多特征融合的越英端到端语音翻译方法

马候丽[1,2]，董凌[1,2]，王文君[1,2]，王剑*[1,2]，高盛祥[1,2]，余正涛[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1341584939@qq.com,46761956@qq.com,175360805@qq.com,
1528906057@qq.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

摘要

语音翻译的编码器需要同时编码语音中的声学和语义信息，单一的Fbank或Wav2vec2语音特征表征能力存在不足。本文通过分析人工的Fbank特征与自监督的Wav2vec2特征间的差异性，提出基于交叉注意力机制的声学特征融合方法，并探究了不同的自监督特征和融合方式，加强模型对语音中声学和语义信息的学习。结合越南语语音特点，以Fbank特征为主、Pitch特征为辅混合编码Fbank表征，构建多特征融合的越-英语音翻译模型。实验表明，使用多特征的语音翻译模型相比单特征翻译效果更优，与简单的特征拼接方法相比更有效，所提的多特征融合方法在越-英语音翻译任务上提升了1.97个BLEU值。

关键词： 端到端语音翻译；特征融合；越南语；语音表征；音高特征

A Vietnamese-English end-to-end speech translation method based on multi-feature fusion

Houli Ma[1,2],Ling Dong[1,2],Wenjun Wang[1,2],Jian Wang[1,2], Shengxiang Gao[1,2],Zhengtao Yu[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
1341584939@qq.com,46761956@qq.com,175360805@qq.com,
1528906057@qq.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

Abstract

Speech translation encoder requires to represent both acoustic and semantic information, a single speech feature whether Fbank or Wav2vec2 feature is insufficient. Based on the difference between hand-crafted Fbank feature and self-supervised wav2vec2 feature, this paper proposes a representation fusion method by cross-attention mechanism, and explores different self-supervised features and fusion method. Multiple features complement each other, strengthen the modelling ability of acoustic and semantic information in speech. In addition, combined with Vietnamese language features, the Fbank feature is used as the main, and the Pitch feature is used as the supplementary, hybrid encoding representation to construct a Vietnamese-English speech translation model. Experiments show that the proposed framework outperforms baselines with

*王剑（通信作者）：1528906057@qq.com
基金项目：国家自然科学基金（61732005, U21B2027, 61972186）；云南高新技术产业发展项目（201606）；云南省重大科技专项计划（202103AA080015, 202002AD080001-5）；云南省基础研究计划（202001AS070014）；云南省学术和技术带头人后备人才（202105AC160018）

gains of up to 1.97 BLEU, the speech translation effect of using multi-feature is better than that of single-feature, and the proposed multi-feature fusion method is more effective for performance improvement than the simple feature concatenation.

Keywords: end-to-end speech translation , feature fusion , Vietnamese , speech representation , pitch feature

1 引言

语音到文本的翻译旨在将源语言的语音翻译为目标语言的文本 (Stentiford and Steer, 1990)，广泛应用于会议演讲、跨境旅游、同声传译等各个领域。然而构建语音翻译数据集对标注者和成本要求较高，目前仅有少数语言对有公共语料库，如英语到德语、法语等语言以及少数欧洲语言 (Di Gangi et al., 2019; Iranzo-Sánchez et al., 2020)。大多数语言对缺少语音翻译标注语料，如越南语-英语尚无公开的语音翻译数据集，研究工作相对匮乏，迫切需要开展相关研究。

端到端语音翻译使用一个模型直接将源语言语音映射到目标语言文本，避免了级联方式固有的错误累积、高延迟等缺陷 (Bérard et al., 2016)，因此备受研究者关注。端到端语音翻译模型同时进行跨模态跨语言的映射，且训练数据较稀缺，翻译性能与级联模型仍存在较大差距。此外，语音数据受说话人情绪、音量、口音和外界噪声等因素产生多变性 (Han et al., 2021; Liu et al., 2020)，限制了端到端语音翻译模型的性能。而特征提取是语音翻译的重要步骤,特征的好坏直接影响翻译的效果。因此，探索有效的语音表征对于语音翻译任务至关重要。

语音翻译中常使用人工设计的Fbank特征或基于自监督的Wav2vec2特征 (Baevski et al., 2020)作为模型输入。如图 1左侧为Fbank特征的提取过程，在一次分帧的基础上，进行逐帧变换。在采样率为16K、帧长为25ms、帧移为10ms的设置下，每帧Fbank特征覆盖400个采样点，能够表示语音声学信息中的局部特征，但Fbank特征提取方法根据人声预先设置，不可动态学习，具有一定的局限性。图 1右侧为Wav2vec2特征提取过程，Wav2vec2模型通过堆叠的7层不同步长和卷积核大小的卷积神经网络，进行逐层循序计算提取特征。每帧Wav2vec2特征覆盖3240个音频采样点，堆叠的CNN增大了语音声学信息的覆盖范围，有利于对语音中语义信息进行表征和学习。但由于Wav2vec2表征在大规模无标注语音上进行自监督预训练，表征在目标任务上的表现性能高度依赖训练域和目标域的相关性 (Berrebbi et al., 2022)。单一的Fbank或Wav2vec2特征作为模型的输入时，不能满足模型同时对语音中声学信息和语义信息建模的需求。

