure 4　Cool Edit Pro software



图 5　麦克风

Figure 5　Microphone



图 6　声卡

Figure 6　Audio interface



图 7　调音台

Figure 7　Audio mixer



图 8　高清摄像头及固定三脚架

Figure 8　HD camera and tripod

1.4　多模态数据同步与对齐方法

为保证文本语料、语音音频、嘴唇视频和超声舌成像四类模态之间具有稳定的一一对应关系，本研究采用统一编号、同步触发、时间标记、后期校准的多层次同步策略。

首先，在采集前为每个发音单元设置唯一 ID，并将该 ID 同时写入文本语料、语音音频、嘴唇视频和超声舌成像文件名中，从文件组织层面建立 4 类模态之间的对应关系。例如，ID 为 1 的“bā八”在除了文本语料外的 3 类模态中均以“1-bā八”作为核心文件名，从而便于后续检索、比对和批量处理。

其次，在采集过程中，通过外置同步触发装置同时启动超声舌成像、音频采集和嘴唇视频采集设备，并在音频轨道中嵌入同步脉冲信号，作为多模态时间对齐的统一参考点。该同步脉冲在音频波形中表现为明显的瞬时峰值，可用于确定各模态数据的共同起始时刻。对于超声舌成像和嘴唇视频，分别根据采集帧率建立帧序列时间轴；对于语音音频，则依据 44.1 kHz 采样率建立采样点时间轴。通过将同步脉冲所在时间点设为共同零点，可以把音频采样点、嘴唇视频帧和超声舌像帧统一映射到同一时间坐标系中。

再次，在后期处理阶段，使用 Praat 对语音音频进行波形和语谱图检查，标注发音起止点、静默

段和异常发音片段；使用 MATLAB 或 Python 脚本读取音频、嘴唇视频和超声舌成像文件的时间信息，并依据同步脉冲位置、采样率和帧率对各模态数据进行时间轴校准。具体流程如图 9 所示。

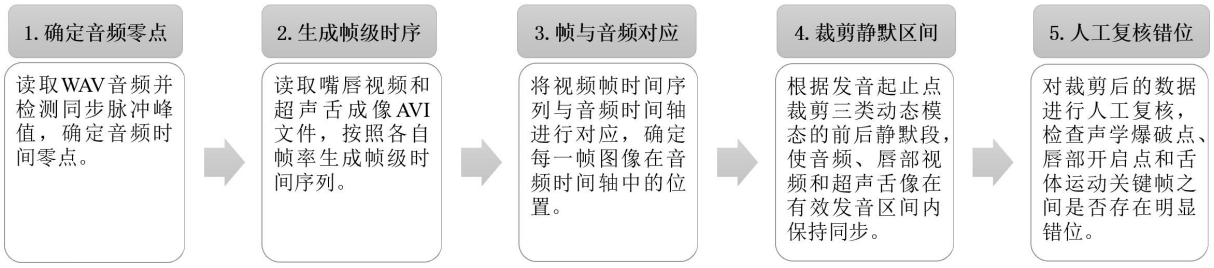


图 9　不同模态之间的对齐处理流程

Figure 9　Alignment workflow across different modalities

在具体对齐过程中，本研究不单纯以文件时长一致作为同步判断标准，而是结合音频波形、唇部运动和超声舌体运动的关键事件进行交叉验证。例如，在“bā”等双唇塞音音节中，语音音频中的爆破瞬间、嘴唇视频中的双唇开启时刻，以及超声舌像中元音阶段舌体运动的开始时间应当具有合理的时间对应关系。对于出现音频削波、视频卡顿、探头滑动或关键事件无法匹配的样本，本团队将其重新采集。通过上述处理，可以较好保证多模态数据在样本层面、时间轴层面和关键发音事件层面的同步一致性。

2　数据样本描述

2.1　典型样本选取

为便于读者直观理解数据集的构成与格式，本节选取典型发音单元 ID 为 1 的“bā”（声母 b，韵母 a，第一声阴平，常用汉字“八”）作为示例进行详细说明。理由如下：一是该音节包含双唇塞音声母和低元音韵母，涉及舌位从静止到低位前伸的典型过渡，能较好体现发音过程中舌运动特征；二是“a”元音在超声图像上舌面轮廓清晰，便于展示图像数据特征；三是该音节为汉语普通话基础音节，具有普遍代表性。

2.2　样本的数据结构与示例

本数据集以发音单元为基本组织单位，每个发音单元对应一个声韵调组合，包含文本语料、语音音频、嘴唇视频、超声舌成像 4 类数据。数据集共包含 1024 个有效发音单元，按模态类型、声调、音节类别进行三级目录组织。下面是 ID 为 1 的“bā”的具体示例：

（1）文本语料。所有文本语料分为 5 个 PDF 文件存储（如图 10 所示），其中“1-318 第一声（阴平）.pdf”文件中第二行内容包含多层信息：ID 号 1、声母 b、韵母 a、拼音 ba、声调第一声（阴平）、拼音声调组合 bā、常用汉字“八”等（如图 11 所示）。

（2）语音音频。音频采用 WAV 格式存储，文件命名为“1-bā八.wav”（如图 12 所示），采样率 44.1 kHz，量化精度 16 bit，单声道。音频时长约 1.3 秒。使用 Praat 软件打开 bā的语音音频对应的语谱图如图 13 所示。从语谱图中可以观察到，起始部分为爆破瞬间的能量集中区（垂直条纹），随后进入元音段，呈现出清晰稳定的共振峰结构，尤其是前两个共振峰（F1 和 F2）的走向，能够反映舌位从闭合到打开的过渡特征，可用于后续的语音分析与建模。

名称	修改日期	类型	大小	创建日期
1-318第一声（阴平）文本语料.pdf	2026/3/12 10:18	Adobe Acrobat 文档	384 KB	2026/3/12 10:18
319-564第二声（阳平）文本语料.pdf	2026/3/12 10:19	Adobe Acrobat 文档	319 KB	2026/3/12 10:19
565-866第三声（上声）文本语料.pdf	2026/3/12 10:19	Adobe Acrobat 文档	407 KB	2026/3/12 10:19
867-1200第四声（去声）文本语料.pdf	2026/3/12 10:20	Adobe Acrobat 文档	399 KB	2026/3/12 10:20
1201-1204特殊单韵母文本语料.pdf	2026/3/12 10:20	Adobe Acrobat 文档	79 KB	2026/3/12 10:20

