EmoAra：保留情绪的英语语音转录与跨语言翻译，结合阿拉伯语文本转语音

Besher Hassan IbrahimAlsarraj Musaab Hasan Yousef Melhim Shahem Fadi Shahem Sultan

电子邮箱：Besher.Hassan@mbzuai.ac.ae, Ibrahim.Alsarraj@mbzuai.ac.ae

arXiv:2602.01170v1 [cs.CL] 1 Feb 2026

摘要

本项目致力于应对跨语言交流中的挑战，尤其是在银行客户服务领域，因为理解情感语境至关重要。通过消除语言障碍、促进共情沟通并提升客户满意度，该系统展示了其在银行客户服务中的实际应用。系统整合了**语音情感识别（SER）**、**自动语音识别（ASR）**、**机器翻译（MT）**和**文本转语音（TTS）**，实现保留情绪的多语言交流。系统使用**卷积神经网络**进行情感检测，使用**Whisper** 进行英语转录，使用**MarianMT**进行英译阿，并使用**MMS-TTS-Ara** 进行阿拉伯语语音合成。系统处理英语语音、检测情绪、转录并翻译文本，随后生成阿拉伯语语音回复，同时保留情感细微差别。主要结果包括：情感检测的F1分数为**94%**，BLEU分数为**56**，翻译质量的BERTScore F1分数为**88%** ；在翻译准确度、流畅度和领域专用术语方面，人工评估得分为**81%** 。如需了解更多详情，请查看GitHub代码仓库。

1 引言

1.1 项目概述

I在日益全球化的世界中，企业面临着跨越语言和文化障碍进行有效沟通的挑战。本
p项目旨在构建一个情感识别系统，用于客户服务领域，尤其是银行业。通过提取客户
f在口语交互中的情感，然后转录英语音频，将文本翻译成阿拉伯语，并使用文本转语音（TTS）系统
f进行口头呈现。结合新兴语音情感识别（新兴 SER）、自动语音识别（ASR）和机器翻译
(机器翻译（MT）结合文本转语音（TTS），实现多语言交流，同时保留情感完整性。本项目将与语言无关，并
e以情感为导向。情感导向系统是旨在识别和处理内容、理解并保留其中情感语境的系统
c交流。在客户服务场景中，通常会关注客户语音中的情绪基调，这有助于判断客户
n的具体需求和感受。对于银行业而言，这一点的重要性更加突出，因为情绪有助于评估紧迫程度、满意度、
o或不满意。不幸的是，语言障碍以及缺乏实时情感反馈，阻碍了银行代表
p 及时妥善地处理问题。本项目利用情感和语言理解来减轻这些障碍，使该行业能够以情感方式回应任何语言中的情绪化语音。

1.2 主要贡献

本项目的主要贡献可概括如下：

- 跨语言交互中的情感：与许多忽略翻译中情绪基调的现有系统不同，本项目的翻译充分考虑情感，确保原语言中的情绪也能在目标语言中得以体现，从而让客户在接受服务时感到更加自然，也更能感受到共情。

- 银行客户服务中的实际应用：我们的系统一个典型的应用场景是在银行与金融业中实现跨语言沟通，从而提供高效的服务，同时满足理解客户情绪的需求。

第38届神经信息处理系统大会（NeurIPS 2024）。

1.3 相关工作

Doctors Livingstone 和 Russo（2018年）提供了 Ryerson 情感语音与歌曲视听数据库（RAVDESS），该数据库现已成为评估语音情感识别系统的广泛使用的基准 [10]。借助这一数据集，研究人员使用卷积神经网络和循环神经网络，对更加多样的情绪进行了情感识别。

Huang 等人（2021年）展示了如何通过迁移学习过程微调情感检测模型，通过重新训练现有模型来提高准确性 [7]。

Hannun 等人（2014年）推出了 DeepSpeech 系统，该系统利用在海量语音数据集上训练的循环神经网络，并取得了令人瞩目的成果 [5]。近期，Baevski 等人（2020年）提出了 wav2vec 2.0，它支持对网络进行自训练，从而生成比以往采用端到端架构所获得的更优语音表征 [1]。

机器翻译（MT）也取得了飞速进展，其中尤其注重特定的低开销语言模型。Junczys-Dowmunt 等人（2018年）推出了 MarianMT 系统，这是一款快速、强大且设计精良的神经机器翻译系统，并在多个语言对的神经翻译任务中取得了最佳成果 [8]。

多年来，文本转语音（TTS）系统不断发展，Google Cloud 的 TTS 服务等技术能够生成近乎自然的语音。系统支持包括阿拉伯语在内的多种语言，并已应用于不少实际场景。然而，大多数现有 TTS 系统的设计仍将其功能与情感识别和机器翻译系统分开。这使得这些系统在动态且情感复杂的应用场景中效果较差 [2]。

