基于蒙古文文本语义辅助的噪声鲁棒蒙古语语音情感识别方法研究

刘欢，梁凯麟，左昊麟，刘瑞 *
内蒙古大学计算机学院，内蒙古，中国
happinessmonster@163.com, liangkailin98@foxmail.com
zuohaolin_0613@163.com, liurui_imu@163.com

摘要

噪声环境下语音情感识别（Speech Emotion Recognition，SER）旨在从带有背景噪声的语音信号中挖掘情感特征并自动预测说话人的情感状态。尽管这项技术在英语、汉语等语言方面取得了迅速的进展，但对于像蒙古语这样的小语种，在噪声环境下的语音情感识别研究仍处于起步阶段，缺乏相关数据集和方法的研究。为了推动蒙古语语音情感识别的发展，本研究首先构建了一个单说话人语音情感识别数据集。之后为了实现噪声环境下准确的蒙古语语音情感识别，我们提出了一种基于文本-语音双模态的带噪蒙古语语音情感识别基线模型 MonSER。文本信息为噪声语音信号提供额外的语义信息。具体来说，我们的模型首先对带噪语音信号进行频谱特征提取，之后使用多语种预训练模型 XLMBert 对语音信号对应的蒙古文文本信息进行编码。随后将上述提取的双模态信息进行融合，并输入分类器进行情感类别的预测。我们利用该数据集进行模型训练并测试模型的有效性。实验结果表明，我们的双模态模型在多种噪声环境下的蒙古语语音情感识别准确率明显优于只以语音为输入的单模态语音情感识别系统。同时，为了模拟实际场景中文本可能缺失的情况，我们提出了两种文本 mask 策略，该文本实验也进一步验证了文本语音双模态的有效性。

关键词： 蒙古语语音情感识别；噪声；文本

Research on Noise-Robust Mongolian Speech Emotion Recognition Methods Based on Mongolian Text Semantics

Huan Liu, Kailin Liang, Haolin Zuo, Rui Liu *
Inner Mongolia University, School of Computer Science, Inner Mongolia, China
happinessmonster@163.com, liangkailin98@foxmail.com
zuohaolin_0613@163.com, liurui_imu@163.com

Abstract

Speech Emotion Recognition (SER) in noisy environments aims to extract emotional features from speech signals with background noise and automatically predict the

* 通信作者：刘瑞（liurui_imu@163.com）

speaker's emotional state. While this technology has rapidly advanced in major languages like English and Chinese, research on SER in minor languages such as Mongolian is still in its early stages, lacking relevant datasets and methodologies.To advance Mongolian Speech Emotion Recognition (SER), this study initially constructed a single-speaker SER dataset. Subsequently, to achieve accurate Mongolian SER in noisy environments, we proposed a dual-modality Text-Audio-based Noisy Mongolian Speech Emotion Recognition baseline model (MonSER).Textual information provides additional semantic information for noisy speech signals. Specifically, our model first extracts spectrogram features from noisy speech signals, then encodes the corresponding Mongolian text information using the multilingual pre-trained model XLMBert. The extracted dual-modal information is then fused and inputted into a classifier for emotional category prediction. We trained the model using this dataset and tested its effectiveness.Experimental results demonstrate that our dual-modality model achieves significantly higher accuracy in Mongolian Speech Emotion Recognition across various noise environments compared to single-modal speech emotion recognition systems that rely solely on speech inputs. Furthermore, to simulate the potential absence of text in real-world scenarios, we proposed two text masking strategies. This textual experiment also further validates the effectiveness of the text-speech bimodal approach.

Keywords: Mongolian Speech Emotion Recognition , Noise , Text

1 简介

带噪语音情感识别 (Speech Emotion Recognition, SER) 是一项旨在从背景噪声中的语音信号提取情感特征并预测说话人情感状态的技术。不同于传统的语音情感识别任务，带噪 SER 从现实应用的角度出发 (Win and Khine, 2020)，模拟了真实世界中的嘈杂环境。其目标在于在复杂的噪声环境中准确识别用户情感的状态 (Nam and Lee, 2021)(Tiwari et al., 2020)，从而帮助机器更好地理解用户的情感需求，为用户提供更加智能和贴心的服务 (Chatterjee et al., 2021)。它在各种人机交互场景 (Chenchah, Lachiri, 2016)(Alnuaim et al., 2022) 中具有巨大的应用潜力。例如，智能助手 (Xu, 2019)、客服机器人 (Bi et al., 2022)(Lee et al., 2020)、情感驾驶辅助系统 (Haghighat et al., 2023)、远程教育与会议 (Barron-Estrada et al., 2019)(Wang et al., 2020)(El Hammoumi et al., 2018)、健康监测和辅助诊断 (Singh and Srivastava, 2023)(Zisad et al., 2020) 等领域中。因此，这一领域的研究具有重要的意义和价值。