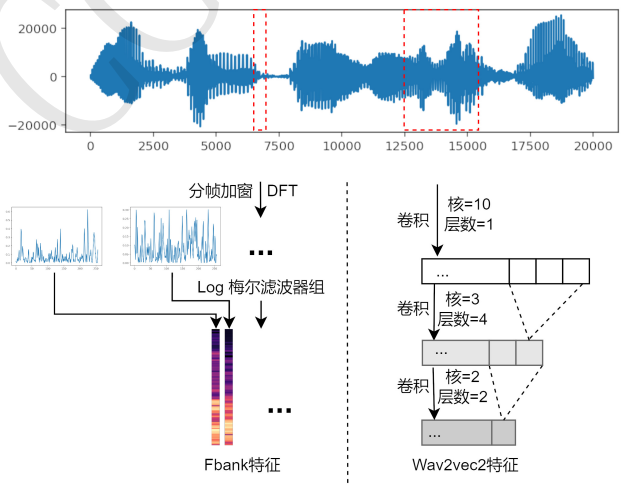


图 1. Fbank和Wav2vec特征提取过程比较

越南语属音调语言，以单音节为主，每个音节的元音带有6种音调之一。音调不同则词义不同，音调错误会产生句子歧义 (Nga et al., 2021)，音调信息通常使用音高特征（Pitch）

表示。对于越南语而言，音调信息在一定程度上有助于表征语音中的声学信息和语义信息。 Huy Nguyen (2019)在越南语语音识别的实验上也表明，融合音调信息的语音特征能够提高识别的准确性。因此，音调信息对于构建越南语-英语语音翻译模型至关重要。

为此，本文对Fbank以及预训练Wav2vec2特征进行实验，分析其在越南语-英语语音翻译和越南语语音识别任务上的表现性能。实验表明不同特征表现效果及其可视化结果均存在差异性。基于两种特征间的差异，提出通过对Wav2vec2特征和Fbank特征分别进行编码后，使用交叉注意力机制进行特征融合的语音翻译模型，帮助编码器同时编码语音中的声学信息和语义信息，弥补单一特征表征能力不足的缺陷。并使用混合Pitch的编码块和Fbank编码块交替编码的方式，在Fbank特征编码的过程中显式的加入越南语音调信息，提高编码器对于音调信息的敏感性，进一步辅助编码器对语义信息的建模。此外，本文还探究了不同的特征融合方式对于语音翻译性能的影响, 为语音翻译任务选择最优融合方式。

本文贡献如下：（1）探究和分析了Fbank特征和Wav2vec2特征在越南语语音识别和越-英语音翻译任务上的表现。（2）基于Wav2vec2特征和Fbank特征间的差异，提出Wav2vec2和Fbank特征相融合的方法进行不同特征间的相互补充，加强编码端的表征能力，并比较了不同融合方法和特征对语音翻译效果的影响。（3）结合越南语语言特点，使用不同编码块交替编码的方式在Fbank特征中加入越南语音高特征，增强语音翻译模型对声学信息和语义信息的编码能力。（4）提出多特征融合的语音翻译框架，将Fbank、Wav2vec2、Pitch三种特征进行有效融合构建编码端语音表征，提升了越南语-英语语音翻译质量。

2 相关工作

2016年， Bérard et al. (2016)提出端到端的语音到文本的翻译的设想。随着深度学习的发展，端到端的语音翻译逐渐得以实现， Duong et al. (2016)建立直接的语音翻译模型，在不使用任何源语言文本前提下，使用单个模型建立源语言语音到目标语言文本的映射。但端到端的语音翻译模型面临着严重的资源不足问题，为此研究者们分别从数据层面和方法层面来缓解资源不足和映射困难。数据层面主要从音频数据增强和数据合成两方面弥补资源不足问题，其中语音增强是语音类任务普遍使用的方法，包含SpecAugment对语音的语谱图在时域和频域进行掩蔽防止模型过拟合 (Park et al., 2019)，对原始语音进行变速等操作。数据合成分别通过为文本翻译语料生成音频数据、为语音识别语料生成目标本文数据，以及为语音翻译语料生成多说话人音频数据等方式 (Post et al., 2013)。但合成大规模语料需耗费大量时间和成本，同时由于机器合成数据分布单一，需要设计合适的比例将合成数据与真实数据混合，且对模型性能的提升仍非常有限。方法层面的探索主要通过不同方法引入额外数据、训练子模块来提高语音翻译的性能。 Anastasopoulos and Chiang (2018)和 Liu et al. (2020) 采用多任务和预训练-微调的方法，通过整合额外的ASR和MT数据来提升语音翻译的性能。此外，研究者们还在课程学习、元学习、知识蒸馏等深度学习方法上进行探索 (Kano et al., 2017; Liu et al., 2019; Indurthi et al., 2020)。这些方法借鉴了相关领域的方法，都能在一定程度上提高语音翻译的性能。越南语-英语的语音翻译语料和文本翻译的训练语料有限，但现有多任务和预训练的方法不能直接迁移，且使用单一的语音特征，对语音中的声学和语义信息的表征存在一定局限性。因此本文提出多特征融合的越-英语音翻译方法，旨在通过融合自监督语音表征，结合越南语语音的音高特征，对传统Fbank特征进行补充，在不使用额外数据以及额外训练步骤的情况下，最大程度提高越英语音翻译的性能，降低训练成本的同时满足低资源语言语音翻译的需求。