图 10　文本语料文件

Figure 10　Text corpus files

ID	声母	韵母	拼音	声调	拼音声调组合	常用汉字
1	b	a	ba	第一声（阴平）	bā	八
2	b	o	bo	第一声（阴平）	bō	波
3	b	i	bi	第一声（阴平）	bī	逼
4	b	ai	bai	第一声（阴平）	bāi	掰
5	b	ei	bei	第一声（阴平）	bēi	杯

图 11　文本语料 bā示例

Figure 11　Example of “bā” in the text corpus

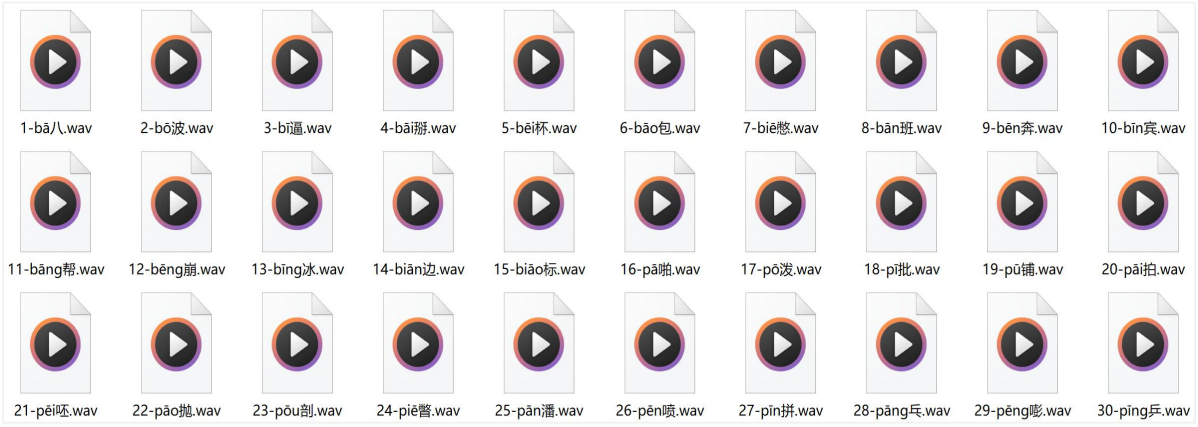


图 12　语音音频文件（ID：1-30）

Figure 12　Speech audio files (ID: 1-30)

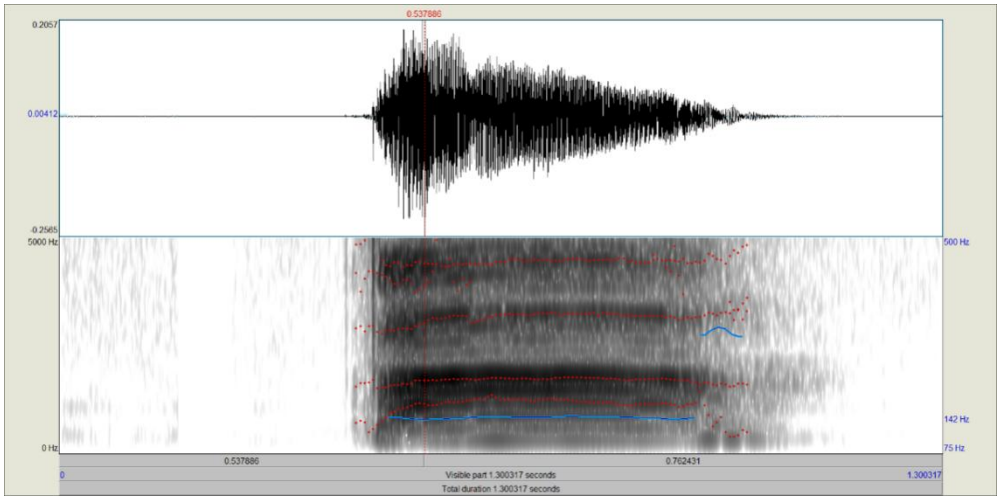


图 13　语音音频文件 bā语谱图

Figure 13　Spectrogram of the “bā” speech audio file

（3）嘴唇视频。首先需要正常录制发音人整个头部视频，以 AVI 格式存储，文件名为“1-bā八.avi”，分辨率为 1920×1080，帧率 30 fps，时长与音频严格同步。然后从头部视频中截取嘴唇部分，分辨率为 1280×720，如图 14 所示。“bā”的嘴唇视频从正面记录发音过程中唇形变化，包含双唇闭合时准备阶段（如第 1–3 帧）、双唇闭合时 b 声母成阻阶段（如第 4–7 帧）、双唇开启伴随口型张开（元音 a，如第 9–23 帧）的全过程，如图 15 所示。

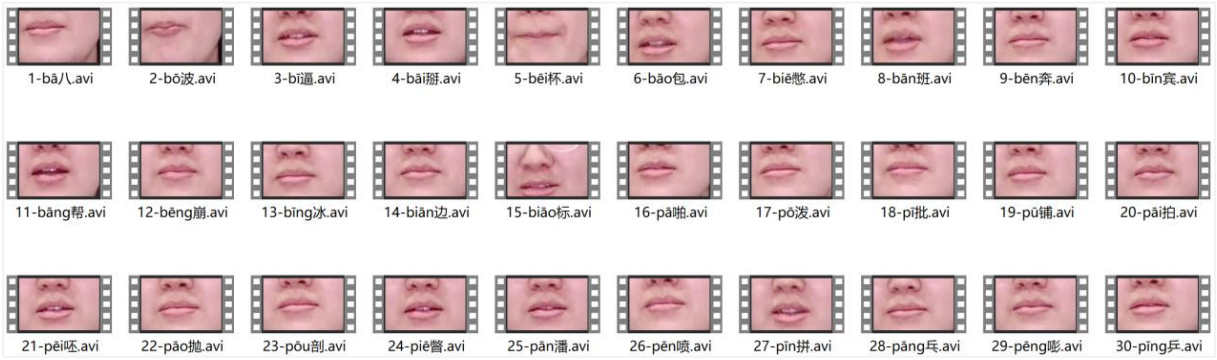


图 14　嘴唇视频文件（ID：1-30）

Figure 14　Lip video files (ID: 1-30)



