基于多尺度建模的端到端自动语音识别方法

陈昊，张润来，张裕浩，高成浩，许晨，马安香，肖桐*，朱靖波
东北大学计算机科学与工程学院自然语言处理实验室，沈阳，中国
methanechen@126.com, me_henrychang@163.com, yoohao.zhang@gmail.com,
{gaochrishao, xuchenneu}@outlook.com,
{maanxiang, xiaotong, zhujingbo}@mail.neu.edu.cn

摘要

近年来，基于深度学习的端到端自动语音识别模型直接对语音和文本进行建模，结构简单且性能上也具有显著优势，逐渐成为主流。然而，由于连续的语音信号与离散的文本在长度及表示尺度上存在巨大差异，二者间的模态鸿沟问题是该类任务一直存在的困扰。为解决该问题，本文提出了多尺度语音识别建模方法，该方法从利用细粒度分布知识的角度出发，建立多个不同尺度形式的文本信息，将特征序列从细粒度的低层次序列逐步对齐预测出文本序列。这种逐级预测的方式能够有效降低预测难度，缓解模态鸿沟带来的影响，并通过融合不同尺度下特征，提高语料信息的丰富性与完整性，进一步增强模型推理能力。本文在LibriSpeech小规模、大规模和TEDLIUM2数据集上实验，相比基线系统词错误率平均降低1.7、0.45和0.76，验证了方法的有效性。

关键词： 多尺度建模；特征融合；语音识别

An End-to-End Automatic Speech Recognition Method Based on Multiscale Modeling

Hao Chen, Runlai Zhang, Yuhao Zhang, Chenghao Gao, Chen Xu, Anxiang Ma, Tong Xiao*, Jingbo Zhu
NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
methanechen@126.com, me_henrychang@163.com, yoohao.zhang@gmail.com,
{gaochrishao, xuchenneu}@outlook.com,
{maanxiang, xiaotong, zhujingbo}@mail.neu.edu.cn

Abstract

End-to-end automatic speech recognition models based on deep learning that directly model speech and text have become mainstream due to their simple structure and remarkable performance. However, a persistent challenge in such tasks is the modality gap between continuous speech signals and discrete text, which arises from the significant differences in length and representation scale between the two modalities. To address the problem, this paper proposes a multi-scale speech recognition modeling method that builds multiple scales of text information from the perspective of using more fine-grained distribution knowledge. This progressive prediction approach effectively reduces the difficulty of prediction, mitigates the impact of the modality gap. The approach also enhances the model's inference capability, enriches and complements the information in the speech data by fusing features from different scales. Our method is effective on LibriSpeech small-scale, large-scale, and TEDLIUM2 datasets, showing an average reduction in word error rates of about 1.7, 0.45 and 0.76 compared to baseline.

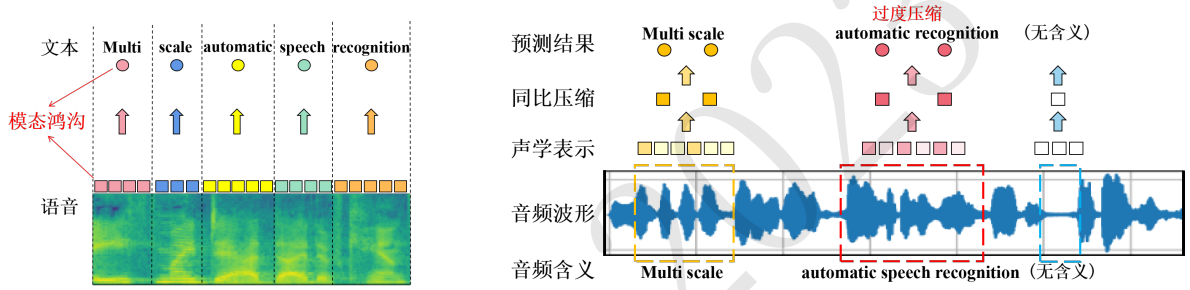
Keywords: Multiscale modeling , Feature fusion , Speech recognition