语音特征表示的好坏会直接影响语音翻译的效果，语音信号的不确定性以及噪声增加了语音特征提取的难度。传统的语音特征提取采用信号处理的方法对语音信号进行频域分析，目前使用较多的是Fbank和MFCC两种特征。MFCC特征由于DCT变换造成一定程度的语音信息丢失，因此基于深度神经网络的语音识别和语音翻译模型中更多使用的是Fbank特征 (Mohamed, 2014) 。但由于实际的语音数据以及场景的复杂性，人工设计方式提取的语音特征在一定程度具有局限性。近几年在语音识别和说话人识别领域出现使用基于深度学习的方式提取语音特征 (Tüske et al., 2014)。由于神经网络可以从原始波形中提取出更合适的语音表征，研究者们基于CNN构建声学模型，直接使用原始波形作为输入 (Ravanelli and Bengio, 2018)。由于原始波形是长序列数据，训练时对内存资源和硬件要求高，使得训练过程更加具有挑战性。 Baevski et al. (2020)采用自监督的方法在大量无标注的原始音频上学习Wav2vec2语音表征，通过多层卷积神经网络和Transformer模型提取语音表征，使用该表征在下游语音识别任务

中经少量标注数据上进行微调，在Librispeech的测试集上实现了4.2%的词错率。

与以往使用单一语音特征的语音翻译模型相比，本文根据Fbank特征和Wav2vec2特征的差异性，探索两种特征的融合方法，并结合越南语语音特点有效融入语音音高特征，提升越南语-英语语音翻译模型效果。

3 方法

本文提出一种有效融合多特征的越-英语音翻译模型，模型由编码器和解码器组成。对越南语音频分别提取Fbank特征、Wav2vec2特征和Pitch特征，作为编码器输入。编码器由交替特征编码层、Wav2vec2编码层和表征融合层组成，使用交替特征编码层对Fbank和Pitch两种频谱特征进行混合编码输出频谱表征，通过加入越南语的音调信息增强模型对语义信息的表征能力；同时使用Wav2vec2编码层对Wav2vec2特征进行编码输出自监督表征；将频谱表征和自监督表征输入到表征融合层，通过交叉注意力机制学习不同类型表征间的对齐和融合，得到最终的编码器输出表征。最终，将编码器输出表征输入解码器，输出目标语言文本词序列，具体过程如下所述。

3.1 越南语音频多特征提取

音频特征提取是语音翻译模型构建的重要基础环节，模型需要根据输入的语音特征同时对声学信息和语义信息建模，使用单一特征同时对两类信息进行建模存在较大挑战。本文对音频序列分别提取Fbank特征、Pitch特征以及Wav2vec2三种特征，作为模型的输入。其中，Fbank特征和Pitch特征为人工特征，Wav2vec2特征为自监督方法所提取的特征。训练语料为$S = \{(x, y)\}$，其中$x = (x_1, \ldots, x_m)$为音频序列，$y = (y_1, ..., y_n)$为目标语言文本序列，m和n分别为源音频序列和目标文本序列的长度。特征提取过程如式 (1)所示：

$$i = \text{Extractor}_i(x) \in \mathbb{R}^{d_i}, i \in \{FilterBank, Pitch, Wav2vec2\}, \tag{1}$$

Fbank特征：使用torchaudio包[1]，设置帧移为10ms，帧窗口大小为25ms，提取80维的Fbank特征序列为$f = (f_1, f_2, ...f_{l_f})$，其中$f \in \mathbb{R}^{d_f}$，$d_f$为Fbank特征维度，$l_f$为序列长度；

Pitch特征：使用pySPTK工具[2]中的SWIPE算法进行提取，搜索频率范围设置为50Hz至400Hz，提取的Pitch特征序列为$p = (p_1, p_2, ...p_{l_p})$，其中$p \in \mathbb{R}^{d_p}$，$d_p$为Pitch特征维度，$l_p$为序列长度；

Wav2vec2特征：开源的w2v2-vi模型[3] 在100小时的越南语有声读物进行预训练，使用该模型的第7层CNN输出的512维向量进行实验，特征序列为$w = (w_1, w_2, ...w_{l_w})$，其中$w \in \mathbb{R}^{d_w}$，$d_w$为Wav2vec2特征维度$d_w$，$l_w$为序列长度。

3.2 多特征融合编码器

与以往对单一特征进行编码的语音翻译模型不同，本文提出在编码端对Fbank、Wav2vec2以及Pitch三种语音特征进行编码，模型编码器由Wav2vec2特征编码层、Fbank-Pitch交替特征编码层和表征融合层三部分组成。其中，利用Fbank-Pitch交替特征编码层显式加入Pitch特征，进一步辅助编码器对语义信息的建模。多特征融合编码器如图 2所示。