图 15　bā发音时的嘴唇口型变化

Figure 15　Lip shape changes during the pronunciation of “bā”

（4）超声舌成像。超声数据以 AVI 格式存储，文件命名为“1-bā八.avi”，每个 AVI 视频可以转换为内含按时间顺序编号的 JPG 图像 98 帧（对应 33 fps 采集频率，总时长 1.8 秒）。单帧图像分辨率为 864×768 像素，显示舌体正中矢状面动态。为了把超声舌成像中附带的设备信息等干扰因素去除，可以把图像尺寸裁剪到 644×496 像素，只保留中间舌头区域，便于后续使用，如图 16 所示。第 1–10 帧为发音起始前舌头处于自然静息状态，第 11–40 帧对应声母 b 成阻阶段（舌体收紧），第 41–70 帧对应元音 a 发音阶段（舌体从前上方向后下方运动），第 71–98 帧为音节结束后的恢复过程，如图 17 所示。超声舌成像清晰呈现舌体运动轨迹，为鼻韵母舌运动提供典型样本。

3　数据质量控制和评估

为确保数据集的可靠性、准确性和可用性，本研究在采集准备阶段、数据采集过程中及后期处理阶段分别实施了系统化的质量控制措施，并对数据质量进行了量化评估和应用性验证。

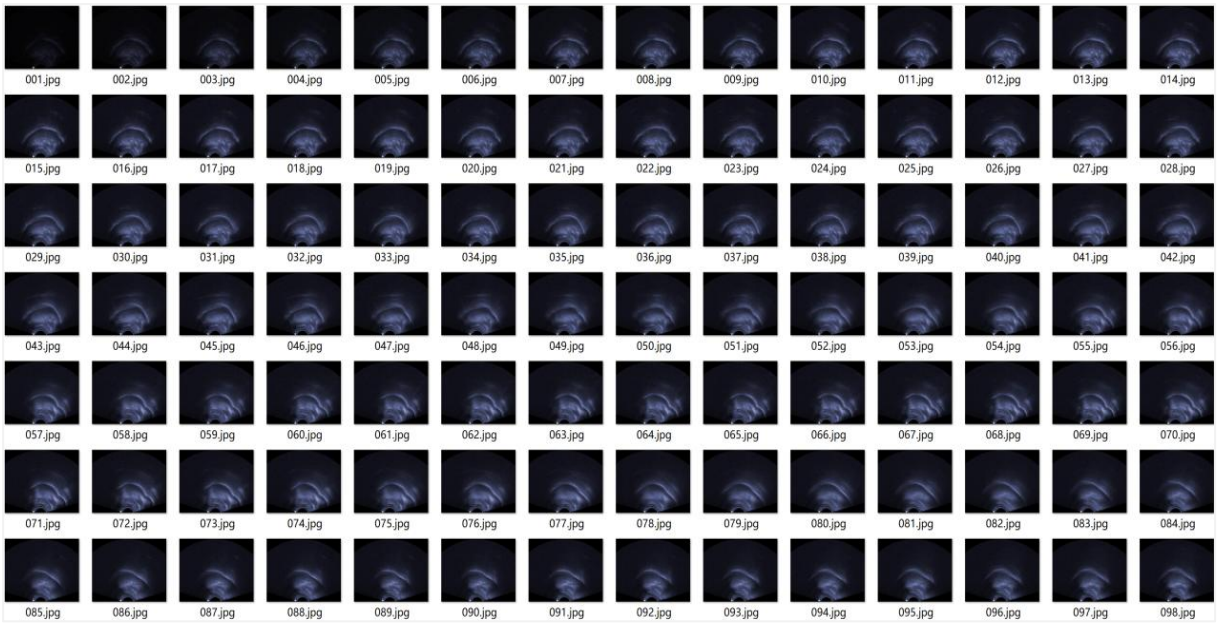


图 16　bā发音时的超声舌成像序列图

Figure 16　Ultrasound tongue image sequence during the pronunciation of “bā”

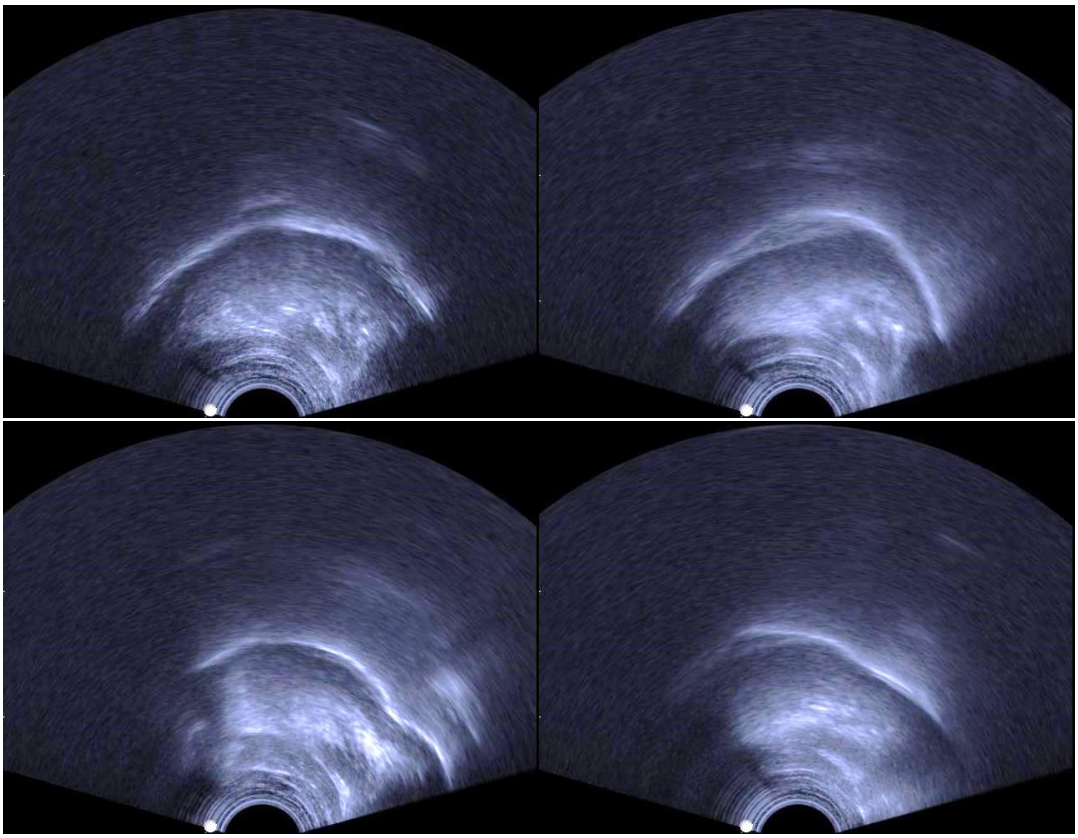


图 17　bā发音时的第 10、40、70、85 帧的超声舌成像（对应上左、上右、下左、下右图）

Figure 17　Ultrasound tongue imaging at frames 10, 40, 70, and 85 during the pronunciation of“bā” (corresponding to top-left, top-right, bottom-left, and bottom-right images, respectively)

