# 面向藏语语音情感识别的数据集构建与验证

## ——以卫藏方言TSEC-4528为例

柏梦琳[1, 2]　黄荣兆[1, 2]　边巴旺堆[1, 2]
1.西藏大学（信息科学技术学院）　西藏拉萨　850000
2.西藏大学（国家级实验教学示范中心）　西藏拉萨　850000

**摘　要：**为解决藏语语音情感数据集稀缺问题，一个面向藏语语音情感数据集TSEC-4528（Tibetan Speech Emotion Corpus-4528）被构建，涵盖语料设计、录制环境与过程及语音数据处理等内容，语音文件分为愤怒、恐惧、快乐、中性、悲伤五大情感标签。通过录音法采集卫藏方言情感语音，经筛选保留4528条数据。经Kappa算法评价法分析，样本Kappa值为0.74。在CNN、DNN、LSTM三种基础模型上，该数据集准确率分别为63.86%、69.43%、70.42%。该数据集旨在补充藏语数据库，提升藏语语音情感识别泛化性，推动其在藏区的应用与发展。

**关键词：**藏语；语音情感识别；数据集构建；有效性验证

### 引言

语音情感识别是实现自然、智能化人机交互的核心技术，其性能的提升高度依赖于大规模、高质量标注的语音数据集[1]。近年来，针对英语、汉语等主流语言的语音情感数据库已得到充分发展，并成功推动了相关算法的研究与实际应用。然而，对于我国丰富的少数民族语言，特别是藏语而言，公开可用的标准化语音情感数据仍极为匮乏。藏语方言体系复杂，内部差异显著，这种数据缺失不仅制约了藏语语音情感识别技术的算法研究与模型训练，更直接阻碍了其在智慧教育、远程医疗、文化遗产数字化等特定领域服务于藏区社会的潜力挖掘。

针对上述问题，我们以使用广泛、具有代表性的卫藏方言（拉萨话）为基础，设计并构建了一个中等规模的藏语语音情感数据集，命名为TSEC-4528。该数据集包含愤怒、恐惧、快乐、中性、悲伤五种基本情感类别，共计4528条经过严格质量筛选的语音样本。在构建过程中，我们系统规划了语料文本，并在受控的录音环境下，组织母语者进行情感演绎式录制，以确保数据的有效性和规范性。

为验证所构建数据集的质量与可靠性，本研究采用Kappa[2]一致性系数对标注结果进行了量化评估。同时，为了初步探索该数据集的实用价值与基线性能，我们选择了卷积神经网络（CNN）[3]、深度神经网络（DNN）[4]和长短期记忆网络（LSTM）[5]三种经典模型，在其上进行了情感识别实验，以评估不同模型架构在该数据集上的表现。

### 一、数据集构建

本实验挑选了12名精通拉萨话的藏族青年参与，分别为6名男性和女性，且均确认未受感冒或其他可能影响发音正常性的因素影响。为了获取高质量的录音效果，本研究选择了一个安静且无外界干扰的录音环境。鉴于同一情感类别通常存在多种不同的表述术语，我们将不同术语进行统一处理，如“恼火”、“生气”、“怒”等统一表示为“愤怒”这一情感。

在录音环节，参与者被要求按照提供的语料文本进行发音，同时，鼓励并引导他们以自然、流畅的方式表达出愤怒、恐惧、快乐、中性以及沮丧这五种不同的情感状态。要求每位参与者均录制了五种情绪状态，每种情绪状态下贡献100条语音记录，总计收获了6000条原始语音文件。文件为wav格式，单声道，16kHz采样，16位采样深度，码率256kbps。经过质量筛选、内容校验后，最终情感语音的总数确定为4528条，各类别情感语音数量如图1所示，因此，本文将此数据集命名为“TSEC-4528”。