*通信作者：肖桐（xiaotong@mail.neu.edu.cn）

1 引言

自动语音识别(Automatic Speech Recognition, ASR)任务旨在将连续输入的语音信号转换为相应的输出文本，广泛应用于会议演讲、智慧办公、智能系统等日常生活领域。传统的语音识别系统需要预先对语音信号进行处理并提取特征，联合训练好的声学模型、语言模型、发音词典共同寻找由特征序列决定的最优状态序列，从而得到识别结果(Huang et al., 2001)。然而该方法不仅需要大量的人工对不同语言发音知识进行总结，随之也带来了模型更新困难、泛化能力弱、降噪能力不强等问题，进而难以满足复杂交流场景的需要。近年来，随着深度学习的兴起，神经网络模型在多个人工智能任务均取得了重要成就，在语音识别任务中基于端到端自动语音识别模型(End-to-End Automatic Speech Recognition, E2E ASR)(Graves and Jaitly, 2014)通过直接建模语音到文本的映射，模型结构简洁且大大简化了训练过程，在众多工作中已被证明能够取得更加优异的性能，逐渐成为主流。

然而，这种基于端到端语音到文本的模型在对音频进行建模时存在一个自然的问题：语音通过连续的声音信号进行传递，而文本是通过离散的符号序列进行传递，在图1(a)中可以发现，相同含义的语音和文本之间无论是序列长度还是内容的表示尺度均存在较大的差异，这种模态鸿沟(Modality Gap)(Fang et al., 2022)无疑给该类任务造成了巨大的困扰。例如，在语音处理中通常以帧级别为最小单元，即使文本处理任务中最小单元使用字符级别，二者序列长度上仍存在着数十倍的差距，同时，基于帧级别的特征信息并不足以预测出字级别文本信息，这两个问题导致了模型难以将二者进行准确对齐，从而影响预测结果。



(a) 语音与文本的模态鸿沟　　(b) 不同音频片段所需压缩方式不同

图1.语音与文本模态差异

为了减轻端到端自动语音识别模型输入与输出之间差距带来的压力，一种简单的方式是将音频特征信息进行压缩，进而对齐到文本的建模粒度上，这一定程度上能够减轻模态差异造成的干扰。在计算机视觉领域中同样面临着模态鸿沟的问题，研究者们采用金字塔结构(Fan et al., 2021)帮助模型逐渐地从图像中抽取有用信息进而对齐到需要的文本粒度上。然而，语音识别任务由于文本中不同词或者字的发音长度不尽相同，这要求金字塔结构需要根据不同的文本调整压缩比例，无疑是十分困难的。这个原因也导致了直接使用金字塔结构进行降采样会使得音频特征中重要信息被压缩。图1(b)从音频波形的角度可直观看到语音不同片段需要的降采样策略不能完全一致，采用统一的压缩策略会使得信息集中的地方被过度压缩。针对这一现象，本文展示了金字塔结构（如图2）在语音识别任务中存在的过度压缩问题，进而提出多尺度语音识别建模方法。该方法从利用更细粒度的分布知识的角度出发，建立多个不同尺度形式的文本信息供模型学习，利用细粒度层次上下文知识指导粗粒度数据的处理，进而防止模型在金字塔结构处理过程中一些低层次信息被过度压缩。本文为待识别的语音特征建立对应的词级别文本以外，同时建立对应的音素级别、字符级别文本信息共同参与训练，使模型在合适的粒度上进行对齐。这种逐级预测的方式不仅缓解了语音与文本之间粒度差距过大难以对齐的问题，并且能够通过融合不同尺度空间下的文本信息，使得语料信息更为丰富完整，缓解语音数据稀缺的问题(Zhang et al., 2022b)，例如字符以及子词级别文本侧重语料语义的理解，音素级别则更适应于声学信号的表达，二者信息能够有效互补，弥补了模型对音频过度压缩带来的损失，进一步改善自动语音识别效果。本文在LibriSpeech小规模和大规模数据以及TEDLIUM2数据集上进行了实验，本文方法相比基线系统词错误率平均约降低1.7、0.45和0.76，验证了所提出方法的有效性。

本文主要有如下贡献：

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

（1）本文利用语言发音的先验知识设计了多尺度文本金字塔结构，并发现金字塔结构在语音识别任务中存在的过度压缩问题。

（2）本文提出多尺度语音建模方法，引入子词、音素、字符等细粒度级别信息，并利用连接主义时序分类预测不同尺度的对齐效果，实现模型从细粒度的低层次序列逐步对齐预测出词序列，缓解了语音文本间模态鸿沟问题。