在现有的带噪语音情感识别研究中，面临着两个关键的挑战：1）任务相关的数据集支持；2）噪声环境下鲁棒的情感识别方法。为了更好地支持带噪 SER 的研究，先前的研究人员在数据集和方法论方面进行了大量的工作。在数据集方面，主流语言的情感语音库较为丰富，如 Maribor(Ambrus, 2000)、Belfast(McGilloway, 2000)(Douglas-Cowie, 2000)、EMO-DB(Burkhardt et al., 2005)、CASIA(Liu et al., 2011)、IEMOCAP(Busso et al., 2008) 等数据集，涵盖了英语、斯洛文尼亚语、法语和西班牙语、德语、中文等多种语言，情感种类也从 5 类情感到 8 类情感不等；然而，这些数据集通常只提供情感标签，没有对音频中的噪声信息进行标注，同时也不支持

蒙古语的语音情感识别任务。另一方面，尽管刘瑞 (刘瑞 et al., 2022) 等人构建了当前最大规模的蒙古语语音合成数据库，以及刘志强 (Liu et al., 2022) 等人构建了一个针对蒙古语语音识别的 IMUT-MC 语音语料库，但却没有专门针对蒙古语带噪语音情感识别领域的数据集。在方法论方面，一些工作利用语音增强或深度学习的网络架构进行带噪 SER 任务。具体而言，SR(Bandela and Kumar, 2021) 等人采用了密集的非负矩阵分解方法来对嘈杂语音信号进行去噪处理，并用支持向量机分类器对音频进行情感分类 (Jain et al., 2020)(Al Dujaili et al., 2021)。Tawari 等人 (Tawari and Trivedi, 2010) 提出了基于音调和强度轮廓的倒谱分析方法，以解决在不同环境（如汽车设置）中准确识别人类情感的挑战。MingkeXu 等人 (Xu et al., 2021) 提出了基于多头注意力机制的头部融合方法，旨在提高情感识别的准确性并增强模型在噪声环境下的鲁棒性。但在蒙古语应用场景下的带噪 SER 研究却相对滞后。这主要是因为蒙古语是一种使用人数较少的语言，因此相关的研究并不多，这导致缺乏专门针对蒙古语特性优化的情感识别模型和算法。上述问题为带噪情感识别的准确性和模型的泛化能力带来了额外的挑战。

本研究首次构建了一个专门针对蒙古语的带噪语音情感识别数据集 Noisy Mongolian Speech Emotion Recognition Dataset（NMonSER）。该数据集包含 1001 条蒙古语音频数据，覆盖 8 种情绪：生气、乏味、厌恶、高兴、自然、悲伤、害怕和惊讶，涵盖了文本、声音双模态。这些数据由一位 22 岁的专业蒙古语播音员在标准录音室内录制，并使用 Adobe Audition 软件处理，确保高质量录音。为了贴近生活中带噪环境下的音频，我们将录制的数据在不同信噪比（10db、5db、0db、-5db、-10db）下添加了 Noisex-92 标准噪声库 (Varga and Steeneken, 1993) 中的典型噪声（如 Pink 噪声、Babble 噪声等），形成了适用于噪声环境下的蒙古语音频数据集 MonSER-DN。此数据集通过精细的预处理和特征提取流程（包括音频重采样、短时傅里叶变换、Mel 频率转换等）生成了 3D 特征表示，旨在支持复杂环境下的情感识别研究。此外，文本预处理模拟了噪声环境下文本信息缺失的情况，使得数据集更贴近真实应用场景。数据集的构建考虑了真实环境中的多种挑战，如噪声干扰和信息缺失，使其成为研究蒙古语情绪识别的有价值资源。

此外，我们还提出了一种针对带噪蒙古语语音情感识别的多模态分析方法 MonSER，旨在实现带噪环境下的蒙古语语音情感识别。具体来说，我们主要通过文本特征提取、音频特征提取和模态融合三个核心模块来实现。在文本特征提取阶段，采用了 XLMBert 预训练模型 (Yuyang and Wu, 2022)，该模型通过 SentencePiece 分词和位置编码，有效捕捉蒙古语文本的语义信息。音频特征提取模块则通过卷积神经网络深入挖掘预处理音频信号的时频特征，强化了模型对音频时序变化的感知能力。最后，模态融合模块通过结合文本和音频的嵌入表示，并经过分类器进行分类，实现了高效的情感识别。整个模型不仅深入考虑了文本和语音特征的处理，还通过多个实验证明，我们的方法增强了情感识别的准确性和鲁棒性，展现了多模态情感分析在带噪蒙古语情感识别中的应用潜力。

综上，本研究不仅填补了带噪蒙古语语音情感识别领域的研究空白，而且通过开发基于语音-文本双模态的带噪语音情感识别模型 MonSER，展现了多模态情感分析在带噪蒙古语情绪识别中的应用潜力。我们的研究目标是通过自动学习最佳特征表示，有效捕获数据内部隐藏的情感信息，从而促进蒙古语技术的发展，更好地服务于少数民族人民。

本文的主要贡献如下：

1. 我们首次构建了一个面向蒙古语的单一说话人带噪语音情感识别数据集 NMonSER，填补了带噪语音情感识别领域在小语种研究方面的空白。这一数据集的构建为未来的蒙古语语音

情感识别研究提供了重要的资源。

2. 我们进一步提出了蒙古语带噪语音情感识别模型 MonSER，实现了带噪环境下的蒙古语情感识别，并且在模型中首次使用了 XLMBert 去提取原始蒙古语文本中的情感信息，提高了系统的鲁棒性。