3.1　采集准备阶段

在发音人筛选环节，采用专业语言学教师主观评测与客观语音分析相结合的方式，确保发音人

汉语普通话水平达到一级乙等，无任何言语、听力障碍。实验前对所有采集设备进行校准：超声探头通过仿体模测试验证成像清晰度，音频链路采用标准声源校准信噪比，视频摄像头进行白平衡和对焦校准。探头固定装置经预实验验证，在发音人头部轻微移动时仍能保持探头与下颌相对位置稳定。语料设计阶段，由两名语音学专家对所有发音单元的合规则性进行逐一审核，确保无不符合普通话音系规则的音节。

3.2 数据采集阶段

采集过程中实时监控多路信号质量。超声图像由实验人员实时观察，确保舌面轮廓清晰、探头位置未偏移，若出现图像模糊或探头滑动则暂停实验并重新调整。音频信号通过电平表实时监控，确保无削波失真。每完成一个发音单元，系统自动记录并同步标记。每采集 1 个单元后回放检查音视频质量和同步性，出现误差时调整设备重新采集。

3.3 后期处理阶段

数据整理完成后，实施多级质控流程。首先，进行有效性筛查，由两名语言学专业人员独立对所有发音单元进行主观评价，剔除发音错误、超声图像模糊、音频失真的样本，及时补录有问题的发音单元。其次，语音音频质量评估方面，计算所有有效样本的信噪比，确保无削波失真现象；通过频谱分析确认音频频率响应平坦，未引入采集设备导致的额外共振峰畸变。再次，在嘴唇视频质量评估方面，逐帧检查视频清晰度、曝光均匀性及唇部区域完整性，确保无运动模糊或焦点漂移；通过脉冲信号验证视频与音频的帧级同步精度。最后，超声图像舌面轮廓提取质量采用重测信度评估，随机选取 10 个发音单元的连续 10 帧图像（共 100 帧），由同一操作者间隔一周后重新提取轮廓，计算两次提取坐标的均方根误差，平均值为 1.32 mm（图像像素尺寸约 0.5 mm/像素），误差主要集中于舌尖区域，整体轮廓一致性良好。总之，数据完整性评估显示，所有有效样本均包含完整的四类数据，无数据缺失或断点，最终形成高质量、高可靠性的标准化数据集。

3.4 基于数据集的语音技术验证

前面第 3.1 节到第 3.3 节描述了在数据采集与处理的过程中进行的质量管控与指标测评，这些措施能够保障多模态数据集的可靠性与规范性。本节在此基础上将多模态数据集分别应用于语音识别与语音合成这两类典型语音工程任务，通过任务完成效果对多模态数据集开展功能性验证，考察其在真实应用场景中的实际表现，并分析不同模态特征在语音处理过程中呈现出性能差距与相互补充关系。

关于特征方法选择，本部分采用梅尔频率倒谱系数（MFCCs）、二维离散余弦变换（DCT）和 STRAIGHT 声码器等传统特征与工具，主要基于以下考虑：第一，这些方法在语音识别、语音合成和发音生理数据处理中具有较高的可解释性和稳定性，便于对不同模态数据的基本可用性进行初步验证；第二，传统特征对数据规模要求相对较低，更适合本研究这种以单一标准发音人为核心的数据集验证任务；第三，MFCCs 可与超声、唇部视觉特征直接维度匹配，保证多模态融合实验的公平性与可比性；第四，MFCCs、DCT 等特征便于与既有语音工程和超声舌成像相关研究进行横向比较，有助于说明数据集本身的质量和模态价值。与此同时，本科研团队也认识到，HuBERT、AV-HuBERT、WavVec 2.0、data2vec、Whisper、WavLM 等模型已在当前语音识别和语音表示学习中得到广泛应用，在后续研究中可进一步基于这些深度自监督模型提取高层语音表征，并与超声舌成像、唇部视频等

发音生理模态进行融合，以提升识别、合成和跨模态重建性能。

（1）语音识别。首先，在并行数据流上提取特征，并根据语音音频流的起始状态对超声舌成像和唇部视频中的初始和结束处的静默状态进行裁剪。针对语音音频流，提取 13 维 MFCCs 与其一阶、二阶导数组合成 39 维特征向量对语音音频流进行表示。针对超声舌成像和嘴唇视频流，将二维离散余弦变换（DCT）应用到每个超声舌成像和唇部视频帧上。其次，使用 Kaldi 语音识别工具包[15]训练声学模型，对上一阶段提取的相应特征初始化模型，在单音素和三音素训练之后，使用线性判别分析（LDA）和最大似然线性变换（MLLT）处理输入特征，使用特征空间 MLLR (fMLLR)自适应训练，按照 Kaldi 工具包的 nnet3 训练时延神经网络（TDNN）[16]，使用一个简单的域内二元语言模型进行解码。最后结果如表 2 所示，基于语音音频特征进行的语音识别的单词错误率最低，基于超声舌成像特征进行的语音识别的单词错误率居中，基于嘴唇视频特征进行的语音识别的单词错误率最高，而基于超声舌成像特征和嘴唇视频特征进行的语音识别的单词错误率大大降低。