基于这些理念并开展相关研究工作，本项目旨在将自动语音识别、机器翻译、情感语音识别和文本转语音方法整合到一个统一、连贯的系统中。具体而言，我们希望打造一个强大的端到端多语言系统，通过应用迁移学习方法和预训练模型，在语言转换流程中保留内容的情感特征。

2 问题陈述

在当今互联互通的世界中，服务提供者与客户必须懂得如何跨语言有效沟通，这一点至关重要；尤其是在银行业等领域，客户的情感可能是确定所提供服务的重要依据。遗憾的是，大多数自动化客户服务系统都无法检测和解读语音中蕴含的情感，无论语音使用的是首选语言还是其他语言。这一障碍可能导致沟通不畅、客户满意度降低，以及无法开展及时有效的沟通。我们的目标是通过开发一个系统来弥补这一不足：它不仅能够转录和翻译口语，还能保留客户互动中的情感语境。通过实施多语言、情感驱动的语音识别与翻译流程，我们的解决方案将帮助客户服务团队更具同理心、更准确地回应客户需求，不受语言差异的影响。

3 所提方法

3.1 探索性数据分析

3.1.1 翻译数据预处理

为准备翻译数据集以对 Marian 机器翻译模型进行微调，我们采用了预处理步骤，使数据与模型兼容：

- **文本规范化**：转换为小写、删除数字和标点符号，并去除多余空格。
- **语言特定过滤**：为确保提交数据的一致性，阿拉伯语数据仅保留阿拉伯文，英语数据仅保留英语字母。
- **分词和填充**：使用 MarianMT分词器对句子进行分词，然后将序列截断至最多 128 个词元，以兼顾内存效率和文本覆盖率。

这些预处理步骤通过规范内容、保持语言一致性并减少训练过程中的错误，为数据集的微调做好了准备。

3.1.2 音频文件

共有1440个.wav格式的音频文件。如图1所示，标记为“angry”和“calm”的音频文件具有明显不同的频率和振幅特征。左侧声谱图展示了“愤怒”语音样本的对数频率功率谱，其特点是强度较高、频率变化范围较宽。右侧声谱图表示“平静”情感，其强度较低，频率过渡也更加平滑。这些差异凸显了情感语音独特的声学特征，尤其是能量在不同频率和时间上的分布。