Fbank-Pitch交替特征编码层：Fbank特征序列f经下采样$D(\cdot)$后，叠加位置编码pos_f，与下采样后的Pitch特征序列$D(p)$共同作为交替特征编码层的输入，编码输出隐层状态序列h_1，如式 (2)，下文简称该序列为频谱表征；

$$h_1 = \text{AlternatedEncoder}(D(f) + pos_f, D(p)) \tag{2}$$

Wav2vec2特征编码层：对于Wav2vec2特征序列w，使用CNN作为编码器，并通过维度转换，得到隐层状态序列h_2，下文简称自监督表征，编码过程如式 (3)所示；

$$h_2 = \text{Wav2vec2Encoder}(w) \tag{3}$$

[1] https://pytorch.org/audio

[2] https://github.com/r9y9/pysptk

[3] https://huggingface.co/dragonSwing/wav2vec2-base-pretrain-vietnamese

表征融合层：本模块输入为频谱表征h_1和自监督表征h_2，使用交叉注意力机制进行表征间的融合和对齐，输出融合表征向量。在不增加表征长度和特征维度的情形下，通过交叉注意力机制自动学习两种表征间的对齐，进行相互补充和增强。特征融合过程如式 (4)所示。

$$h_x^A = \text{FusionLayer}(h_1, h_2) \tag{4}$$

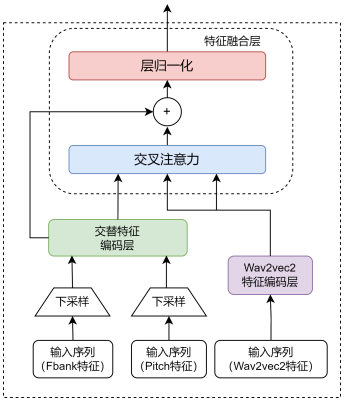
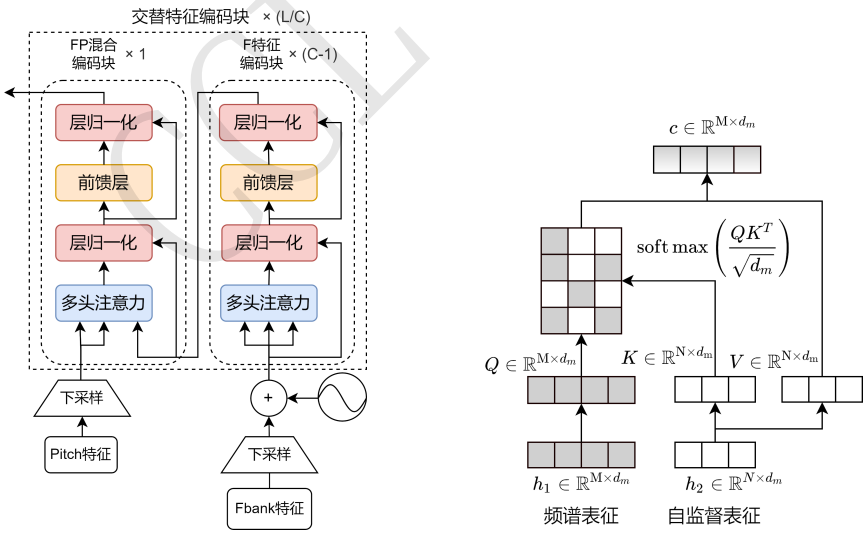


图 2. 多特征融合的编码器

3.2.1 Fbank-Pitch交替特征编码层

越南语中，音调可用于区分词义，语音中的音调信息通过音高特征（Pitch）表示。在语音特征中显式加入音调信息可增强模型对于语义信息的建模能力。不同于以往使用单一的Fbank特征作为Transformer编码块输入的工作，本文根据越南语的语音及音调特点，以Fbank特征为主，Pitch特征为辅，使用两种编码块交替编码的方式进行特征编码。两种编码块包括以Fbank作为输入，基于自注意力的Transformer编码块，和以Fbank和Pitch作为输入，基于交叉注意力的Transformer编码块，下文简称F特征编码块（F-Block）和FP混合编码块（FP-Block）。交替编码的方式在不增加编码块个数和模型复杂度的基础上，融合Pitch信息进行更有效的编码。



(a) Fbank-Pitch交替特征编码层　　(b) 特征融合层中的交叉注意力机制

图 3. 多特征融合编码器子模块

如图 3（a）所示，交替特征编码层共包含L个编码块，交替周期为C，则包含(L/C)个交替周期。每个交替周期内有$(C-1)$个F-Block和1个FP-Block, 其中i为当前编码块的块数，编码块的设置如式 (5)。F-Block专注对Fbank特征进行编码，而FP-Block采用交叉注意力机制同时

对Pitch特征和Fbank特征进行混合编码。F-Block和FP-Block均对Fbank特征进行编码，间隔多个块对Pitch特征进行编码，该设计与实际发音相符，即区分词义和句义主要通过不同音素的发音，辅以不同的音高。本文设置$C=3$，$L=12$，具体说明见4.1.3。

$$Block_i = \begin{cases} F-Block, & i\%C \neq 0 \\ FP-Block, & i\%C = 0 \end{cases} \tag{5}$$