表 2　不同模态特征的语音识别性能

Table 2　Speech recognition performance using different modal features

特征模态	单词错误率（WER/%）
音频	4.23
超声舌成像	28.67
嘴唇视频	45.41
超声舌成像+嘴唇视频	19.25

（2）语音合成。从发音运动器官（如舌头、嘴唇）图像中提取特征信息，将其映射到声学参数，最终合成语音。首先，对于嘴部视频和超声舌成像，使用多媒体处理工具 FFmpeg 对视频流重新采样以匹配超声流的帧速率。对于声学特征，使用 STRAIGHT 声码器[17]提取 41 维 MFCCs 和一维基频（F0），基频在清音帧内插值并转换为 logF0 和二进制清音/浊音标志，所有特征都连接起来形成一个 43 维的声学向量，声学特征被归一化。其次，训练阶段使用编码器-解码器架构，编码器将超声舌成像帧或唇部视频帧转换为 512 维向量。当超声舌成像和唇部视频都用作输入时，分别采用单独的编码器，接下来将两个编码器的输出连接起来，并发送到解码器。基于 ESPnet[18]开源的 ASR（automatic speech recognition）模型对合成的语音样本进行解码。最后结果如表 3 所示，基于超声舌成像特征进行的语音合成的 MCD（Mel Cepstral Distortion，表示的是合成后语音的 MFCC 特征与标准输出语音的 MFCC 特征的差距，单位为分贝）和单词错误率较低，基于嘴唇视频特征进行的语音合成的 MCD 和单词错误率较高，而基于超声舌成像特征+嘴唇特征进行的语音合成的 MCD 和单词错误率大幅降低。

表 3　不同模态特征的语音合成性能

Table 3　Speech synthesis performance using different modal features

特征模态	梅尔倒谱失真（MCD/dB）	单词错误率（WER/%）
超声舌成像	4.63	36.72
嘴唇视频	5.18	49.35
超声舌成像+嘴唇视频	4.34	20.11

综上所述，实验对前期经过多阶段质量控制的多模态数据集进行了实际应用验证，不仅佐证了前文质控措施的有效性，确保数据集能够满足实际任务需求，更从应用层面对数据集进行了功能性

验证。通过多模态特征融合后性能的显著提升，凸显了多模态数据在语音信息处理中的互补价值及其相较于单一模态数据的独特优势——其内在的跨模态关联特征能够有效支撑复杂的语音技术任务，而将发音生理过程（舌运动）与外部可视发音动作（唇动）及声学产出相结合的特性。这也为探索视听与生理多模态交互机制、发展高鲁棒性语音接口提供了独特的实验平台，同时为后续相关研究中多模态数据的应用与质量管控提供了实践支撑与理论参考。

4 数据价值

本数据集在音系覆盖范围、多模态集成度、数据同步精度及质量控制标准化等方面具有显著特色，与国内外现有同类数据集相比展现出独特优势，具备广阔的基础研究与实际应用前景，可为汉语普通话发音生理机制探索、语音技术创新及跨语言比较研究提供高质量的基础资源支撑。

4.1 数据的特色与优势

与国内外同类数据集[7-9,12-13]相比，本数据集的创新性与核心价值主要体现在以下两个方面：

在数据覆盖范围上，国际同类数据集虽包含多语言超声舌像数据，但汉语普通话样本极少，且未覆盖完整声韵调组合。国内数据集方面，学者们多采集汉语普通话元音超声数据，未覆盖辅音及声调；有的双模态数据集缺少嘴唇视频模态，且未实现多模态精准同步。本数据集完整覆盖汉语普通话所有声韵母搭配及四个声调，同时集成文本、音频、嘴唇视频、超声舌成像 4 大模态，实现了汉语普通话音系的全方位覆盖与多维度生理信息的融合，显著优于现有同类数据集。

在数据加工与来源上，本数据集突破现有研究局限，在来源上选取汉语普通话一级乙等标准发音人，从源头保障发音规范性；在加工过程中采用同步触发技术，结合 Praat、MATLAB 等专业工具，实现多模态数据的对齐与标准化预处理，构建结构化文件管理体系，解决了现有数据集同步性差、加工不规范的问题。在质量控制上，采用采集前校准、采集时监控、后期多级校验的全流程质控体系，数据可靠性更高。

4.2 数据的学术价值与实践意义

在基础研究层面，本数据集可以为深入探究汉语普通话发音生理机制提供基础数据。研究人员不仅可以系统地分析声母、韵母、声调这三类语音单位的发音生理特征以及它们之间的相互作用关系，而且可以分析不同声调条件下元音发音舌位的前后高低变化情况，还可以分析辅音成阻部位与后续元音之间的协同发音现象。本数据集可以支持多维度分析，超声舌成像可用于舌体轮廓提取和声道面积函数计算，语音音频可用于声学特征分析，嘴唇视频可用于口型开合变化研究，而把超声舌成像、嘴唇视频以及语音音频多模态数据融合在一起就可以探究发音运动与声学输出之间的映射关系。此外，本数据集还能够为二语习得研究提供汉语普通话标准发音的生理参照，帮助学习者发现自身发音与母语者之间的运动差异。

在跨语言研究层面，本数据集能够为跨语言发音生理对比研究提供重要的参考。本数据集与英语、藏语等语言的同类数据[19]相比，研究人员可以从发音运动层面揭示汉语普通话作为声调语言的类型学特征，进一步探讨不同语言音系结构对发音器官运动模式的塑造作用。本数据集使用了国际通用的数据格式和开放共享的发布方式，方便国内外研究人员下载使用。

在应用技术开发层面，本数据集在语音工程领域的多个技术方向上都可以使用。在语音合成方

面，研究人员可以对多模态发音生理数据进行参数化建模以便提升语音合成的自然流畅度[20]；在语音识别方面，可以把多模态数据作为视听融合识别的训练数据，这样可以提升噪声环境下的识别鲁棒性；在言语障碍康复方面，本数据集不仅可以为评估患者发音功能提供一些参照，也可以为患者康复训练提供具体的量化目标[21]。随着人工智能技术的不断发展，本数据集还可用于训练发音运动预测模型、构建虚拟发音人以及为语音脑机接口等前沿技术提供基础数据。此外，多模态深度学习技术作为人工智能的重要分支，通过融合图像、文本、语音、视频等数据能够大幅提升多模态模型的感知与推理能力，比如，有的研究人员基于图神经网络的跨模态关系捕捉方法[22]构建了发音器官运动图结构，这样就增强了生理数据与声学信号的关联性；有的研究人员基于迁移学习的跨模态知识迁移框架[23-24]将富集模态（如英语语音）的知识迁移至低资源语言（如藏语拉萨话）建模中；也有研究人员基于深度学习的端到端技术[25]在汉语普通话的超声舌成像中重建了语音。