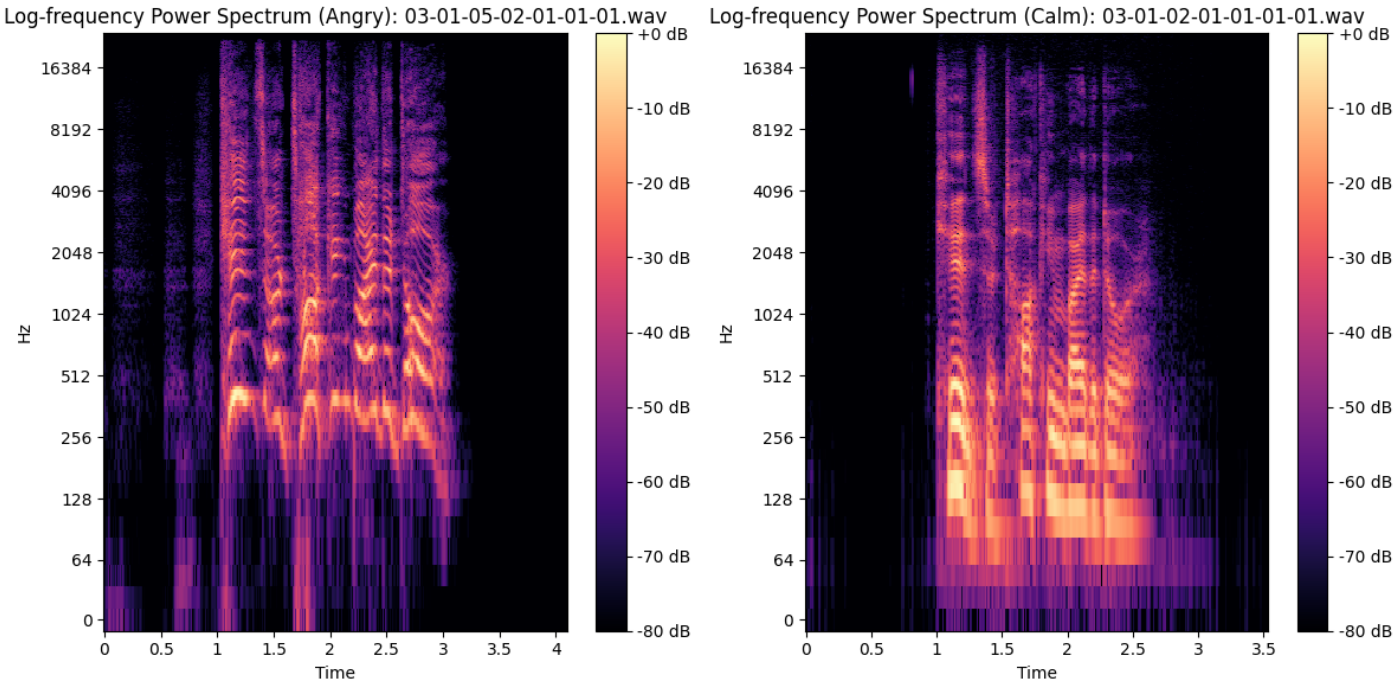

图1：不同情绪的对数频率功率谱

3.2 方法

我们构建了一个由四个组件组成的端到端系统，如图2所示：用于提取情感的**卷积神经网络** 模型。（我们实现并测试了长短期记忆网络和 *ResNet50*模型，以替代卷积神经网络，更多详情请参见A）。随后，使用经过预训练的**Whisper** [13] Transformer模型识别语音并将其转换为文本。然后，使用**MarianMT** [4] 预训练并经过微调的Transformer模型，将英语句子翻译成阿拉伯语句子。最后，使用 **MMS-TTS-Ara** [11] 预训练模型将文本转换为语音。它们在处理流水线中的具体作用如下：

3.2.1 用于语音情感识别的 CNN 模型：

该组件采用基于 TensorFlow/Keras 构建的人工神经网络模型，用于识别音频数据中的情感。实现如下：

数据准备： 为了获得更好的泛化能力，我们从（emotion-features.csv）中创建了数值特征，并对音频文件应用了**数据增强**技术，生成同一音频文件的多个变体，以模拟不同的声学条件。以下是所应用的技术：

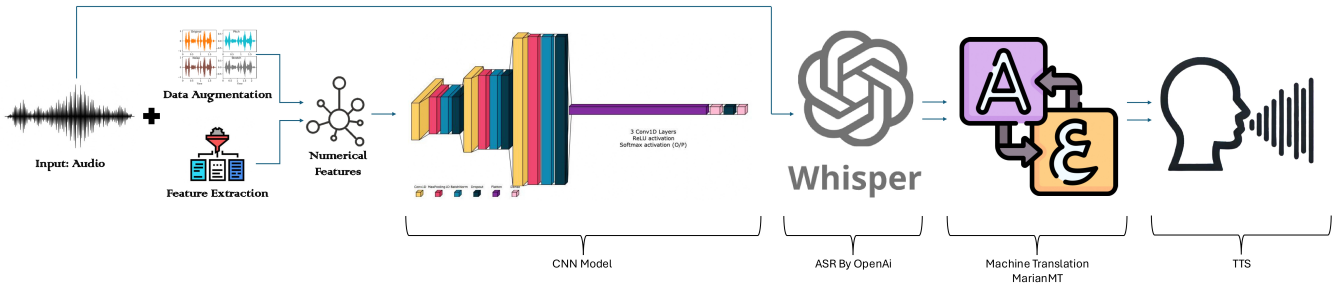

图2：系统架构

- **添加噪声：** 向音频信号中添加高斯噪声。

$$y_{\text{noisy}} = y + \text{noise_factor} \cdot \mathcal{N}(0, 1) \tag{1}$$

其中，y是原始信号，$\mathcal{N}$(0，1)是标准正态分布。

- **音高变换：**通过升高或降低频率来调整信号的音高。
- **时间拉伸：**加快或减慢音频信号。
- **时间偏移：**按信号长度的百分比移动信号。

数据增强后，使用**特征提取**将音频波形转换为数值表示，这是一种降维步骤，用于仅保留数据中的相关信息。使用了几种特征提取方法：

- **过零率 (ZCR)**通过 m 测量信号穿过零振幅线的频率，以捕捉信号的噪声程度。

$$\text{ZCR} = \frac{1}{T} \sum_{t=1}^{T} 1 \left[y_t \cdot y_{t-1} < 0 \right] \tag{2}$$

其中，T 为样本数量，1 为指示函数。

- **均方根能量（RMSE）**用于测量音频信号的能量。

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} y_t^2} \tag{3}$$

- **梅尔频率 倒谱系数（MFCC）**捕捉信号的频谱特性，并表示其音色 e.

$$\text{MFCC}[k] = \sum_{n=0}^{N-1} \log(X[n]) \cdot \cos\left(\frac{N\pi k}{N} (n - 0.5) \right) \tag{4}$$

其中，X[n] 为梅尔尺度功率谱。

特征提取 完成后，数据即可输入 CNN 模型。CNN 的构建如下 :

独热编码： 情绪标签（Y）属于类别型数据，使用独热编码器进行独热编码。这会将标签转换为适用于多分类的二进制向量。

特征缩放： 使用（标准化器）对所有特征进行归一化，将数据缩放至零均值和单位方差，以确保优化过程中的数值稳定性。

为 CNN 输入重塑形状： T 这些特征被重塑为 (样本数, 特征数, 1)，使其与一维卷积层兼容 rs.