3.2.2 基于多头交叉注意力的表征融合层

Fbank特征根据人耳对声学信号的感知，手工设计结构来提取特征，对复杂的音频的特征提取具有局限性；基于自监督-预训练方式得到的自监督表征缺乏对具体任务和数据的适应性。为更好的对音频进行表征，在编码器中，将带有音调信息的频谱表征和自监督表征使用交叉注意力机制进行融合，使得不同类型特征间相互补充，满足语音翻译任务需要同时对声学信息和语义信息建模的要求。

将Fbank-Pitch交替特征编码层输出的频谱表征h_1，和Wav2Vec2特征编码层输出的自监督表征h_2，通过多头交叉注意力机制进行特征融合。多头交叉注意力计算过程如图 3（b）所示，将频谱表征h_1作为q，自监督表征h_2作为k和v，首先通过式 (6)的线性变换分别得到向量Q,K,V，其中W_i^Q, W_i^K, W_i^V均为随机初始化的参数矩阵；

$$Q = qW_i^Q, K = kW_i^K, V = vW_i^V \tag{6}$$

然后经过式 (7)计算单头注意力得到向量序列$head_i$，其中d_m模型的隐层维度，与向量Q,K,V的维度相等；

$$\begin{aligned} head_i &= \text{Attention}(Q,K,V) \\ &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \end{aligned} \tag{7}$$

再将各个头的向量序列经过式 (8)运算进行拼接，输出音频序列的向量表征c，其中h为多头注意力的头数，W_i^O为随机初始化的参数矩阵；

$$\begin{aligned} c &= \text{MultiHead}(Q,K,V) \\ &= \text{Concat}(head_1, head_2, \cdots, head_h)W^O \end{aligned} \tag{8}$$

最后，通过残差网络将c与h_1相加后经层归一化得到最终的编码器输出表征h_x^A，如式 (9)所示。

$$h_x^A = \text{LayerNorm}(c + h_2) \tag{9}$$

3.3 解码器

本文使用的解码器遵循Transformer 解码器的模型架构。首先将目标语言的文本序列映射为词嵌入向量，继而通过自注意力网络和交叉注意力网络关注到编码器的输出，经过前馈神经网络映射得到解码器的输出，如公式 (10)所示:

$$\begin{aligned} h_e &= \text{Emb}(y) \\ h_y &= \text{Decoder}(h_x, h_e) \end{aligned} \tag{10}$$

通过softmax函数将输出映射到目标语言的词表上，得到目标语言文本序列的预测概率，如公式 (11)所示:

$$p(y|x) = \text{softmax}(h_y) \tag{11}$$

最终，整个语音翻译模型的损失函数如式 (12)所示。

$$L_{st} = -\sum_{(x,y)\in S} \log p(y|x) \tag{12}$$

4 实验设置与结果分析

4.1 实验设置

4.1.1 数据集

实验所采用的数据集来自VLSP2019[4]中的越南语语音识别数据集，该语料包含约416小时的有声小说音频以及人工校对的越南语文本，具体设置见表 1。为进行越南语到英语的语音翻译，调用Google机器翻译服务将越南语文本翻译为英语文本。

数据集	时长/h	句数/k
训练集	395.5	300.0
开发集	13.1	10.0
测试集	7.2	5.5

表 1. 实验数据集

4.1.2 数据预处理

对于语音的Fbank特征，在训练集上使用SpecAugment的LB策略增强语音数据 (Park et al., 2019)，以保证更好的泛化性和鲁棒性。其中，语音的Fbank特征和Pitch特征序列使用均值和方差归一化处理。对于语音的Wav2vec2特征，使用开源的w2v2-vi模型提取512维的语音特征进行实验，下文称w2v2-vi特征。由于普通Transformer模型自注意层的计算复杂度为输入长度的平方，为了对输入数据更有效的计算，实验中采用卷积神经网络对输入序列进行下采样，通过设置不同的层数合理控制输入模型的序列长度，使不同特征的序列长度基本保持一致。所有的卷积层均使用相同配置，步长为2，卷积核大小为5，Wav2vec2、Fbank和Pitch特征下采样的输出维度分别为256、256和32。过滤了小于5帧大于3000帧的音频。

对于目标语言文本，区分大小写同时保留标点。句子使用词表大小为4k的Unigram SentencesPiece模型 (Sennrich et al., 2015)进行分词，采用256维的词嵌入并叠加了位置嵌入。

4.1.3 模型配置与评价指标

为保证实验的公平性，本文所进行的实验均基于Fairseq的Transformer-S2T-S框架[5]，所提方法基于该框架实现。模型的基本配置中，编码器有12层，解码器有6层，多头注意力头数为4，隐层变量维度为256，前馈网络的维度为2048，dropout为0.1。所有实验的训练配置参数均为：使用Adam优化器 (Kingma and Ba, 2014)，其中$\beta_1 = 0.9, \beta_2 = 0.997$；使用标签平滑率为0.1的交叉熵损失作为目标函数 (Müller et al., 2019)；学习率最大阈值为1e-3，学习率预热为10000，使用inverse sqrt 动态调整学习率。整个训练过程在1张Tesla T4 GPU上进行。解码使用大小为5的束搜索算法，使用区分大小写的ScareBLEU[6]作为模型性能的评价指标。

