中文手机录音音频语料库

MAGICDATA Mandarin Chinese Read Speech

MAGICDATA Mandarin Chinese Read Speech Corpus was developed by MAGIC DATA Co., Ltd. and freely published for non-commercial use. The corpus consists of 755 hours of scripted read speech data by 1000 native speakers of the Mandarin Chinese spoken in mainland China. In order to accommodate demands of variety, the corpus comprises three subsets: a training set (712.09 hours), a dev set (14.84 hours), and a test set (28.08 hours). All of the three sets are recorded indoor by smart phone and outputs are PCM formatted. The domain of recording texts is diversified, including interactive Q&A, music search, SNS messages, home command and control, etc.

MagicData中文手机录音音频语料库包含755小时的中文普通话朗读语音数据，其中分为训练集712.09小时、开发集14.84小时和测试集28.08小时。本语料库的录制文本覆盖多样化的使用场景，包括互动问答、音乐搜索、口语短信信息、家居命令控制等。采集方式为手机录音，涵盖多种类型的安卓手机；录音输出为PCM格式。1000名来自中国不同口音区域的发言人参与采集。MagicData中文手机录音音频语料库由MagicData有限公司开发，免费发布供非商业使用。