CNN 模型由以下各层构成：

输入层： 接收形状为（特征，1）的输入数据，表示经过缩放的特征。

卷积层： 音频信号经过三个一维卷积层（带有滤波器），通过 ReLU 激活提取特征，并引入非线性，以解决梯度消失问题。

批归一化： 对每一层的输出进行归一化，以加快训练速度。

最大池化和丢弃层：每个 卷积层之后都紧接着设置最大池化层，以减少特征维度 ，以及用于卷积层和用于 d ense 层的丢弃，以防止过拟合。

全连接稠密层：卷积输出展平后，使用了包含 256 个神经元的全连接层和 ReLU 激活，并在对图像进行分类后以多分类方式完成 softmax 输出。

该模型使用 Adam 优化器进行编译，该优化器以训练深度神经网络时的高效性而著称。由于该问题属于多分类任务，因此采用分类交叉熵作为损失函数。总体而言，卷积神经网络能够有效捕捉序列数据中的模式。使用一维卷积层与音频特征的结构相匹配。

3.2.2 OpenAI 的 Whisper

Whisper 是 OpenAI [13] 开发的一款先进的自动语音识别（ASR）系统。其架构基于 Transformer 模型。主要组件：

编码器-解码器架构： Whisper 采用编码器-解码器方法，其中编码器负责处理原始音频，而解码器则根据编码特征生成转录文本。在这一序列到序列框架中，解码器接收语音信号并生成文本词元。

我们的项目使用 **Whisper Base** [14]。Whisper 的不同变体在速度与准确率之间各有权衡。总的来说，Whisper 非常稳健 t 能够应对口音、噪声和背景 干扰等变化，因此它是一个可靠的自动语音识别模型。

3.2.3 英语到阿拉伯语翻译

本节重点介绍如何通过使用微调后的 Marian 机器翻译模型，提升银行领域的英语到阿拉伯语翻译质量。首先，尝试使用多语言模型进行翻译。具体步骤包括评估预训练、准备训练数据集、微调，以及采用多种评估指标进行评估。使用了领域专用数据和先进的基于 Transformer 的架构。

数据集准备

该数据集包含一个定制创建的银行领域翻译数据集和一个英语到阿拉伯语通用翻译数据集。这些数据分别按80%、10%和10%的比例划分为训练、验证和测试子集。

初步模型与挑战

本工作的目标是开发一个高性能的翻译模型。由于数据集规模有限，训练定制构建的Transformer模型面临重大挑战。初步工作是在24,000个英阿句对上训练Transformer模型。该方法提供了基本的翻译能力，但受限于数据集规模，不足以实现有效泛化——尤其是对于英语和阿拉伯语这类复杂语言对。构建高性能Transformer模型通常需要数百万句对才能达到预期效果。为应对这些限制，我们对预训练模型MarianMT进行了微调，以在现有小规模数据集上优化翻译性能。

采用Marian机器翻译模型

Marian 机器翻译是由Helsinki-NLP [6], 提供的预训练翻译模型，它基于高效的Transformer序列模型。也可以将其理解为由编码器、解码器、多头自注意力机制、层归一化以及逐位置前馈网络组成。编码器将源语言文本的上下文和句法结构捕获为一系列连续表示。然后，编码器将源文本转换为若干连续表示。解码器随后通过关注这些表示并建模目标序列中的依赖关系来生成目标文本。这强调了预训练模型和额外数据的重要性，从而使MarianMT成为合适的选择。

超参数实验

为了优化 Marian 机器翻译模型的训练，我们针对一些关键超参数进行了实验，例如学习率、批大小、训练轮数、束搜索宽度和梯度累积步数，并测试了多种配置，以找出能够获得最佳翻译质量的超参数值。

数学公式

模型通过最小化预测词元 $\mathcal{L}\hat{y}$ 与参考词元 y 之间的交叉熵损失来进行训练，其定义如下：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log P(\hat{y}_i | y_{<i}, x)$$

其中， x 是输入序列， $y_{<i}$ 是前序词元， N 是词元数量。

MarianMT 的高效性以及其在不同语言数据集上的预训练，为其提供了强大的基线。此外，Transformer 架构对长程依赖关系的建模能力，对于执行英语-阿拉伯语等复杂翻译至关重要。

3.2.4 使用 MMS-TTS-Ara 进行文本转语音

如图 2 所示，在整个流程的最后阶段采用预训练的 MMS-TTS-Ara 模型，将翻译后的阿拉伯语文本转换为自然流畅的语音。该模型集成了**文本编码器**、**序列生成器**和**神经声码器**，以生成高质量的音频输出。