（3）本文提出多尺度特征融合方法，引入更多元化特征，有效补充相同语义下基于不同尺度特点的信息，提高信息完整性与丰富性，改善了由于压缩而导致的语义信息丢失的问题。

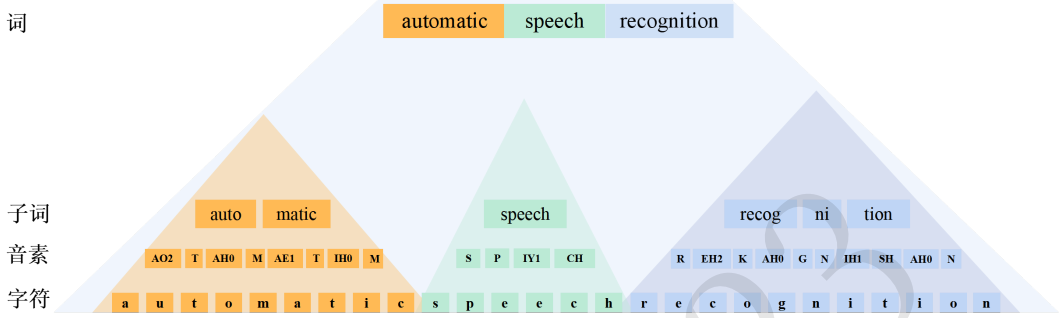


图2.多尺度文本金字塔结构

2 相关工作

语音识别技术的发展可追溯到上世纪五十年代，从最开始的模板匹配阶段到统计模型阶段，早期研究普遍集中在隐马尔科夫模型(Hidden Markov Model, HMM)(Baum et al., 1970)与统计模型例如高斯混合模型(Gaussian Mixture Model, GMM)(Dempster et al., 1977)的结合，由于模型泛化能力弱、算法复杂度高且数据质量难以保证等问题，该类模型在日常对话、新闻播报等场景下的识别率只能达到80%左右，应用普遍受到限制。在2010年后通过引入深度神经网络(Deep Neural Networks, DNN)(LeCun et al., 2015; Goodfellow et al., 2016)后发现，DNN能够将相邻帧建模到一起作为输入，更适合处理语音信号这种复杂信号。这种基于输入与输出的端到端建模方法大大简化了建模过程，通过目标函数直接训练神经网络模型，并对语音识别过程进行优化，提高语音识别精度。目前主流端到端模型主要包括连接主义时序分类(Connectionist Temporal Classification, CTC)(Graves et al., 2006)、循环神经网络转换器(Recurrent Neural Network Transducer, RNN-T)(Chorowski et al., 2014; Chorowski et al., 2015; Chan et al., 2016)和基于注意力机制的端到端语音识别(Attention Based Encoder-Decoder, AED)(Rao et al., 2017)等，然而不同端到端模型并不尽相同，例如CTC通过引入Blank机制，借鉴隐马尔科夫和动态规划思想实现标签硬对齐，而基于注意力机制的方法则通过序列模型实现输入与输出的软对齐，不同特点的模型在不同工作中均取得显著优势。2017年Transformer(Vaswani et al., 2017)架构问世，将Transformer应用于语音领域也进一步取得了明显的提升效果(Zhou et al., 2018; Dong et al., 2018)。

在语音任务中，通常以连续的帧级别特征作为标准的模型输入，而由于音频本身具有信息稀疏以及长序列的特点，导致生成的特征序列比对应文本序列长的多。这种过长序列不仅会导致捕获长距离依赖关系变的更为困难，注意力机制分配也会受到影响。为减少音频序列长度，目前多通过对特征序列进行降采样的方式(Guo et al., 2020)，通过堆叠多层卷积操作将多个相邻位置进行压缩，该方法操作简单且计算代价较低，但忽略了音频本身信息分布不均以及音频与文本之间的差异，会导致特征信息的过度采样以及遗漏问题。随着降采样策略不断发展(Xu et al., 2023b)，计算机视觉领域中提出了“金字塔”策略(Fan et al., 2021)解决类似问题。相关研究发现通过在更低分辨率下关注图像重要特征不仅能够减少计算需求，同样也可以帮助模型理解上下文以指导高分辨率下的处理(Rosenfeld and Thurston, 1971; Burt and Adelson, 1987)。