3. 实验结果表明，我们的 MonSER 模型在各种噪声环境下对蒙古语语音情感识别的准确率明显高于仅使用语音作为输入的单模态语音情感识别系统，并且关于文本比例的实验也进一步验证了我们所提出方法的有效性。

2 数据集构建及处理

在本小节中，我们将详细介绍数据集的构建过程，包括文本收集、音频录制以及噪声添加等步骤。文本收集及音频录制等具体细节将在 2.1 小节中进行详细介绍。带噪音频制作细节等将在 2.2 小节详细介绍。

2.1 文本收集及音频录制

为构建 NMonSER 数据集，首先要进行文本信息的收集。选择文本的主题需要侧重于能够引发和表现出丰富情感的场景。我们从现有的蒙古文电子书、新闻以及相关影视剧中共筛选并收集了常用蒙古文 1001 条，涉及到日常生活、文化、体育、娱乐等方面的内容，这些文本中包含清晰的情感指向性，能够明确地表达或引起特定情感反应，有助于构建一个全面且效果良好的情感语料库。整个语料库数据统计情况如表 1 所示，这些文本共 (Total) 包含 82556 个字符，平均 (Mean) 每句 82 个字符，最短 (Min) 句 14 个字符，最长 (Max) 句 238 个字符；总共含有 13725 个单词，平均每句话单词量为 14 个，最大单词数量 44 个，最小单词数量 2 个。所有文本以 UTF-8 编码保存在 TXT 文件中。

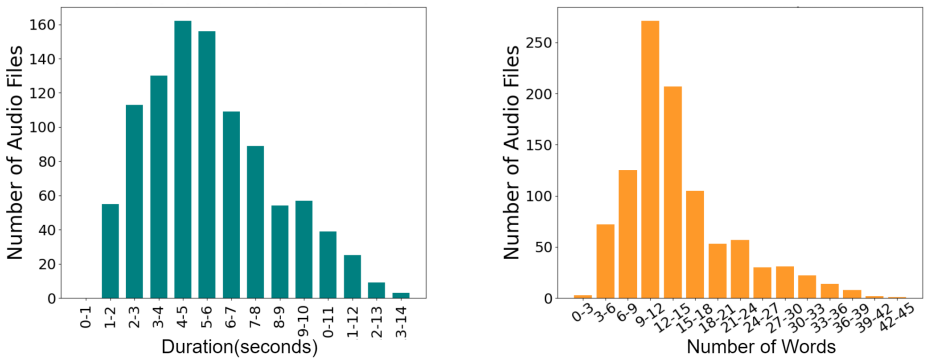
Category	Character				Word			
Statistics	Total	Mean	Min	Max	Total	Mean	Min	Max
	82556	82	14	238	13725	14	2	44

表 1: NMonSER 数据统计详情表

在语音情感识别任务中，我们避免选择过长的句子以减少情感状态的波动，确保模型更好地学习和识别情感状态。因此，如图 1 所示，我们的句子持续时间主要集中在 4 到 6 秒之间，单词数量集中在 9 到 15 之间。为了录制音频，我们邀请了一位 22 岁的蒙古语专业播音员，并有一位志愿者负责检查语法和情感表达的准确性。录音过程在标准录音工作室中进行，使用 Adobe Audition 软件进行录制。我们录制了 8 种情感的纯净音频数据，分别为生气 72 条、乏味 73 条、厌恶 71 条、高兴 72 条、自然 501 条、悲伤 72 条、害怕 72 条和惊讶 70 条，共计 1001 条蒙古语音频数据。最终录制数据总时长约 1.57 小时，所有音频以 WAV 格式存储，采样率为 44.1kHz，采样精度为 16bit。

2.2 噪声添加及语料库结构

为了生成不同信噪比的带噪情感语音，首先需要准备生活中的常见噪音数据和干净的情感语音数据。我们选取的噪声样本来源于 Noisex-92 标准噪声库，其中包括 Pink 噪声、Babble 噪



（a）MonSER-DN语料库时间长度分布直方图　（b）MonSER-DN语料库单词个数分布直方图

图 1: NMonSER 语料库数据统计图

声、M109 噪声和 F16 噪声等典型噪声样本。干净的情感音频选择上述 2.1 小节中所介绍的 1001 条蒙古语纯净的情感音频数据。其次，我们对数据集进行预处理。为了确保干净音频与噪声频率相匹配，我们将干净音频下采样至 16kHz。又由于噪声音频和干净音频的时长不能一一对应，所以我们进行了音频数据对齐操作：若干净音频数据的长度大于噪声数据长度，则利用向上取整的方法将噪声数据重复多次，直到与干净音频数据长度相匹配，确保噪声数据长度足够覆盖干净音频数据；若噪声数据的长度大于干净音频的长度，则将噪声数据截取到与干净音频数据相同的长度；若干净音频数据和噪声数据长度相同，则直接进行后续的带噪音频生成工作。接着，我们分别计算得到干净音频信号功率 P_s 和噪声功率 Pn，具体计算如下：

$$P_s = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2, P_n = \frac{1}{M}\sum_{m=0}^{M-1}|y(m)|^2 \tag{1}$$