- **文本编码器：** 使用基于 Transformer 的架构，将阿拉伯语文本转换为潜在表示，同时捕捉语音和句法细节。
- **序列生成器：** 将这些表示映射为梅尔频谱图，对语音的时频特征进行编码。
- **神经声码器：** 利用 HiFi-GAN 等先进模型将梅尔频谱图转换为波形音频，生成语调清晰、如同人类的语音。

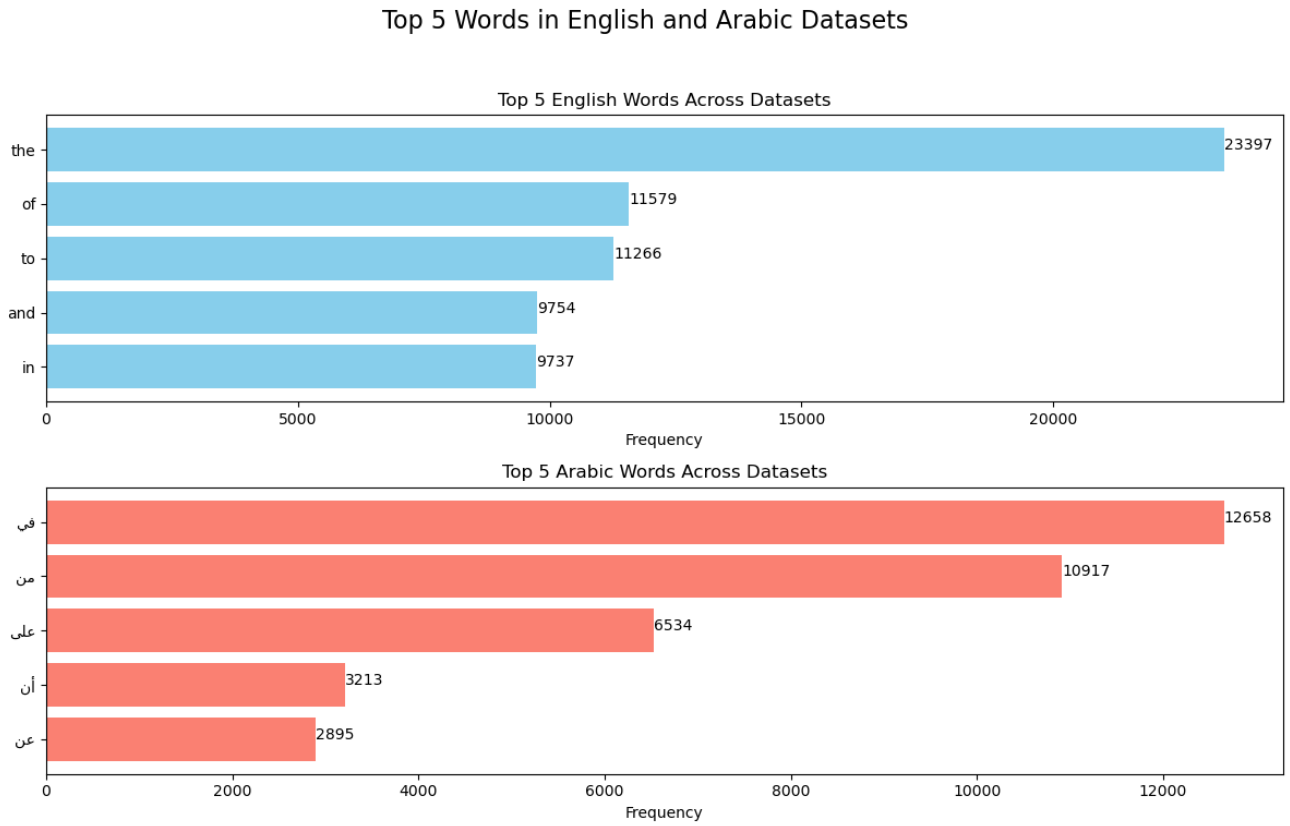

图3：英语和阿拉伯语数据集中的前5个高频词

该过程包括文本规范化、分词以及韵律调整，确保保留上游组件中的情感语境。MMS-TTS-Ara 针对阿拉伯语进行了优化，可生成流畅、自然的输出，使其非常适合银行客户交互。此步骤通过将文本无缝转换为语音，完成了多语言通信流水线。