5 数据使用方法和建议

为了增强数据集的可复用性，本部分详细描述了不同模态数据的导入方式、格式转换和基本处理流程。

语音音频数据以 WAV 格式保存，可直接导入 Praat、MATLAB、Python 或 Kaldi 等工具中使用。使用 Praat 时，可通过“Open—Read from file”导入 WAV 文件，随后查看波形和语谱图，并根据研究需要提取基频、共振峰、时长、强度、语音起止点等声学参数。若用于机器学习或语音识别任务，可使用 Python 中的 librosa、torchaudio 等工具读取 WAV 文件，并进一步提取 MFCCs、谱图、梅尔谱图或自监督语音表征等特征。

嘴唇视频数据以 AVI 格式保存，可使用 FFmpeg 转换为逐帧图像序列，也可直接使用 OpenCV 读取。基本流程为：首先将 AVI 文件按原始帧率转换为 PNG 或 JPG 图像序列；其次根据唇部区域位置裁剪感兴趣区域；再次结合 OpenCV、MediaPipe 或其他计算机视觉工具提取唇部开合度、唇宽、唇高、面积变化或关键点坐标等参数。对于需要与音频或超声舌像融合分析的任务，应保留帧序号和时间戳信息，确保唇部运动参数能够与音频时间轴和超声帧序列对应。

超声舌成像数据以 AVI 格式及相关 XML 文件保存。用户可先使用 FFmpeg 将超声 AVI 转换为按时间顺序编号的图像序列，再根据本研究提供的裁剪区域保留舌体运动的核心图像区域。随后，可将图像序列导入 EdgeTrak 或 AAA 软件进行舌面轮廓半自动提取，也可使用 MATLAB 或 Python 进行自定义图像处理。具体的处理步骤如图 18 所示。

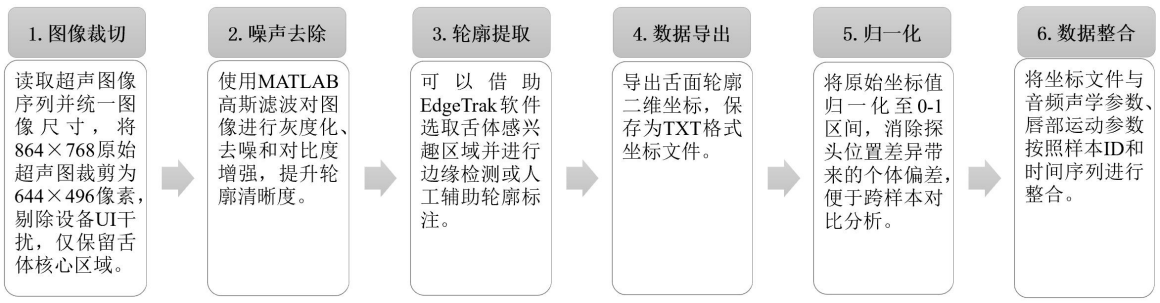


图 18　超声舌轮廓的提取和预处理流程

Figure 18　Workflow for ultrasound tongue contour extraction and preprocessing

在多模态融合分析中，建议用户采用“ID 匹配+时间轴匹配”的方式组织数据。ID 匹配用于保

证同一发音单元的四类模态能够对应，时间轴匹配用于保证同一发音过程中的音频采样点、嘴唇视频帧和超声图像帧能够在时间维度上对齐。对于不同帧率的数据，可根据音频起始点和同步标记，将视频帧统一转换到音频的时间轴中，再根据研究任务选择插值、重采样或关键帧匹配等方式完成融合。

需要说明的是，本研究提供的是原始数据及标准化整理后的多模态数据资源，使用者可根据具体研究目的选择不同层次的处理方式。若研究重点为语音学机制分析，可重点提取舌面轮廓、舌高点、舌前后位置、唇部开合度及基频、共振峰等参数；若研究重点为语音工程或多模态学习，可将音频谱图、超声图像帧和唇部视频帧作为多模态输入，构建语音识别、语音合成、发音运动预测或跨模态重建模型。此外，本数据集也可以与电声门图仪（EGG）、电磁发音仪（EMA）等其他模态数据结合进行更复杂的多模态融合分析。

使用者须注意，本数据集采集自单一发音人，虽具有发音标准和数据一致性高的优势，但个体差异因素（身高、体重、年龄、性别、母语或方言、受教育程度等）需在跨人群研究中加以考虑。本数据集遵循伦理规范，已做匿名化处理，不含任何可识别个人身份的信息，可公开共享。本数据集使用的许可协议是 CCBY-NC-SA 4.0（NC 指非商业性使用，SA 指衍生数据应遵循原数据的使用许可方式继续共享），禁止用于商业用途，仅可用于学术研究与非营利性实践，使用过程中需注明数据集来源和引用本论文，确保数据合理合规复用。

6 AIGC 使用声明

在撰写本手稿期间，作者未使用 AIGC 工具。所有手稿和数据均由作者独立完成，作者对论文的整体内容、数据真实性及学术诚信负全部责任。

作者分工职责

张金溪（1985—），男，山东省沂水县人，博士，副教授，研究方向为实验语音学、语音工程、深度学习。主要承担工作：数据采集、校对及评估，论文撰写，数据质量管理。

于洪志（1947—），女，山东省龙口市人，教授，博士生导师，研究方向为自然语言处理。主要承担工作：数据审核，数据质量控制与评估。

张晶（1989—），女，甘肃省庆阳市人，博士，副教授，研究方向为实验语音学、藏语语音学。主要承担工作：文本语料设计，发音人选择，数据筛选。

参考文献

[1] 张金溪, 李永宏, 寇赟. 基于超声成像技术的发音生理舌体研究方法综述[J]. 计算机应用研究, 2021, 38(1): 15–22. DOI:10.19734/j.issn.1001-3695.2019.11.0651. [ZHANG J X, LI Y H, KOU Y. Research method of pronunciation physiological tongue based on ultrasonic imaging technique[J]. Application Research of Computers, 2021, 38(1): 15–22. DOI:10.19734/j.issn.1001-3695.2019.11.0651.]