其中，$x(n)$ 表示干净音频信号，$y(m)$ 表示噪声音频信号，N 是干净音频信号的长度，M 是噪声音频信号的长度。然后我们通过不同信噪比（10dB、5dB、0DB、-5dB、-10dB）计算缩放系数 k，用于调整噪声音频的幅度，使其符合指定的信噪比，具体的计算公式如（3）：

$$k = \sqrt{\frac{Ps}{\frac{Pn}{10}(SNR/10)}} \tag{2}$$

其中，SNR 是信噪比，k 是用于调整噪声幅度的缩放系数，Ps 是干净音频信号功率，Pn 是噪声功率。最后，我们将调整好后的噪声数据与干净音频数据相加，得到混合后的音频数据。

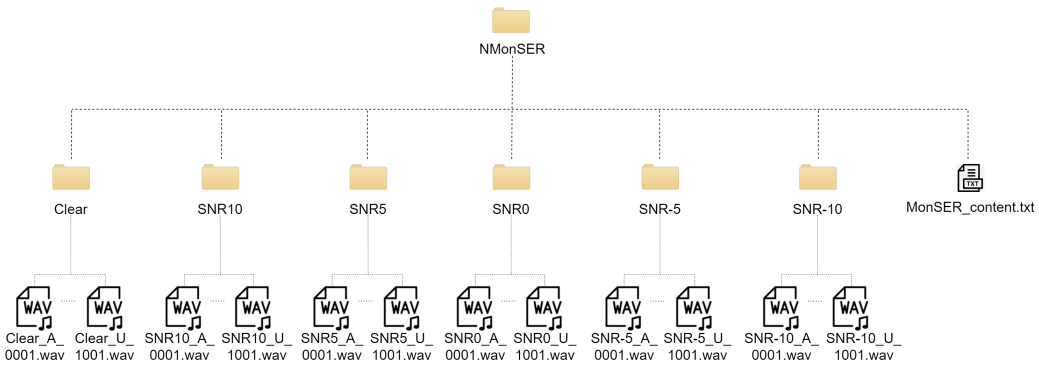


图 2: NMonSER 语料库结构图

NMonSER 语料库的文件结构如图 2 所示，将叠加噪声后的音频数据保存为 WAV 格式的文件，采样率为 16kHz，编码为 16 位。为了更好地管理数据，我们根据不同的信噪比对这些文件进行分类，并将它们放置在相应命名的文件夹中，每个文件夹中有 1001 条音频。此外，我们使用了 ID 序列来命名文本语料，并将其存储在名为 MonSER_content 的 txt 文件中。对于音频文件的命名，则是由信噪比、情绪标签以及文本 ID 序列组合而成。这样的文件管理结构有助于我们更加清晰地组织和管理 NMonSER 语料库中的数据。

3 模型架构

在本节中，提出并详细解释了所提出的 MonSER 模型，该模型的框架结构如图 3 所示。主要由三个模块所构成，包含蒙古文文本编码器、蒙古语语音编码器和语义增强模块。

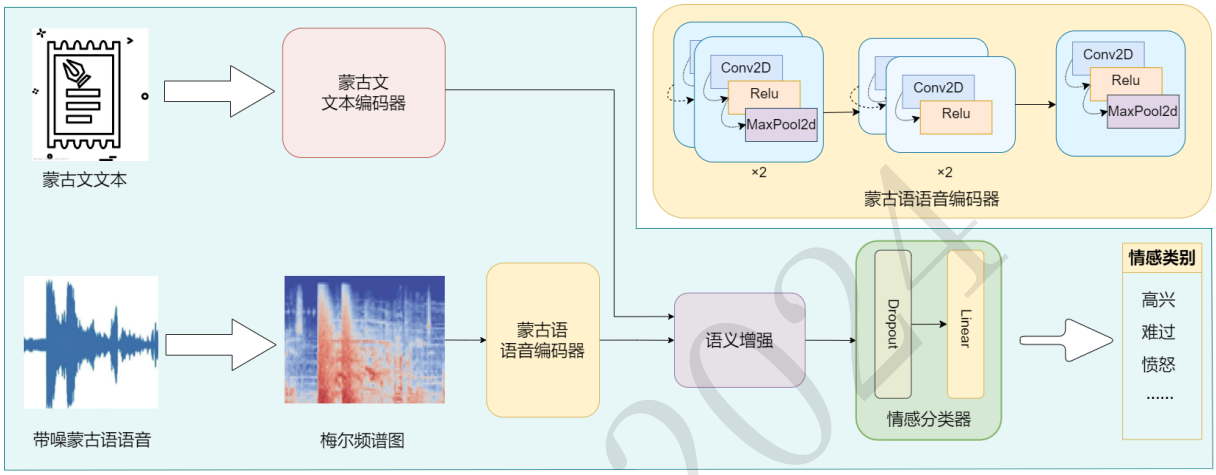


图 3: MonSER 模型结构图

3.1 蒙古文文本编码器

在蒙古文文本编码器中，我们主要运用了 XLMBert 对文本进行处理。XLMBert 是基于 BERT 模型的预训练语言模型，专门用于处理多语言文本，包括蒙古语。它通过在大规模文本语料上进行预训练来学习通用的语言表示，然后可以在各种下游任务上进行微调。XLMBert 对蒙古语的编码过程，大致包括以下步骤：首先，XLMBert 使用 SentencePiece 分词器对蒙古语文本进行分词处理，将文本切分为词片段（subword units）。然后，将分词后的词片段映射为词嵌入向量。XLMBert 模型在预训练过程中学习了词嵌入向量，能够将每个词片段转换为一个固定维度的向量表示，包含了词汇信息和语义信息。接下来，在输入序列中加入位置编码，以区分不同位置的词片段。位置编码通常是一种特定模式的向量，用于表示每个词片段在序列中的位置信息。将词嵌入向量与位置编码相加，得到最终的输入序列表示。这个输入序列会经过多层 Transformer 编码器，进行多层的自注意力机制和前向传播操作，从而逐步提取输入文本的语义特征，该语义特征最终会被表示为 768 维向量。