4 实验设置

4.1 MarianMT微调

为了将Marian MT模型融入银行业领域，我们实施了一套系统化的微调流程，旨在提升翻译质量，同时最大化训练效率。用于微调的翻译数据集中功能词出现频率很高，例如英语中的“the”“of”和“to”，以及阿拉伯语中的类似词语，如图3所示。这些高频使用的词语对于帮助模型理解任何语言最基本的结构规则及其语法至关重要，从而使翻译更加准确、流畅。就句子复杂度而言，阿拉伯语和英语银行数据集的平均句长分别为10个词和11个词，而通用英阿翻译数据集的平均句长更长，阿拉伯语为15个词，英语为17个词。这种句长范围可能增强了模型的灵活性，使其能够高效处理短句和复杂句。

4.2 数据集

4.2.1 RAVDESS数据集

具体而言，表1展示了RAVDESS [9] 数据集，该数据集将用于训练我们的语音情感识别（SER）模型。每个文件名都包含标识符，格式遵循RAVDESS官方网站的说明：(XX-XX-XX-XX-XX-XX-XX.wav)

模态（ 01 = 02 = 03 = 01 = 02 = 01 = 02 = 03 = 04 全视听、仅视频、纯音频）。声道（语音、歌曲）。情感（中性、平静、快乐、悲伤、愤怒、恐惧、厌恶、惊讶）。情感强度（正常、强烈）。注意：中性情感没有强烈强度。语句（01 = “孩子们正在门边说话”、 02 = “狗正坐在门边”）。重复（01 = 1次重复、 02 = 2次

重复）。演员（01 至 24。编号为奇数的演员为男性，编号为偶数的演员为女性）。数据集被划分为训练集、验证集和测试集。划分比例为：测试集 7.5%、验证集 7.5%、训练集 85%。

数据集	**RAVDESS**
模态	音频、视频（全音视频）、纯音频
语言	英语（北美口音）
样本数量	1440个音频文件
参与者人数	24名演员（12名男性，12名女性）
情感类别数	8

表1：RAVDESS数据集的特征

4.2.2 阿拉伯语到英语翻译数据集

在翻译部分，为了有效地微调 MarianMT 模型，我们使用了两个数据集。第一个是来自 Kaggle 的英语到阿拉伯语翻译句子 [12] 数据集，内容涵盖现代标准阿拉伯语中的一般主题。该数据集大小为 6.4 MB，包含超过 24000 个样本，展示了英语文本及其阿拉伯语译文。

4.2.3 Banking 77 数据集

对于数据集的第二部分，我们根据 Hugging Face 数据集仓库中的 **Banking 77** [3] 数据集创建了一个新数据集。原始数据集为英语，因此我们使用 deep translator 中的 Google 翻译和 MyMemoryTranslator 模型将句子翻译成阿拉伯语，随后验证了译文的准确性。该数据集包含 10000 条英语句子及其对应的阿拉伯语译文，大小为 773 KB。

4.3 超参数

为训练 CNN 模型测试了不同的超参数。此外，还针对翻译模型测试了不同的超参数配置，以确定微调的最优设置。以下是所采用的主要方法：

- **高质量：**注重准确性与流畅度，采用较高的束搜索宽度、较低的学习率以及更长的训练轮数。
- **均衡：**采用适中的学习率、较少的训练轮数以及较小的束搜索，以便在合理时间内完成有效训练。
- **快速训练：**采用较高的学习率、较少的训练轮数以及较小的批次大小，以牺牲质量为代价快速完成训练。

在测试这些配置后，微调的最优超参数详见表 2“微调 MarianMT 模型超参数”一节。

4.4 评估指标

4.4.1 情感分类

为了评估语音情感识别 （SER）组件，可以使用多种评估指标，例如：精确率、召回率、F1分数和支持度 rt.

4.4.2 翻译

为了评估翻译质量，我们综合使用了 BLEU分数（通过衡量模型输出与参考文本之间的 n 元语法来评估翻译质量）和 BERTScore（比较上下文嵌入，以评估含义和流畅度）。由于自动评价指标在给出准确分数方面存在局限性，我们对银行领域翻译数据集中的 100 个测试样本进行了人工评估。每条翻译的评分范围为 1 至 3 分，以 0.5 分为增量，并根据准确性、流畅度和领域专用术语进行评定，如表 3 所示：

