MAGIC DATA
Data Set Your Mind

MD Datasets

Dataset Products – Smart Travel

Smart Travel		
Scenarios	**Dataset**	**Hours**
Smart Cockpit	In-Vehicle Mandarin Chinese Scripted Speech Corpus	1,000
	Man-Manchine Interaction Scripted Speech Corpus	6,000
	Mandarin Chinese Scripted Speech Corpus	20,000
	Mandarin Chinese Conversational With Natural Style Speech Corpus	20,000
	Chinese Dialect Scripted Speech & Conversational Speech Corpus	15,000
	Chinese-English Code-Mixing Scripted Speech Corpus	2,000
	Chinese Speaking English Scripted Speech Corpus	7,000
	In-Vehicle Noise Corpus	500

Dataset Products – Smart Customer Service

Smart Customer Service		
Scenarios	**Dataset**	**Hours**
On-line Service Off-line Service	Mandarin Chinese Conversational Telephony Speech Corpus	15,000
	Mandarin Chinese Customer-Service ASR Corpus	5,000
	Chinese Dialect Scripted Speech & Conversational Speech Corpus	15,000
	Mandarin Chinese Conversational With Natural Style Speech Corpus	20,000
	Noisey Environment-Service ASR Corpus	2,000
	Chinese-English Code-Mixing Scripted Speech Corpus	2,000
	Chinese Speaking English Scripted Speech Corpus	7,000

Dataset Products – Smart Meeting

Smart Meeting		
Scenarios	**Dataset**	**Hours**
On-line Meeting Off-line Meeting	Mandarin Chinese Meeting Conversational Speech Corpus	1,500
	Mandarin Chinese Broadcast Speech Corpus	20,000
	Mandarin Chinese Conversational With Natural Style Speech Corpus	20,000
	Noisey Environment-Service ASR Corpus	2,000
	Mandarin Chinese Scripted Speech Corpus	20,000
	Chinese-English Code-Mixing Scripted Speech Corpus	2,000
	Chinese Speaking English Scripted Speech Corpus	7,000

Dataset Products – Social Intelligence

Social Intelligence		
Scenarios	**Dataset**	**Hours**
Video-sharing Platform	Mandarin Chinese Conversational With Natural Style Speech Corpus	20,000
	Mandarin Chinese Broadcast Speech Corpus	20,000
	Mandarin Chinese Scripted Speech Corpus	20,000
	Chinese Dialect Scripted Speech & Conversational Speech Corpus	15,000
	Chinese-English Code-Mixing Scripted Speech Corpus	2,000
	Chinese Speaking English Scripted Speech Corpus	7,000
	Noisy Environment-Service ASR Corpus	2,000

Dataset Products – Smart Home/Smart Device

Smart Home/Smart Device		
Scenarios	**Dataset**	**Hours**
Smart Phone Wearable Device Smart Home Smart Appliances	Far-Field Man-Manchine Interaction Scripted Speech Corpus	2,500
	Mandarin Chinese Scripted Speech Corpus	20,000
	Mandarin Chinese Conversational With Natural Style Speech Corpus	20,000
	Chinese Dialect Scripted Speech&Conversational Speech Corpus	15,000
	Chinese-English Code-Mixing Scripted Speech Corpus	2,000
	Chinese Speaking English Scripted Speech Corpus	7,000
	Indoor Noise Corpus	500

Dataset Products – Global Languages

Languages	Hours	Conversational Hours	Scripted Hours
English	20000	4,000	16,000
Korean	7500	5,500	2,000
Japanese	6500	5,000	1,500
Indonesian	4600	3,000	1,600
Malay	3500	2,500	1,000
Turkish	2500	2,000	500
Filipino	2500	2,000	500
Thai	2100	800	1300
German	1600	1,100	500
Russian	1400	1,000	400
Italian	1200	700	500
Portuguese	1200	1,200	
Spanish	1000	500	500
French	1000	500	500
Arabic	1000	800	200
Vietnamese	1000	1,000	
Urdu	500		500
Hindi	500		500