3.2 蒙古语语音编码器

蒙古语语音编码器使用 3D 频谱图作为输入。相较于 2D 特征，3D 特征在处理时序数据和捕捉时序信息方面具有优势 (Krizhevsky et al., 2012)，能有效捕获数据在时间维度上的演变和

变化，对蒙古语音情感识别至关重要。模块主要定义了卷积神经网络模型，接收经过短时傅里叶变换 STFT 处理的频谱图，通过一系列特征提取层逐步加工 3D 频谱图特征。

具体来说，首层卷积使用了 64 个 11x11 滤波器，步长为 4，填充为 2，旨在捕获 3D 频谱图的低级特征，如边缘和纹理。ReLU 激活引入非线性，最大池化层减小特征图尺寸，保留关键信息。随后，通过更细致的卷积进一步提炼特征，包括第二层 192 个 5x5 滤波器，提高对高级模式的识别能力。第三至五层卷积持续深化对 3D 频谱图的理解，包括第三个卷积层 384 个 3x3 滤波器，以及第四和第五个卷积层各 256 个 3x3 滤波器，旨在捕获更为复杂和抽象的特征，为最终任务提供丰富的特征集。最后的池化层确保特征图的尺寸进一步减小，同时保留最为重要的特征信息。整个特征提取流程有效地从 3D 频谱图中提炼出关键特征，形成 1*256 的数组，作为下一步模态融合的输入条件。通过对每一层的参数和结构优化，模型能够根据具体任务和数据集的需求达到最佳性能。

3.3 语义增强模块

在语义增强模块，主要目标是实现文本和音频信息的有效融合，并对这些融合后的信息进行神经网络处理 (Christy et al., 2020)。首先，将经过编码的文本和音频嵌入表示沿列方向拼接，生成一个综合的特征表示。这种融合方式旨在创建一个全面丰富的特征空间，为后续的情感分类任务奠定基础。接下来，通过 Dropout 层对这个融合的特征表示进行随机的失活处理，以此来减轻模型的过拟合现象。Dropout 通过随机使一部分神经元的输出为零，降低了模型对单一数据特征的依赖，增强了模型的泛化能力。最后，linear 线性层将经 Dropout 处理的特征数据映射到一个新的空间，用于输出最终的分类结果。这一步是通过一个全连接的线性层完成的，它将融合后的特征转换成模型预测所需的具体输出形式，确保了模型可以对不同的信息源进行有效的学习和整合。

综上所述，本节主要介绍了蒙古语语音情感识别设计的 MonSER 模型，包括蒙古文文本编码器、蒙古语语音编码器和语义增强模块。蒙古文文本编码器主要用 XLMBert 模型处理，利用 SentencePiece 分词器进行分词、映射词嵌入向量、添加位置编码，最终提取深层语义特征。蒙古语语音编码器主要利用卷积层等结构提取频谱图特征，捕获语音情感特性。语义增强模块将文本和音频信息结合，形成综合特征，通过 Dropout 层减少过拟合，线性层映射为情感分类输出。这一设计有效地实现了噪声环境下对蒙古语语音情感的识别。

4 实验设计

4.1 数据预处理

数据预处理主要包括对原始音频的预处理和蒙古语文本的预处理。首先，对原始音频进行标准化处理，生成对应的频谱图以及其对应的 3D 特征表示。其次，对蒙古语文本进行处理，在实际环境中，由于噪声的存在，获取的文本信息可能会有一定程度的缺失，因此我们分别以不同比例和不同单词个数将文本中的单词替换为'[mask]'，从而模拟文本信息缺失的情况。

4.1.1 音频预处理

对原始音频进行预处理时，首先使用 Python 中 SciPy 库中的 wavfile.read 函数，获取音频文件的采样率和音频数据，对其进行重采样、分帧等预处理。然后对每一帧的音频信号进行短时

傅里叶变换（STFT），得到其频域信息，其离散形式的计算公式为：

$$X(m,\omega) = \sum_{n=0}^{N-1} x(n)w(n-m)e^{-j\omega n} \tag{3}$$

其中，$X(m,\omega)$ 表示频率 ω 处在时间片段 m 的频谱值，$x(n)$ 是输入信号的离散时间序列，$w(n-m)$ 是窗口函数，N 是窗口长度。