在语音任务中，为了促进端到端语音识别任务中词级别的表示学习，可通过让模型逐步学习难度提升的抽象语言序列(Higuchi et al., 2021)，为目标文本构建更细力度尺度信息，将特征序列从细粒度的低层次序列逐步对齐到子词序列，直至最后预测出单词级别序列，Jelinek and F. (1976)的工作也表明逐步提高语言信息的抽象水平对于训练语音识别模型是一种合理的方式。

除了对特征序列进行相关处理外，语音特征表示自身对语音下游任务同样具有显著的影响(Deng et al., 2013)，单一特征往往难以包含语音中语言、情感、韵律等多元化信息，近年来研究者们开始探索将多种特征经过融合用于语音下游任务。由于不同声学特征可通过不同角度对语音任务起到帮助作用，故而语音任务中可通过不同特征融合互补发挥各自优势，该方式保留了大部分信息，但同时也增加了特征维度。在先前工作中，Yoon et al.(2018; 2019)通过将韵律特征与MFCC特征融合应用于语音情感识别任务；袁文浩et al. (2021)将时域特征与频域特征融合实现语音增强；Zhang et al. (2022a)通过融合不同频率下Mel滤波器组特征在说话人识别任务中达到了最先进的性能。

3 本文方法

3.1 基线

模型结构 在基线中，本文主要采用Transformer模型用作端到端语音识别任务，与标准Transformer模型结构基本一致，在编码器输入时，通过堆叠两层步长为2的卷积模块将输入的语音特征进行压缩，将长度压缩为原先1/4，以此降低序列长度，节约计算资源，模型总体采用12层编码器，6层解码器的结构。

连接主义时序分类 连接主义时序分类(CTC)是一种在符号序列上训练而不需要对齐的递归网络方法。传统的基于隐马尔科夫模型的深度神经网络语音识别模型都需要预先建立输入语音特征与输出标签之间的对应关系，耗时耗力且难以保证对齐的准确性。不同于传统的深度神经网络语音识别模型，CTC是一种端到端的模型训练方法，通过将模型的输出层进行扩展，使输出文本和标签建立对应关系，模型能够直接对输入的语音特征进行训练，从而输出预测序列的概率。Xu et al. (2023a)将CTC应用在语音任务的监督学习取得了显著效果。

给定输入序列X=[x_1,x_2,x_3...,x_T]以及对应的标签数据Y=[y_1,y_2,y_3...,y_U]，分别对应自动语音识别任务中音频特征序列以及文本序列，CTC返回给定输入序列X的所有可能Y的输出分布，根据输出概率输出最有可能的结果。令$p(l|x)$表示输入为x，输出为序列l的概率，$p(l|x)$形式化定义如下：

$$P(l|x) = \sum_{\pi \in F^{-1}(l)} p(\pi|x) \tag{1}$$

其中，$\pi \in F^{-1}(l)$代表所有经过F^{-1}变换（将神经网络输出的原始预测序列转换为最终输出序列）后是l的路径π，y_t^k表示π路径下t时刻的概率值，对于任意一条路径π都有：

$$P(\pi|x) = \prod_{t=1}^{T} (y_t^k) \tag{2}$$

虽然音频序列远远大于文本序列，但是由于引入了空白位置对齐的方式，这种CTC建模方式能够实现两者的一一对应。这种特性使得本文可以根据当前音频的建模粒度，将音频序列对齐到相应的文本表示粒度，如字符、音素、子词等。

3.2 金字塔结构

金字塔结构是一种在计算机视觉中广泛使用的图像处理方法，通过将图像分解为不同尺度的子图像，对图像进行多尺度分析。这种结构的核心在于将降采样模块适当地加入到模型结构中，使其逐步地聚合信息，之后再利用其他的模块对压缩之后特征进行处理。由于在信息聚合的过程中，特征长度在逐渐减少，而每个位置上的信息逐渐增加，最后得到表示更有利于尺度更小的文本任务。如在图3中，本文在编码端中插入了一些降采样模块，这些模块通过聚合相邻单元的信息获得更细粒度的音频特征表示。为便于模型对于进入多尺度层向量表示序列的处理，本文采用层标准化(Layer Normalization)(Ba et al., 2016)的方式，规范向量输出，利用正则化的方式，增强模型对于不同尺度数据的适应性，提高模型训练的速度和性能，降低了过拟合的风险。这个结构作为语音**金字塔结构**模型。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