根据部分标注情况记录如表1所示，记录了5个评估者对听到的同一语音所标注的情感类别。

根据计算k值需要，将记录表整理为类别统计矩阵，每一行是一个样本，每一列是该类别被评估者选中

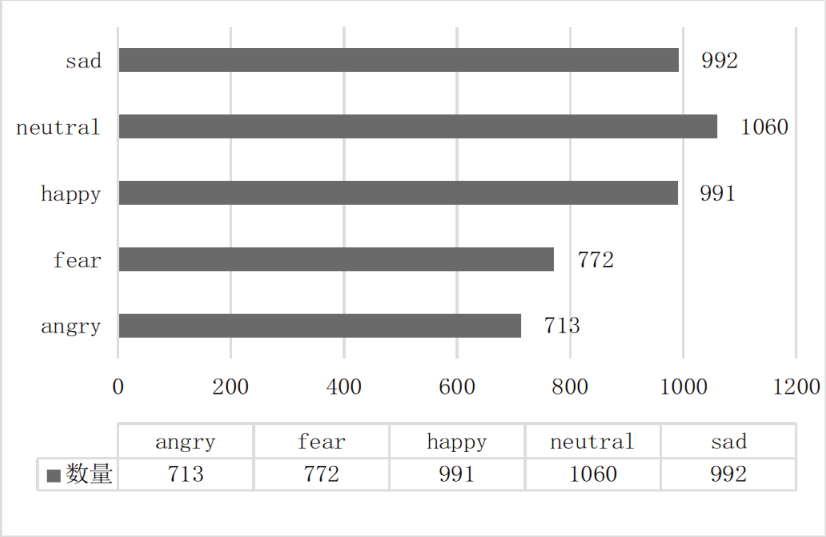


图1　TSEC-4528中各类型情绪数量

表1　部分语音标注情况

| 文件名 | 实际标签 | 评估者1 | 评估者2 | 评估者3 | 评估者4 | 评估者5 |
|---|---|---|---|---|---|---|
| 1.wav | 沮丧 | 沮丧 | 沮丧 | 沮丧 | 沮丧 | 沮丧 |
| 2.wav | 愤怒 | 愤怒 | 愤怒 | 愤怒 | 愤怒 | 愤怒 |
| 3.wav | 快乐 | 快乐 | 愤怒 | 快乐 | 快乐 | 快乐 |
| 4.wav | 恐惧 | 恐惧 | 恐惧 | 恐惧 | 恐惧 | 恐惧 |
| 5.wav | 沮丧 | 沮丧 | 沮丧 | 沮丧 | 沮丧 | 沮丧 |
| … | … | … | … | … | … | … |
| 360.wav | 恐惧 | 恐惧 | 恐惧 | 恐惧 | 恐惧 | 恐惧 |

的次数。原本k值体现的是多为评估者对同一样本的一致性，一般是对样本未知的情况，由于本数据集是采用表演型采集的情感数据，即标签已知的情况，k是为了验证标签的可靠性与合理性，本实验将原标签当成第六位“评估者”。k越接近1[6]，说明评估者整体越贴近真实标签。

表2　Kappa系数指标解释

| Kappa系数 | 一致性程度解释 |
|---|---|
| 0.00到0.20 | 极低的一致性 |
| 0.21到0.40 | 一般的一致性 |
| 0.41到0.60 | 中等的一致性 |
| 0.61到0.80 | 高度的一致性 |
| 0.81到1.00 | 几乎完全一致 |

本研究所设计的藏语语音情感数据集所抽样的360个样本的k约为0.74。参照表2中的Kappa评价指标，可以看出，标注者与本设计的藏语语音情感数据集之间的一致性达到了一个相对较高的层次。

## 二、实验结果与分析

实验采用本文建立的藏语语音情感数据集（TSEC-4528），其中含有5大类情绪的语音数据共4528条。按照约4：1的比例将数据集划分成训练集和测试集，详细信息见表3。本文将基线实验所用的MFCC特征（采用openSMILE提取，39维，FS=12.5ms，FL=50ms）。