然后使用 Mel 滤波器组对频域信息进行加权，得到每个 Mel 频率段的能量，计算公式为：

$$M(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{4}$$

其中，$M(f)$ 表示频率 f 对应的 Mel 频率。

接着再将每个 Mel 频率段的能量取对数，得到 Log-Mel 频谱图，计算公式为：

$$log-Mel(f) = \log(1+\frac{E(f)}{E_0}) \tag{5}$$

其中，$E(f)$ 表示 Mel 频率段 f 的能量，E_0 是一个小常数，用于防止对数变换时出现零值。

最后，为了更好的处理时序数据和捕捉时序信息，我们根据提供或默认的均值和标准差，对输入的频谱图进行标准化处理，将得到的语谱图转换为 3D 特征表示，用于输入模型中进行训练和预测。

4.1.2 文本预处理

在实际应用场景中，面对环境噪音等因素的干扰，我们常常会遇到文本信息不完整的挑战(Fan et al., 2023)(Liu et al., 2024)。为了模拟这种情况并深入理解在含噪环境下处理文本的复杂性，我们采纳了一项独特的方法。具体来说，我们通过将文本中的文字以不同的比例（20%、40%、60%、80%）替换为'[mask]' 标记，有效地重现了文本信息丢失的情境。具体操作如图 4(a) 所示。

然而，这种方法在处理少于 5 个词的短句时遇到了困难，主要表现在对短句进行'[mask]' 替换时，不同比例下可能出现相似或相同的情况。例如，对于仅含有四个词的句子，当替换比例为 20% 时，可能会与原句完全相同；对于三个词的句子，20% 的替换比例下也可能出现相同的现象，而 40% 和 60% 的替换比例则呈现相似的替换效果；最极端的情况是，对于只有两个词的句子，在 20% 和 40% 的替换比例下可能完全不变，而在 60% 和 80% 的替换比例下也可能出现相似的替换效果。这种现象可能导致 XLMBert 在处理不同文本缺失比例的情况下生成相似或相同的编码，进而在含噪声环境下的语音情感识别中得到类似的准确率，影响对模型性能的评估。为了应对这一挑战，我们进一步筛选了长度超过 5 个词的数据样本，并采用更细致的掩码策略，即在每个句子中随机替换 1 到 5 个词，具体操作如图 4(b) 所示。通过这种方法生成的新数据集，我们能够更准确地模拟和训练模型，以提升结果的可信度。

4.2 参数配置

在本实验中，我们采用了 XLMBert 对文本数据进行处理和编码，提取的文本特征为 768 维。音频特征提取部分采用了一系列卷积神经网络层和最大池化层，从输入的音频数据中提取特征，音频特征维度为 256 维。接着，我们将融合后的文本和音频特征映射到 1024 维度的特征空间，通过全连接层进行分类，得到最终的输出结果。具体模型参数如下：首先是卷积层，输入通道为 3，输出通道为 64，卷积核大小为 11x11，步长为 4，填充为 2，采用 ReLU 激活函数。

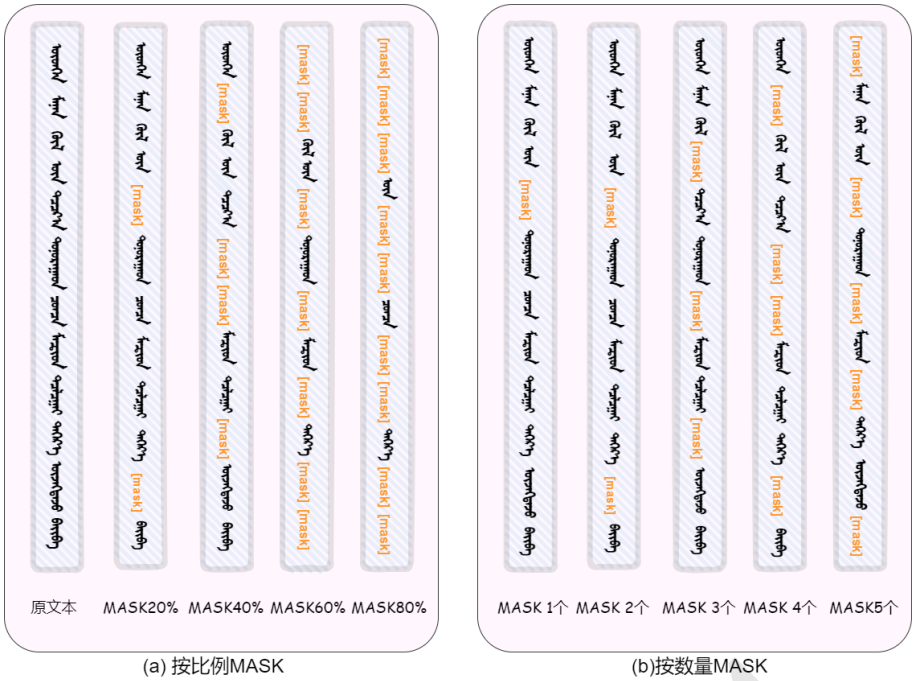


图 4: 蒙古文文本 mask 策略，(a) 按比例 MASK (b) 按数量 MASK

接着是最大池化层，池化核大小为 3x3，步长为 2。之后的卷积层和 ReLU 激活函数与前述类似，用于进一步提取和加工音频特征。紧接着是自适应平均池化层，将特征图大小调整为 12x12。然后是全连接线性层，将融合后的文本和音频特征映射到 1024 维度的特征空间。为了防止过拟合，我们使用了 Dropout 层，在训练过程中随机失活一部分神经元。最后是全连接线性层，用于最终的分类任务输出，输入特征维度为 1280。我们使用了学习率为 0.001 的 Adam 优化器，并采用了交叉熵损失函数。Batch size 设置为 32，Dropout 率为 0.5。初始的 epoch 值设为 150，但我们实际上应用了提前停止（early stopping）函数，在损失连续 20 次未下降时停止模型训练。