人工评估能更深入地了解银行业领域翻译的质量，弥补自动评价指标未能捕捉到的细微差异，并提升整体评估效果。

4.5 基线比较

我们建立了基线模型，以评估通过微调和额外处理所实现的性能提升。这些基线为评估所开发系统的有效性提供了参考：

- **卷积神经网络** **模型：** 创建了一个基线卷积神经网络模型，未进行微调、数据增强或特征提取。

超参数	值
卷积神经网络模型超参数	
学习率	1×10^{-3}
训练轮数	100
批大小	64
卷积核大小	3
丢弃率	0.15（卷积层），0.25（全连接层）
滤波器数量	64, 128, 256
池化大小	2
优化器	Adam
激活函数	ReLU（隐藏层），Softmax（输出层）
微调 MarianMT 模型超参数	
批大小	单设备批大小为 8，梯度累积步数为 4，因此有效批大小为 32。
学习率	3×10^{-5} 使用 1000 个预热步数，以确保训练稳定。
训练轮数	经过 10 轮次微调，以实现最佳收敛效果。
生成参数	最大序列长度为 300 个词元，束搜索宽度为 8。
权重衰减	0.01，用于正则化。
评估策略	在每轮次结束时执行。
混合精度	启用 FP16，以加快训练并减少内存使用。
梯度累积	累积 4 步以提高内存效率。
保存策略	在每个轮次结束时保存最佳模型，最多保留 2 个检查点。
优化	使用带权重衰减的 Adam 优化器以保持稳定性。
验证批次大小	每设备评估批次大小为 8。

表 2：卷积神经网络和微调后的 MarianMT 模型的超参数。

得分	描述
3	准确度极高、流畅且符合领域特点，堪比人工翻译。
2.5	总体准确，仅有少量措辞或术语不一致。
2	可以接受，传达了主要含义，但存在一些小的错误。
1.5	准确性或流畅度方面存在明显问题，造成轻微困惑。
1	存在重大问题，导致翻译难以理解。

表 3：翻译质量评估标准

• **MarianMT 模型：** 对于翻译任务，使用预训练状态下的 MarianMT 模型作为基线。将其性能与微调后的版本进行比较，以评估翻译质量的提升。

这些基线　为评估模型开发过程中引入的改进所产生的影响奠定基础　t.

4.6 计算资源

该项目使用 RTX 3060 GPU、Ryzen 3700X CPU 和 64GB 内存运行，同时也将使用 Google Colab 提供的 A100 和 T4 等 GPU 作为计算资源支持。

5 结果与讨论

5.1 CNN 模型

上文表 4 展示了实现的 CNN 模型和基线卷积神经网络在 F1分数上的表现。结果表明，实现的 CNN 模型在所有情绪类别上的表现均优于基线。两种模型中，平静、恐惧和悲伤情绪类别的 F1分数最低，这表明两种模型在对这些情绪进行分类时都遇到了一定困难，可能是因为它们的声学特征存在重叠。另一方面，在所有不同的情绪类别中，愤怒情绪类别在两种模型中均取得了最高的 F1分数，这可能是由于此前图 1 所示的愤怒情绪类别具有独特且强度较高的功率谱。更重要的是，结果证明，特征提取、微调和数据增强对于开发语音情感识别系统至关重要。基线卷积神经网络未使用这些特征，因此难以对不同的情绪状态进行分类，在所有评估指标上的表现都较差。然而，实现的

情绪类别	基线卷积神经网络（F1分数）	实现的卷积神经网络（F1分数）
愤怒	0.66	0.97
Calm	0.49	0.92
厌恶	0.52	0.96
恐惧	0.48	0.92
快乐	0.56	0.94
中性	0.52	0.93
Sad	0.50	0.92
惊讶	0.53	0.95
平均F1分数	**0.53**	**0.94**

表4：基线模型和实现的卷积神经网络模型在各情绪类别上的F1分数

CNN 能够取得良好的 F1分数，因此具备实际应用价值。这表明，在开发最先进模型时，利用特定特征和稳健的训练技术至关重要。测试了 *CNN* 以外的不同模型，更多详情请参见A。

5.2 翻译

从头训练 Transformer模型的 BLEU分数为 23.00，这凸显了我们使用小型数据集时可能面临的挑战；使用预训练的 Marian 机器翻译模型后，该分数略有提升，达到 25.48。在领域专用数据集和通用数据集上进行微调，带来了显著提升——如表5所示，最佳模型的 BLEU分数达到 56，BERTScore F1分数达到 88.7%。尽管取得了这些进步，模型仍可能难以处理很长的句子：由于训练数据集不包含超过 20 个词的句子，模型可能会截断翻译或误解上下文，这限制了它捕捉长序列依赖关系的能力。模型处理较短句子的效果更好。整合领域专用数据集和通用数据集，使模型能够学习银行业特定术语，同时保持更广泛的语言能力，从而能够对主要句子结构和上下文进行泛化，并取得更高的 BLEU 和 BERTScore 评估指标。第 4.4.2 节讨论的人工评估也支持这些发现：按照准确性、流畅度和领域专用术语这几项标准，对银行业领域的 100 个测试样本进行评分（标准详见第 3 节），模型的人工评估平均得分为 81%，与自动评价指标高度一致，证实了其执行领域专用翻译任务的能力。

模型	BLEU分数	BLEU置信区间	BERTScore F1分数
从头训练的Transformer	23.78	(22.94–24.5)	67%
预训练Marian机器翻译模型	25.48	(23.98–27.09)	73%
微调的 MarianMT 模型	56.00	(54.5–57.5)	88.7%