为选定最优的交替周期，选择训练集的20%，将交替特征编码器作为编码器进行初步的语音翻译实验。交替特征编码块中有12个编码块组成，为保证两种编码块均匀分布，交替期周期C分别在2、3、4、6中选择，其中Pitch特征编码比例依次减少。实验结果如表 2所示，其中FP/F表示编码器中FP-Block与F-Block总个数比例。由表可知，在交替周期为3时，编码器获得最佳翻译效果，下文实验均采用该设置。

C	2	3	4	6
P/FP	6/6	4/8	3/9	2/10
BLEU	7.32	8.30	7.96	8.09

表 2. 不同交替周期C在测试集上的BLEU值

[4] https://vlsp.org.vn/
[5] https://github.com/pytorch/fairseq
[6] https://github.com/mjpost/sacrebleu

4.2 实验结果

4.2.1 Fbank特征与Wav2vec2特征在ASR和ST的比较

遵循fairseq的端到端语音到文本 (Wang et al., 2020)的模型设置，先在该编码器-解码器模型上，分别使用Fbank特征和w2v2-vi特征进行语音翻译和语音识别任务，对两种特征均采用2层卷积网络进行下采样。

特征	ASR(WER)	ST(BLEU)
Fbank特征	3.98	37.59
w2v2-vi特征	4.13	36.89

表 3. 特征比较实验

在测试集上的实验结果如表 3所示，Fbank特征在ASR任务上的词错率较w2v2-vi特征低0.15，在ST任务上的BLEU值高0.98。在高资源设置下，使用越南语语音的Fbank特征在ASR任务和ST任务上的性能略优于w2v2-vi特征。

Nguyen et al. (2020)验证了自监督表征在低资源设置下(小于100小时)明显优于fbank表征，在中等资源设置下两种特征的效果接近（100小时至300小时）。在附录 A中，进一步对Wav2vec2特征和Fbank特征，在MuST-C的英语-越南语数据集在不同资源设置下进行比较，实验结果在中等资源设置下两种特征在语音翻译上的性能差异较小，同时表明在高资源设置下（大于300小时），Fbank特征优于Wav2vec2特征。

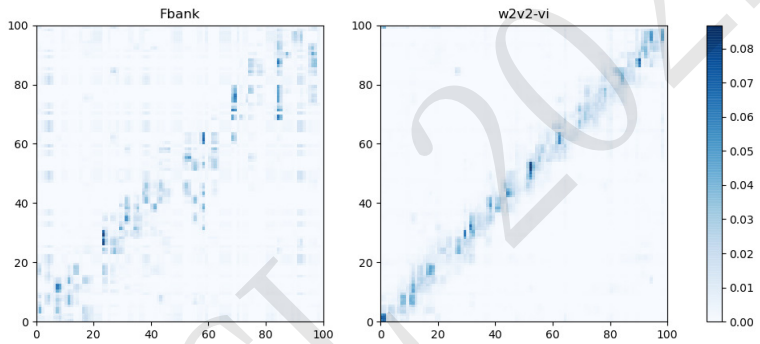


图 4. 两种语音特征编码端的注意力可视化图

Fbank特征在语音翻译任务和语音识别的表现更优，这是因为Fbank特征的计算过程是使用人工设计的结构对语音信号进行降噪和时频分析的过程，能达到较好的效果。而w2v2-vi特征取得的效果相对较差，这是因为该特征编码器预训练阶段在大规模的无标签语音数据上进行自监督的学习，所提取的特征对于语音数据有较好的泛化性，但缺少对目标任务和数据集的适应性。两种特征分别输入到语音翻译模型编码器，其注意力向量序列压缩后的可视化图如图 4所示，由图可知，Fbank特征的注意力更加分散，w2v2-vi特征在编码端学习到的注意力更加集中，这表明不同的特征提取方式所提取的特征存在差异性，促使编码器在训练过程中关注不同的信息。相比于使用人工结构提取的Fbank特征，w2v2-vi在自监督-预训练阶段，从原始波形中动态学习特征提取。从最终输入编码器的序列长度看，经下采样后的Fbank特征序列长度是Wav2vec2特征序列的两倍，而过长的序列长度容易分散注意力。观察w2v2-vi的特征也可发现其注意力更加聚集，说明从原始波形中提取的Wav2vec2特征更加有助于对语音信号进行建模，可以潜在地帮助神经网络发现更好的、更易适应于语音翻译模型的语音表示。

4.2.2 不同融合方法及不同自监督特征的对比实验

为验证融合的有效性，设置单一特征和融合特征两组实验，同时对常用的融合方法和不同的Wav2vec2特征进行比较，选择最优的特征融合方式。使用w2v2-vi和XLSR-53 [7] 两