训练数据集产品-智慧出行

智慧出行		
场景	数据集	数据量(h)
智能座舱	车载环境中文朗读数据集	1,000
	人机交互类朗读数据集	6,000
	中文普通话朗读数据集	20,000
	中文自然对话数据集	20,000
	方言朗读&对话数据集	15,000
	中英混合数据集	2,000
	中国人说英语数据集	7,000
	车载噪音数据集	500

训练数据集产品-智能客服

智能客服		
场景	数据集	数据量(h)
远程客服 面对面客服	中文电话信道对话数据集	15,000
	中文客服对话数据集	5,000
	方言朗读&对话数据集	15,000
	中文自然对话数据集	20,000
	带噪环境数据集	2,000
	中英混合数据集	2,000
	中国人说英语数据集	7,000

训练数据集产品-智能会议

智能会议		
场景	数据集	数据量(h)
云端会议 线下会议	中文会议对话数据集	1,500
	中文访谈数据集	20,000
	中文自然对话数据集	20,000
	带噪环境数据集	2,000
	中文普通话朗读数据集	20,000
	中英混合数据集	2,000
	中国人说英语数据集	7,000

训练数据集产品-智能社交

智慧社交		
场景	数据集	数据量(h)
短视频	中文自然对话数据集	20,000
	中文访谈数据集	20,000
	中文普通话朗读数据集	20,000
	方言朗读&对话数据集	15,000
	中英混合数据集	2,000
	中国人说英语数据集	7,000
	带噪环境数据集	2,000

训练数据集产品-智能家居/终端

智能家居/终端		
场景	数据集	数据量(h)
智能手机 穿戴设备 智能家居 智能家电	远场人机交互朗读数据集	2,500
	中文普通话朗读数据集	20,000
	中文自然对话数据集	20,000
	方言朗读&对话数据集	15,000
	中英混合数据集	2,000
	中国人说英语数据集	7,000
	家居噪音数据集	500

训练数据集产品-外语

语种	总时长(h)	对话式时长(h)	朗读式时长(h)
英语	20,000	4,000	16,000
韩语	7,500	5,500	2,000
日语	6,500	5,000	1,500
印尼语	4,600	3,000	1,600
马来语	3,500	2,500	1,000
土耳其语	2,500	2,000	500
菲律宾语	2,500	2,000	500
泰语	2,100	800	1,300
德语	1,600	1,100	500
俄语	1,400	1,000	400
意大利语	1,200	700	500
葡萄牙语	1,200	1,200	
西班牙语	1,000	500	500
法语	1,000	500	500
阿拉伯语	1,000	800	200
越南语	1,000	1,000	
乌尔都语	500		500
印地语	500		500

Magic Data R&D Center works on conversational AI data and read speech data comparison

Compared with read speech data, conversational AI data word accuracy is improved up to 84%

Word Accuracy		3,000 hrs Training Data	
		Read speech	Conversational data
Scenarios	Customer services	43%	79%
	Broadcasting	70%	82%
	Navigation command	51%	64%

The more the conversational data is used, the higher the word accuracy comes

Word Accuracy		Conversational Data	
		3,000 hrs	30,000 hrs
Scenarios	Customer services	79%	83%
	Broadcasting	82%	89%
	Navigation command	64%	71%

Conclusion：Compared with read speech training data, conversational training data has better performance on machine learning.

Magic Data R&D Center

Magic Data研发中心进行的对话数据与朗读数据对比试验

等小时的对话式数据比朗读数据字正确率提升84%

字正确率		3,000 小时训练数据	
		朗读数据	对话数据
场景	客服对话	43%	79%
	直播社交	70%	82%
	车载导航	51%	64%

对话数据小时数越多，字正确率越高

字正确率		对话式数据	
		3,000 小时	30,000 小时
场景	客服对话	79%	83%
	直播社交	82%	89%
	车载导航	64%	71%

结论：与朗读数据相比，对话数据使机器学习表现更佳。

Magic Data 研发中心