表5：BLEU 和 BERTScore 结果及置信区间

结果凸显了微调在领域专用翻译中的作用：BLEU、BERTScore 和人工评估指标均有所提升。结合领域专用数据集与通用数据集，显著提升了银行业等专业领域的翻译质量。

6 局限性

微调后的Marian机器翻译模型提升了翻译质量，但受数据集规模限制。增加银行领域数据集的样本数量，或可进一步提升准确性和流畅度。未来工作可以探索通过回译或使用 T5 等模型进行释义改写来扩充数据，并采用适配器层或 LoRA 等先进的微调技术。此外，我们可以抽取部分译文，请人工评估其质量，从而获得准确的评估结果。

7 结论

本项目成功展示了在银行客户服务环境中整合情感驱动的语音转录与跨语言翻译。通过采用多个模型，包括用于语音情感识别的卷积神经网络、用于自动语音识别的 Whisper、用于翻译的 MarianMT，以及用于文本转语音合成的 MMS-TTS-Ara，我们开发了一套能够在不同语言间保持情感完整性的端到端系统。该系统在真实银行业场景中的实际应用，体现了它通过解决语言障碍和情感细微差异来提升客户体验的能力。未来的工作将着重于扩充数据集并探索先进的微调技术，以进一步提升翻译的准确性和流畅度。

A 补充结果

B 长短期记忆网络 &ResNet50

情绪类别	长短期记忆网络（F1分数）	ResNet50（F1分数）
愤怒	0.85	0.90
Calm	0.83	0.86
厌恶	0.86	0.91
恐惧	0.84	0.88
快乐	0.88	0.92
中性	0.83	0.87
Sad	0.82	0.85
惊讶	0.89	0.93
平均F1分数	**0.87**	**0.89**

表 6：长短期记忆网络和 ResNet50 模型在各情绪类别上的 F1分数

如表 6 所示，长短期记忆网络和 ResNet50 是另外两个模型，已针对所有不同的情绪类别进行实现和测试。这两个模型的平均 F1分数几乎相同。不过，ResNet50 在所有情绪类别上的得分都更高，这可能是因为其深度残差架构能够从音频信号中提取稳健的特征。正如第 5.1 节中对悲伤、恐惧和平静情绪类别的观察结果所示，长短期记忆网络和 ResNet50 模型在区分不同情绪类别时都面临相同的困难。

参考文献

[1] A. Baevski、H. Zhou、A. Mohamed 和 M. Auli。wav2vec 2.0：一种用于语音表征自监督学习的框架。载于第*34*届神经信息处理系统大会（*NeurIPS*）论文集，2020年。[2] Google Cloud。Cloud 文本转语音文档。https://cloud.google.com/text-to-speech，2021年。[3] Hugging Face Datasets。Banking 77 数据集。https://huggingface.co/datasets/banking77，2020年。通过 Hugging Face Datasets 库访问。[4] Hugging Face。Marian：多语言机器翻译模型。https://huggingface.co/docs/transformers/en/model_doc/marian，2021年。访问日期：2024年11月14日。[5] A. Hannun、A. Maas、D. Jurafsky 和 A. Ng。DeepSpeech：扩展端到端语音识别。*arXiv* 预印本 *arXiv:1412.5567*，2014年。[6] Helsinki-NLP。MarianMT：开放式神经机器翻译模型，2020年。访问日期：2024年11月14日。[7] Y. Huang、X. Yu、J. Ou 和 W. Xu。使用迁移学习进行语音情感识别。载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）论文集，2021年。[8] M. Junczys-Dowmunt、R. Grundkiewicz、K. Heafield 和 A. Birch。Marian：快速神经机器翻译，采用 c++。载于*ACL 2018*系统演示论文集，2018年。[9] UWRF Kaggle。RAVDESS情绪语音音频。https://www.kaggle.com/datasets/uwrfkaggler/ ravdess-emotional-speech-audio，2023年。访问日期：2024年9月16日。[10] S. R. Livingstone 和 F. A. Russo。Ryerson 情感语音与歌曲视听数据库（RAVDESS）。*PLOSONE*，13 0196391，2018 5( ):e 。[11] MMS-TTS-Ara。facebook/mms-tts-ara - 阿拉伯语文本转语音模型。https://huggingface.co/facebook/mms-tts-ara，2024年。https:// huggingface.co/facebook/mms-tts-ara。[12] Samir Moustafa。阿拉伯语到英语翻译句子数据集，2024年。Kaggle。[13] OpenAI。Whisper：鲁棒语音识别模型。https://openai.com/index/whisper/，2022年。访问日期：2024年11月14日。[14] OpenAI。Whisper：鲁棒语音识别模型。https://github.com/openai/whisper，2022年。访问日期：2024年11月14日。