[2] KELSEY C A, MINIFIE F D, HIXON T J. Applications of ultrasound in speech research[J]. Journal of Speech and Hearing Research, 1969, 12(3): 564–575. DOI:10.1044/jshr.1203.564.

[3] 冯胜利. 汉语音法、词法和句法的基础结构特殊性溯源[J]. 语言教学与研究, 2025(5): 50–66.

DOI:10.3969/j.issn.0257-9448.2025.05.006. [FENG S L. Tracing the uniqueness of the basic structures of Chinese phonology, morphology, and syntax[J]. Language Teaching and Linguistic Studies, 2025(5): 50–66. DOI:10.3969/j.issn.0257-9448.2025.05.006.]

[4] 张金溪, 李永宏, 寇赟. 基于超声成像的语言研究综述[J]. 西北民族大学学报(哲学社会科学版), 2019(6): 143–153. DOI:10.14084/j.cnki.cn62-1185/c.2019.06.014. [ZHANG J X, LI Y H, KOU Y. Overview of linguistic study based on ultrasound imaging technology[J]. Journal of Northwest Minzu University (Philosophy and Social Science), 2019(6): 143–153. DOI:10.14084/j.cnki.cn62-1185/c.2019.06.014.]

[5] WESTBURY J, MILENKOVIC P, WEISMER G, et al. X-ray microbeam speech production database[J]. The Journal of the Acoustical Society of America, 1990, 88(S1): S56. DOI:10.1121/1.2029064.

[6] NARAYANAN S, TOUTIOS A, RAMANARAYANAN V, et al. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)[J]. The Journal of the Acoustical Society of America, 2014, 136(3): 1307. DOI:10.1121/1.4890284.

[7] ESHKY A, RIBEIRO M S, CLELAND J, et al. UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions[C]. Interspeech 2018. ISCA, 2018: 1888–1892. DOI:10.21437/interspeech.2018-1736.

[8] RIBEIRO M S, SANGER J, ZHANG J X, et al. Tal: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos[C]. 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen, China. IEEE, 2021: 1109–1116. DOI:10.1109/SLT48900.2021.9383619.

[9] KIMURA N, SU Z, SAEKI T, et al. SSR7000: a synchronized corpus of ultrasound tongue imaging for end-to-end silent speech recognition[C]. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022, 6866–6873. https://aclanthology.org/2022.lrec-1.741.

[10] 郑玉玲, 朱思俞. 普通话语音动态腭位数据库及其研究平台[C]. 第六届全国人机语音通讯学术会议论文集. 北京: 中国中文信息学会, 2001, 319–324. [ZHENG Y L, ZHU S Y. A dynamic palatographic database of Mandarin Chinese and research platform [C]. Proceedings of the 6th National Conference on Human-Machine Speech Communication. Beijing: Chinese Information Processing Society of China, 2001, 319-324.]

[11] 陈彧. 基于超声波检测的汉语普通话基础元音发音的舌体运动研究[D]. 天津: 南开大学, 2012. [CHEN Y. An investigation of togue movement in the production of Mandarin's basic vowels using ultrasound[D]. Tianjin: Nankai University, 2011.]

[12] YANG Y D, SU R F, ZHAO S F, et al. An audio-ultrasound synchronized database of tongue movement for Mandarin speech[J]. Scientific Data, 2025, 12: 607. DOI:10.1038/s41597-025-04917-w.

[13] 张金溪, 于洪志, 张晶. 藏语拉萨话元音发音时的超声舌成像数据集[J/OL]. 中国科学数据, 2026, 11(2). (2026-06-03). DOI: 10.11922/11-6035.csd.2025.0150.zh. [ZHANG J X, YU H Z, ZHANG J. Ultrasound tongue imaging dataset for vowel articulation in Tibetan Lhasa Dialect[[J/OL]. China Scientific Data, 2026, 11(2). (2026-06-03). DOI: 10.11922/11-6035.csd.2025.0150.zh.]

[14] 张金溪, 张克宏, 于洪志. 深度学习在超声舌成像处理中的研究进展[J]. 中国图象图形学报, 2026, 31(2): 391–432. DOI:10.11834/jig.250015. [ZHANG J X, ZHANG K H, YU H Z. Advances in deep

learning techniques for ultrasound tongue imaging processing[J]. Journal of Image and Graphics, 2026, 31(2): 391–432. DOI:10.11834/jig.250015.]

[15] RAVANELLI M, PARCOLLET T, BENGIO Y. The pytorch-kaldi speech recognition toolkit[C]. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom. IEEE, 2019: 6465–6469. DOI:10.1109/icassp.2019.8683713.

[16] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]. Interspeech 2015. ISCA, 2015: 3214 – 3218. DOI:10.21437/interspeech.2015-647.

[17] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNÉ A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3/4): 187–207. DOI:10.1016/S0167-6393(98)00085-5.

[18] WATANABE S, HORI T, KARITA S, et al. ESPnet: end-to-end speech processing toolkit[C]. Interspeech 2018. ISCA, 2018: 2207–2211. DOI:10.21437/interspeech.2018-1456.

[19] ONYANGO O N, WANZARE L, OBUHUMA J. Swahili and code-switched English-Swahili political hate speech detection textual dataset[J]. Data Intelligence, 2025, 7(3): 819-850. DOI: 10.3724/2096-7004.di.2025.0053.

[20] REN Z, ZHU L, ZHAO X, et al. Amdo-Chinese speech translation dataset[J]. Data Intelligence, 2025, 7(3): 786-797. DOI: 10.3724/2096-7004.di.2024.0024.

[21] YANG Y D, SU R F, ZHAO S F, et al. Towards unified diffusion model for speech to ultrasound tongue imaging synthesis[J]. Information Fusion, 2026, 127: 103896. DOI:10.1016/j.inffus.2025.103896.

[22] 邹慧琪, 史彬泽, 宋凌云, 等. 基于图神经网络的复杂时空数据挖掘方法综述[J]. 软件学报, 2025, 36(4): 1811–1843. DOI:10.13328/j.cnki.jos.007275. [ZOU H Q, SHI B Z, SONG L Y, et al. Survey on complex spatio-temporal data mining methods based on graph neural network[J]. Journal of Software, 2025, 36(4): 1811–1843. DOI:10.13328/j.cnki.jos.007275.]