[7] https://github.com/pytorch/fairseq/blob/main/examples/wav2vec

种Wav2vec2模型提取的特征作为融合特征，分别采用拼接融合-L、拼接融合-F和注意力融合三种融合方式进行实验，其中拼接融合-L将频谱表征h_1和自监督表征h_2在长度维度进行拼接，如式 (13)所示，两种特征的特征维度相同，即$d_w = d_z$，通过在长度维度进行拼接最终编码得到的向量$h_x^L \in \mathbb{R}^{d_w \times (M+N)}$ 作为编码器输出。拼接融合-F将h_1和h_2在特征维度进行拼接，如式 (14)，对Fbank特征和Wav2vec2特征分别应用2层和1层卷积下采样，使得二者长度维度相近在相同数量级，长度差异主要在于卷积时的填充。在特征维度进行拼接时，选择长度最大的特征长度作为表征最终输出的长度。经拼接后输出向量在特征维度是原始向量的两倍，再通过线形层变换为原始表征维度。在该融合方式下最终编码得到的向量$h_x^F \in \mathbb{R}^{d_w \times \max(M,N)}$。

$$h_x^L = \text{ConcatL}(h_1, h_2), h_x^L \in \mathbb{R}^{d_w \times (M+N)} \tag{13}$$

$$h_x^F = \text{Linear}(\text{ConcatF}(\text{h}_1, \text{h}_2)), h_x^F \in \mathbb{R}^{d_w \times \max(M,N)} \tag{14}$$

有无融合	方法	w2v2-vi特征	XLSR-53特征
w/o融合	单一特征	36.89	36.37
w/融合	拼接融合-L	38.78	38.44
	拼接融合-F	37.62	38.17
	注意力融合	38.86	38.63

表 4. 融合方法比较比较实验，w/o融合表示无融合，w/融合表示有融合

由表 3和 4可知，w/融合方法比w/o融合方法提升0.73+BLEU值，表明经两类特征融合得到的编码表征，相比于单一的Fbank、w2v2-vi和XLSR-53特征编码的表征能提升语音翻译的性能，通过卷积网络提取的自监督表征和基于频谱的Fbank特征相融合后，有利于编码语音中的局部和全局信息，来更好的表征语音中的声学信息和语音信息，从而提升翻译性能。从表 4的融合方法看，基于注意力融合方法的BLEU值高于拼接融合-L方法和拼接融合-F方法，通过交叉注意力机制学习两种特征之间的对齐关系，编码输出的序列在序列长度和隐层维度保持不变，解码时不增加额外的计算开销，实验结果表明注意力融合方式是最佳的融合方法。从特征类型看，w2v2-vi特征在注意力融合方式下和拼接融合-L方式下略优于XLSR-53的特征，分析可能原因是w2v2-vi的预训练语料为100小时的越南语有声读物，对越南语音频特征提取有优势，而XLSR-53的训练数据为53k小时的多语言音频，其中越南语占比较少，故对越南语音频的特征提取可能产生干扰。而拼接融合-F方式较其他两种方式性能差距较大，其原因可能是在特征维度进行拼接后使用线性层对特征维度进行降维，而使用单层线性层对两种拼接特征进行降维并非最优的降维方式。

4.2.3 不同模型的对比实验

为验证所提方法的有效性，分别使用Fairseq S2T模型、编码器经ASR预训练的ST模型以及MT和ST多任务联合训练的模型作为基线模型在数据集上进行实验，下文简称Fairseq ST基线，ST+ASR PT基线和MTL ST基线。为公平比较，所有模型均不采用额外数据进行预训练或训练。其中，MTL ST基线中MT和ST的损失均为NLL损失，比重分配为4:6，模型基于Tranformer架构。由表 5中Fairseq S2T基线的实验知，在资源充足的情况下Fbank特征的翻译效果优于Wav2vec2特征，故其余基线模型均使用Fbank特征作为输入特征。

实验结果如表 5所示，相比于Fairseq ST基线模型，ST+ASR PT 基线采用经过ASR预训练后的编码器参数来初始化ST模型的编码器，充分利用单语数据来学习语音中的声学信息，提升了1.18个BLEU值。MTL ST基线通过共享解码器参数来进行文本翻译任务和语音翻译任务的联合训练，由于训练过程不采用额外数据,且联合训练过程的损失分配导致单个任务的性能非最优的结果，故而相比与Fairseq ST基线下降3.64个BLEU值。所提模型相比于最优的ST+ASR PT基线提升了0.79个BLEU值，相比于使用Fbank特征的Fairseq ST基线提升了1.97个BLEU值。使用Fbank-Pitch交替特征编码层和表征融合层来融合w2v2-vi特征和Pitch特征，不同特征间的差异性使其相互补充得到更丰富的编码表征，因此翻译质量得到进一步的提升。

方法	特征	BLEU	参数量/M
Fairseq ST	w2v2-vi	36.89	47.0
Fairseq ST	Fbank	37.59	47.5
MTL ST	Fbank	33.95	84.0
ST+ASR PT	Fbank	38.77	47.5
所提方法	Fbank+w2v2-vi+Pitch	39.56	45.4

表 5. 四种不同模型的BLEU值

此外，所提模型同时对输入音频的三种特征进行编码，但参数量略小于Fairseq ST基线，这是因为所提方法在交替特征编码器中，使用F特征编码块和FP特征编码块交替编码，其中FP特征编码块参数量小于F特征编码块，且对Wav2vec2编码层和特征融合层层数少，具体设置见4.1节。所提模型相比于ST+ASR PT基线和MTL ST基线,不需要额外对模型的部分模块进行预训练或联合训练步骤，训练效率更高。