表3　实验中情感数据集详细占比

| 情感数据 | angry | fear | happy | neutral | sad | 总计 |
|---|---|---|---|---|---|---|
| 训练集 | 570 | 618 | 791 | 849 | 794 | 3622 |
| 测试集 | 143 | 154 | 198 | 213 | 198 | 906 |

为了探讨分析本文构建的数据集普适性与模型对其的识别率，本实验选择了卷积神经网络（Convolutional Neural Networks，CNN）、深度神经网络（Deep Neural Networks，DNN）和长短期记忆网络（Long Short-Term Memory，LSTM）三种基础神经网络模型对此数据集进行情感识别实验。

实验程序部署配置Intel（R）Xeon（R）Gold 5117、CPU @2.00GHz、128G RAM、Nvidia Tesla P40上运行。开发语言为Python，深度学习框架为Tensorflow。实验模型参数，将训练epoch设置为120，LSTM模型中hidden设置为128，batch size设置为32，学习率设置为0.001，优化器使用Adam。CNN模型中卷积层为3，滤波器数量设置为32。

为了评估本文构建的TSEC-4528数据集的效果，本文使用了三种不同的模型，分别是CNN、DNN和LSTM，实验结果如表4所示。

表4　TSEC-4528数据集在不同模型上的表现

| 模型/评价指标 | Macro-P | Macro-R | Macro-$F_1$ | 准确率 |
|---|---|---|---|---|
| CNN | 65.40% | 63.60% | 63.58% | 63.86% |
| DNN | 69.82% | 69.49% | 69.62% | 69.43% |
| LSTM | 70.50% | 70.40% | 70.43% | 70.42% |

对比整体发现，三种模型在该数据集上的测试结果均表现良好，LSTM网络表现最为突出，分类准确率达到了70.42%。与此相比，DNN模型的表现仍然不错，准确率为69.43%，而CNN在处理该数据集时表现较差，准确率仅为63.86%，这可能是由于卷积结构对局部细节特征的捕捉能力存在局限。

## 结论

本研究成功构建并系统验证了藏语语音情感数据集TSEC-4528。该数据集以卫藏方言为基础，涵盖愤怒、恐惧、快乐、中性和悲伤五种情感类型，共包含4528条高质量语音样本。通过科学的评估方法，数据集显示出较高的标注一致性与质量可靠性（Kappa=0.74）。在基于LSTM、DNN和CNN三种神经网络模型的基线实验中，取得了最高70.42%的分类准确率，证明该数据集能够有效支持藏语语音情感识别的研究与应用，为解决藏语情感数据匮乏问题提供了重要的基础资源。

然而，本研究构建的数据集仍存在若干局限性：方言覆盖局限于卫藏方言，未能全面反映藏语方言的多样性；仅包含单模态语音数据，应用场景相对受限；采用表演型录制方式，情感表达的自然性有待提升；数据规模相对较小，制约了复杂模型的训练效果。基于这些发现，未来研究应着力构建覆盖卫藏、安多、康巴三大方言的综合性数据库，探索文本、语音、视频相结合的多模态情感数据集，采用更自然的数据采集方法提升情感真实性，并通过持续采集不断扩增数据规模，以推动藏语语音情感识别技术向更广泛、更深入的应用场景发展。

## 参考文献

[1]李良琦，张雪英，段淑斐，等.普通话多模态情感语音数据库构建与评测[J].复旦学报（自然科学版），2024，63（1）：18-31.

[2]FLEISS J L. Measuring nominal scale agreement among many raters[J]. Psychological Bulletin, 1971, 76(5): 378-382. DOI:10.1037/h0031619.

[3]栾春.高校学生入党情感多模态数据集的构建与应用[D].济南：山东师范大学，2024.

[4]Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545. DOI: 10.1109/TASLP.2014.2339736.

[5]LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521: 436-444. DOI: 10.1038/nature 14539.

[6]Wöllmer M, Kaiser M, Eyben F, et al. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework [J]. Image and Vision Computing, 2013, 31(2): 153-163. DOI: 10.1016/j.imavis.2012.03.001.