[23] HE S T, GUO T W, DAI T, et al. Open-vocabulary multi-label classification via multi-modal knowledge transfer[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 847–855.

[24] KUNXI LI T Z. MergeNet: knowledge migration across heterogeneous models, tasks, and modalities[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39: 4824–4832. DOI:10.1609/aaai.v39i5.32510.

[25] LI F J, SHEN F, MA D, et al. End-to-end Mandarin speech reconstruction based on ultrasound tongue images using deep learning[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2025, 33: 140–149. DOI:10.1109/TNSRE.2024.3520498.

论文引用格式

张金溪, 于洪志, 张晶. 基于超声舌成像的汉语普通话多模态发音生理数据集[J/OL]. 中国科学数据, 2026, 11(2). (2026-06-11). DOI: 10.11922/11-6035.csd.2026.0038.zh.

数据引用格式

张金溪，于洪志，张晶. 基于超声舌成像的汉语普通话多模态发音生理数据集[DS/OL]. V1. Science Data Bank, 2026. (2026-03-31). DOI: 10.57760/sciencedb.j00001.01617.

A dataset of multimodal articulatory physiological data for Mandarin Chinese based on ultrasound tongue imaging

ZHANG Jinxi[1], YU Hongzhi[2,3*], ZHANG Jing[1]

1. School of Business and Media, Lanzhou University of Finance and Economics, Lanzhou 730020, P.R. China

2. Key Laboratory of Linguistic and Cultural Computing Ministry of Education (Northwest Minzu University), Lanzhou 730030, P.R. China

3. Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Gansu Province (Northwest Minzu University), Lanzhou 730030, P.R. China

*Email: yhz@xbmu.cn (YU Hongzhi)

Abstract: Articulatory physiological data are the core foundation of Mandarin Chinese phonetic research and speech engineering. At present, the multimodal pronunciation physiological datasets for Mandarin Chinese have several limitations, including incomplete coverage, single-modality acquisition, and lack of synchronization, which are difficult to meet the requirements of high precise research. To address this issue, this study constructs a multimodal pronunciation physiological dataset of Mandarin Chinese based on ultrasound tongue imaging, thereby addressing the deficiency of existing datasets in the fusion of multi-dimensional pronunciation physiological information. The dataset covers commonly used valid syllable units formed by combinations of initials and finals under four tone conditions, forming 1,024 complete pronunciation units. Multimodal data consists of four parts: text corpora, speech audio, lip video, and ultrasound tongue imaging, which can comprehensively reflect the physiological movement characteristics and acoustic performance during the pronunciation process. In the data quality control stage, a combination of manual verification and machine screening is adopted to eliminate invalid data such as non-standard pronunciation, blurry images, and audio distortion, ultimately ensuring a high-quality dataset. The dataset not only provides data support for basic research on the physiological mechanism of Mandarin Chinese pronunciation, the rules of tone changes, and second language acquisition, but also has applications in speech synthesis and recognition, diagnosis and rehabilitation of speech disorders, modeling of pronunciation mechanisms, and training of artificial intelligence speech models. At the same time, it offers a reference for cross-language comparative studies on pronunciation physiology.

Keywords: ultrasound tongue imaging; Mandarin Chinese; multimodal dataset; articulatory physiology; artificial intelligence; speech engineering

Dataset Profile

Title	A dataset of multimodal articulatory physiological data for Mandarin Chinese based on ultrasound tongue imaging
Data corresponding author	ZHANG Jinxi (582361484@qq.com)
Data authors	ZHANG Jinxi, YU Hongzhi, ZHANG Jing
Time range	2026
Geographical scope	Lanzhou City, Gansu Province
Data volume	41.33 GB (290 GB after decompression)
Data format	*.avi, *.xml, *.wav, *.pdf
Data service system	https://doi.org/10.57760/sciencedb.j00001.01617
Sources of funding	The National Natural Science Foundation of China (62366045, 62566057); the National Social Science Fund of China (22BYY165); Gansu Province Central Government Guides Local Science and Technology Development Fund Project (25ZYJA034); Longyuan Young Talents Project: A Study on the Co-evolutionary Features of Multi-ethnic Languages in Gansu Province; Lanzhou City Science and Technology Plan Project (2025-2-42).
Dataset composition	The data files of the speakers include: (1) Text corpus. PDF files representing five categories of pronunciation features, including 1-318 First Tone (High Level) Text Corpus.pdf, 319-564 Second Tone (Rising) Text Corpus.pdf, 565-866 Third Tone (Falling-Rising) Text Corpus.pdf, 867-1200 Fourth Tone (Falling) Text Corpus.pdf, and 1,201-1,204 Special Simple Finals Text Corpus.pdf. The total data size is 1.54 MB. (2) Audio recordings. RAR compressed files containing five categories of pronunciation features, with WAV format audio in the compressed files, namely: 1-318 First Tone (High Level) Audio Recordings.rar, 319-564 Second Tone (Rising) Audio Recordings.rar, 565-866 Third Tone (Falling-Rising) Audio Recordings.rar, 867-1,200 Fourth Tone (Falling) Audio Recordings.rar, and 1,201-1,204 Special Simple Finals Audio Recordings.rar. The total data size is 180 MB. (3) Lip videos. RAR compressed files containing five categories of pronunciation features, with AVI format videos in the compressed files, namely: 1-318 First Tone (High Level) Lip Videos.rar, 319-564 Second Tone (Rising) Lip Videos.rar, 565-866 Third Tone (Falling-Rising) Lip Videos.rar, 867-1,200 Fourth Tone (Falling) Lip Videos.rar, and 1,201-1,204 Special Simple Finals Lip Videos.rar. The total data size is 639 MB. (4) Ultrasound tongue imaging. RAR compressed files containing five categories of pronunciation features. The compressed files include AVI format ultrasound video data and XML annotation files in the compressed files, namely: 1-318 First Tone (High Level) Ultrasound Tongue Imaging.rar, 319-564 Second Tone (Rising) Ultrasound Tongue Imaging.rar, 565-866 Third Tone (Falling-Rising) Ultrasound Tongue Imaging.rar, 867-1,200 Fourth Tone (Falling) Ultrasound Tongue Imaging.rar, and 1,201-1,204 Special Simple Finals Ultrasound Tongue Imaging.rar. The total data size is 40.6 GB (289 GB after decompression).