4.2.4 消融实验

相比表 5所列的三类端到端基线模型，本文提出交替特征编码器和表征融合层来融合额外的Pitch特征和Wav2vec2特征。本节对所提模型进行了消融实验，评估所提方法中不同模块及额外特征对模型性能的贡献。

模型	BLEU
所提方法	39.56
- 交替特征编码块	38.97
- Pitch特征	38.86
- 特征融合层	37.95
- Wav2vec2特征	37.52

表 6. 消融实验结果

由表 6可知，所提出的不同特征编码模块及融合不同特征对模型性能均能带来正向增益，全部使用可以达到最优结果。交替融合模块按实际区分语义的成分的比重将Pitch特征和Fbank特征进行有效融合，与直接采用Fbank特征和Pitch特征在特征维度进行拼接的方式相比，可以带来0.59个BLEU的提升。进一步去掉Pitch特征，直接使用Fbank的自注意编码表征，翻译性能继续下降0.11个BLEU值。这意味着，越南语的Pitch特征中可能包含能提高语音翻译效果的语义信息，证明了Pitch特征对于有音调语言语音建模的必要性。表征融合层将自监督表征和频谱表征使用交叉注意力机制进行深度融合，来增强编码端输出的表征，去掉该层直接将两种表征用式 (14)在特征维度拼接，性能会下降1.97个BLEU值。进一步去掉Wav2vec2特征，即只使用交替融合模块混合Fbank特征和Pitch特征，翻译性能会继续下降0.43，这表明Wav2vec2特征可以对Fbank特征进行补充，该补充对提升语音翻译的性能是有益的。

5 结论

针对单一特征对复杂语音表征能力不足的问题，本文根据Fbank特征和Wav2vec2特征之间的差异性以及越南语语音特点，使用多特征融合的语音翻译框架将Fbank特征、Wav2vec2特征及Pitch特征进行有效融合，使不同特征间相互补充，增强编码端输出的表征对声学信息和语义信息的表征能力，从而提高越英语音翻译的性能。未来的工作将探索其他自监督表征与传统特征的深度融合，在真实的噪音数据集及低资源场景下的表现，并验证所提方法在其他音调语言上的有效性。

参考文献

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, December.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel Lopez-Francisco, Jonathan D. Amith, and Shinji Watanabe. 2022. Combining spectral and self-supervised features for low resource speech recognition and translation.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California, June. Association for Computational Linguistics.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online, August. Association for Computational Linguistics.

Van Huy Nguyen. 2019. An end-to-end model for vietnamese speech recognition. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7904–7908.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. Interspeech 2017*, pages 2630–2634.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pages 1128–1132.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *CoRR*, abs/2010.14920.

Abdel-rahman Mohamed. 2014. *Deep Neural Network Acoustic Models for ASR*. Thesis. 1 online resource.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Cao Hong Nga, Chung-Ting Li, Yung-Hui Li, and Jia-Ching Wang. 2021. A survey of vietnamese automatic speech recognition. In *2021 9th International Conference on Orange Technology (ICOT)*, pages 1–4.

Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Yannick Estève, and Laurent Besacier. 2020. Investigating Self-supervised Pre-training for End-to-end Speech Translation. In *Interspeech 2020*, Shangai (Virtual Conf), China, October.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany, December 5-6.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

F. W. M. Stentiford and M. G. Steer, 1990. *Machine Translation of Speech*, page 183–196. Chapman & Hall, Ltd., GBR.

Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney. 2014. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Proc. Interspeech 2014*, pages 890–894.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Miguel Pino. 2020. fairseq S2T: fast speech-to-text modeling with fairseq. *CoRR*, abs/2010.05171.

A Fbank与Wav2vec2特征比较

为进一步比较Wav2vec2特征和Fbank特征，使用MuST-C的英语-越南语数据集上基于Tranformer进行语音翻译实验，表 7为原始数据集的划分。

数据集	时长/h	句数/k
train	432.9	230.9
dev	2.5	1.3
tst-HE	1.2	0.6

表 7. MuCT-C 英语-越南语数据集

本文对该数据集的训练集进行进一步划分，分别抽取训练集时长的100%、75%、50%划分为新的训练集，分别记作train-1.0、train-0.75、train-0.5如表 8所示。对于英语音频的Wav2vec2使用开源的Wav2Vec 2.0 Base [7]预训练模型进行提取，为加速模型收敛，均采用ASR预训练，由于两种特征序列长度的差异，实验结果采用在相同步数下进行比较。其余实验设置与 4.2.1中相同。

训练集	train-1.0	train-0.75	train-0.5
时长/h	432.9	324.4	216.5
Fbank	22.93	21.09	18.01
Wav2vec2	21.56	20.08	17.19

表 8. 不同资源设置下，两种特征在tst-HE测试集上的BLEU

实验结果如表 8所示，由表可知，在train-1上，Fbank特征的翻译效果要略优于Wav2vec2特征；在train-0.5和train-0.75上，两种特征的翻译性能基本持平，且随着训练集的减少，两种特征的性能差距逐渐减小。