4.3 实验结果

实验结果如表 2 所示。第一行展示了干净语音以及不同信噪比下的带噪音频，信噪比分别为 10db、5db、0db、-5db、-10db。第二行表示仅包含语音信息时对应各类带噪音频的语音情感识别准确度。从表格中可以明显看出，随着信噪比逐渐降低，即噪声逐渐增大，情感识别的准确率也显著下降，从干净语音的 0.9108 降至信噪比为-10db 时的 0.6832。

SNR / MASK	clear	10db	5db	0db	-5db	-10db
-	0.9108	0.8989	0.7921	0.7822	0.7426	0.6832
0%	0.9900	0.9703	0.9307	0.9109	0.8400	0.7822
20%	0.9800	0.9600	0.9208	0.9010	0.8300	0.7600
40%	0.9505	0.9307	0.9208	0.8911	0.8218	0.7524
60%	0.9200	0.9109	0.8700	0.8416	0.8119	0.7200
80%	0.9192	0.9091	0.8317	0.8300	0.8020	0.7129

表 2: 蒙古文按比例 MASK 与不同信噪比音频上的情感识别准确率

表格的第一列显示了无文本信息以及在噪声环境下丢失的文字信息比例，分别达到 0%、20%、40%、60%、80%。加入文本信息后，各类噪声环境下的准确率明显提升，精度提高了 7.14% 到 13.86%。这表明即使在文本信息有所缺失的情况下，各类噪声环境下的情感识别精度仍然高于单一模态下的准确率。然而，在信噪比为 5db 时，当文字缺失比例达到 20% 和 40% 时，对应的精度是相同的。这种现象可以归因于之前提到的短句问题。当句子过短时，对句子按不同比例进行 mask 可能导致 mask20% 和 mask40% 所对应的文本有效信息相近，这可能导致在同一信噪比下，不同 mask 率得到的准确率相同的情况发生。

为了解决上述现象对模型评估的影响，我们采取了一项重要措施：剔除长度小于等于 5 的句子，并仅保留了长度大于 5 的 959 条有效数据，随后再次进行分类实验。具体的结果如表 3 所示。在这个表格中，第一行与之前的表格相同，第一列表示在一句话中丢失的单词数量，范围从 1 个至 5 个递增。

SNR / MASK WORDS	clear	10db	5db	0db	-5db	-10db
0	0.9900	0.9792	0.9583	0.9375	0.8646	0.8021
1	0.9896	0.9583	0.9479	0.9271	0.8438	0.7917
2	0.9688	0.9479	0.9271	0.9062	0.8333	0.7708
3	0.9583	0.9375	0.9167	0.8958	0.8125	0.7500
4	0.9375	0.9271	0.9062	0.8542	0.7917	0.7188
5	0.9271	0.9167	0.8958	0.8333	0.7604	0.7083

表 3: 蒙古文按单词个数 MASK 与不同信噪比音频上的情感识别准确率

通过按照缺失单词个数进行分类实验，我们观察到在同一信噪比下，不同 mask 率得到的准确率不再相同。这说明我们的模型在噪声环境下进行蒙古语语音情感识别的鲁棒性较好，并且具有较高的精度。这一实验结果进一步验证了我们所提出的模型在处理噪声环境下的可靠性和稳健性。

5 结论

综上所述，我们的基于文本-语音双模态的带噪蒙古语语音情感识别模型在实验中取得了较为不错的结果。通过构建数据集、优化特征提取和融合策略，我们成功提高了蒙古语情感识别的准确性和鲁棒性。具体来说，我们通过积极收集和整理蒙古语情感数据，构建了一个适用于情感识别研究的重要资源，为该领域的进一步发展提供了基础。在模型设计方面，我们充分利用了文本信息和语音特征的双模态优势，通过 XLMBert 模型对蒙古语文本进行编码，并结合音频特征提取器对频谱图进行分析，从而实现了对语音情感的精准识别。这项工作不仅在技术上取得了重要进展，也为蒙古语情感识别领域的研究和实际应用带来了新的启示。我们的成功经验和成果将为未来相关工作提供有益的借鉴和参考，推动语音情感识别技术在蒙古语言言等小语种领域的不断创新和应用。

第二十三届中国计算语言学大会论文集，第268页-第279页，太原，中国，2024年7月25日至28日。
卷1：主会论文

致谢

感谢国家自然科学基金青年科学基金项目（N0.62206136）；广东省数字孪生人重点实验室（华南理工大学）开放课题（2022B1212010004）；“一区两基地”超算能力建设项目（内蒙古大学）（项目编号：21300-231510）的资助。

参考文献

Chenchah, Farah, and Zied Lachiri. 2016. Speech emotion recognition in noisy environment. In *Proceedings of the 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE.

Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *interactions*, 26(4):42–46.

Wei Bi, et al. 2022. Enterprise strategic management from the perspective of business ecosystem construction based on multimodal emotion recognition. *Frontiers in Psychology*, 13:857891.

Parian Haghighat, et al. 2023. Effects of an intelligent virtual assistant on office task performance and workload in a noisy environment. *Applied Ergonomics*, 109:103969.

Maria Lucia Barron-Estrada, Ramon Zatarain-Cabada, and Raul Oramas Bustillos. 2019. Emotion Recognition for Education using Sentiment Analysis. *Res. Comput. Sci.*, 148(5):71-80.

Weiqing Wang, et al. 2020. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. *Complexity*, 2020:1-9.

Oussama El Hammoumi, et al. 2018. Emotion recognition in e-learning systems. In *Proceedings of the 2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, IEEE.

Sonali Singh and Navita Srivastava. 2023. Emotion recognition for mental health prediction using AI techniques: an overview. *International Journal of Advanced Research in Computer Science*, 14(3).

Daniel Cen Ambrus. 2000. Collecting and recording of an emotional speech database. Maribor, Slovenia: University of Maribor.

Sinéad McGilloway, et al. 2000. *Approaching automatic recognition of emotion from voice: A rough benchmark*. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Ellen Douglas-Cowie, Roddy Cowie, and Marc Schrö der. 2000. *A new emotion database: considerations, sources and scope*. ISCA tutorial and research workshop (ITRW) on speech and emotion.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Wolfgang Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Interspeech*, Vol. 5.

Cheng-Lin Liu, Fei Yin, and Dong-Yan Huang. 2011. CASIA online and offline Chinese handwriting databases. In *2011 International Conference on Document Analysis and Recognition*, IEEE.

Carlos Busso, Murtaza Bulut, et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335-359.

刘瑞, 康世胤, 李劲东, 飞龙, 高光来. 2022. MonTTS: 完全非自回归的实时, 高保真蒙古语语音合成模型. 中文信息学报, 36(7):86–97.

赵建东, 高光来, 飞龙. 2014. 基于 HMM 的蒙古语语音合成技术研究. 计算机科学, 41(1):80–82.

Zhiqiang Liu, et al. 2022. IMUT-MC: A speech corpus for Mongolian speech recognition. *China Sci. Data*, 7:13.

Surekha Reddy Bandela and T. Kishore Kumar. 2021. Unsupervised feature selection and NMF denoising for robust Speech Emotion Recognition. *Applied Acoustics*, 172:107645.

Ashish Tawari and Mohan Manubhai Trivedi. 2010. Speech emotion analysis: Exploring the role of context. *IEEE Transactions on Multimedia*, 12(6):502-509.

Mingke Xu, Fan Zhang, and Wei Zhang. 2021. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access*, 9:74539-74549.

Andrew Varga and Herman JM Steeneken. 1993. *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems.* Speech communication, 12(3): 247-251.

Deng Yuyang and Dan Wu. 2022. *Chinese-Tibetan Bilingual Named Entity Recognition for Traditional Tibetan Festivals.* Data Analysis and Knowledge Discovery, 7(7): 125-135.

A. Christy, et al. 2020. *Multimodal speech emotion recognition and classification using convolutional neural network techniques.* International Journal of Speech Technology, 23(2), 381-388.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems 25.

Youngja Nam and Chankyu Lee. 2021. *Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions.* Sensors, 21(13):4399.

Upasana Tiwari, et al. 2020. *Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions.* In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Htwe Pa Pa Win and Phyo Thu Thu Khine. 2020. *Emotion recognition system of noisy speech in real world environment.* International Journal of Image, Graphics and Signal Processing (IJIGSP), 12(2):1-8.

Abeer Ali Alnuaim, et al. 2022. *Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier.* Journal of Healthcare Engineering, 2022.

Qi Fan, et al. 2023. *Learning Noise-Robust Joint Representation for Multimodal Emotion Recognition under Realistic Incomplete Data Scenarios.* arXiv preprint arXiv:2311.16114.

Rui Liu, et al. 2024. *Contrastive Learning based Modality-Invariant Feature Acquisition for Robust Multimodal Emotion Recognition with Missing Modalities.* IEEE Transactions on Affective Computing.

Ming-Che Lee, et al. 2020. *Study on emotion recognition and companion Chatbot using deep neural network.* Multimedia Tools and Applications, 79(27), 19629-19657.

Rajdeep Chatterjee, et al. 2021. *Real-time speech emotion analysis for smart home assistants.* IEEE Transactions on Consumer Electronics, 67(1), 68-76.

Sharif Noor Zisad, Mohammad Shahadat Hossain, and Karl Andersson. 2020. *Speech emotion recognition in neurological disorders using convolutional neural network.* In International Conference on Brain Informatics, Cham: Springer International Publishing.

Manas Jain, et al. 2020. *Speech emotion recognition using support vector machine.* arXiv preprint arXiv:2002.07590.

Mohammed Jawad Al Dujaili, Abbas Ebrahimi-Moghadam, and Ahmed Fatlawi. 2021. *Speech emotion recognition based on SVM and KNN classifications fusion.* International Journal of Electrical and Computer Engineering, 11(2), 1259.