为了确定多尺度降采样比例，本文在LibriSpeech-100h数据集上统计了各个尺度文本与输入特征序列的最小长度比（见表1），根据长度关系可发现，输入序列长度至少为字符以及音素级别序列长度4倍，至少为子词级别序列12倍，基于此本文设计了在每次对齐前引入一层卷积神经网络的策略，将序列压缩为原先长度的二分之一，降低模型训练压力并且避免二者差距过大导致性能下降。同时针对编码器输入时的两层卷积优化为一层，在最终对齐到文本输出时总共实现8倍压缩，控制了编码器端对向量序列的压缩比例。

前文已经提到，语音信号是一种时间序列信号，相邻时间点的采样值存在强相关性和依赖性(Rabiner and Juang, 1993)。因此利用这种结构会导致语音中一些小尺度的信息如字符级别将会被过度地压缩，进而导致了关键信息的丢失，影响语音预测结果。

不同尺度文本关系	最小长度比
输入序列/字符级别序列	4.20
输入序列/音素级别序列	4.86
输入序列/子词级别序列	12.31

表1.LibriSpeech-100h数据集不同尺度文本最小长度比

3.3 多尺度语音建模

基于此，本文在每次对特征序列利用卷积神经网络降采样之前，尝试及时地将音频信息对齐到对应的文本粒度，以防止对应尺度的信息在之后的处理中丢失。同时本文引入了不同文本尺度信息，丰富语音整个建模过程。根据先验知识，由于模型在中间位置层时可以同时获得全局和局部的特征信息，具有足够的高层抽象特征和低层原始特征，能够为多尺度信息的提取提供足够的基础(Lin et al., 2017)。如果将多尺度信息引入到其他层中，可能会降低模型的表现能力或者增加模型的复杂度。如图3所示，本文在编码器端中间第六、九层以及编码端的输出位置分别引入字符级别，音素级别以及子词级别的信息，通过在不同的尺度上对输入序列进行建模来捕捉输入序列中更加细致的信息，该结构为本文提出的**多尺度语音建模**方法。考虑到编码端在建模过程中已经有足够的跨模态抽象特征，可以为对应尺度的处理提供足够的信息，因此该方法直接对指定层的输出向量表示序列进行记录，并利用CTC函数辅助该序列与相应尺度文本进行对齐，之后向量序列经过压缩送到下一种尺度的编码端处理或者送到解码端，整个方法模拟了从细粒度的低层次序列逐步对齐到子词序列的过程。

具体而言，整个对齐过程可分为三个阶段：假设输入模型的特征序列为X，进入编码器端第i层特征序列表示为X_i，字符级别多尺度文本为Y_c，音素级别多尺度文本为Y_p，子词级别文本为Y_w，则三个阶段的CTC对齐损失函数可以分别表示为：

1.特征序列对齐到字符级别多尺度文本的CTC对齐损失函数：

$$L_c = -\log P(Y_c|X_6) \tag{3}$$

2.经过字符级别对齐后的特征序列，对齐到音素级别多尺度文本的CTC对齐损失函数：

$$L_p = -\log P(Y_p|X_9) \tag{4}$$

3.经过音素级别对齐后的特征序列，对齐到子词级别多尺度文本的CTC对齐损失函数：

$$L_w = -\log P(Y_w|X_{12}) \tag{5}$$

其中$P(Y_c|X_6)$、$P(Y_p|X_9)$、$P(Y_w|X_{12})$表示给定输入相应语音特征语音特征序列的条件下，该序列与对应尺度文本字符序列匹配的概率。三个阶段的对齐损失函数可以合并为一个多尺度损失函数L_{ms}：

$$L_{ms} = \alpha * (L_c + L_p + L_w) \tag{6}$$

其中α是超参数，作为缩放因子调节损失大小。通过最小化总损失函数L_{ms}，可以优化特征序列与多尺度文本之间的对齐过程，进而将多尺度信息的引入到建模过程中。这种方法能够提高语音识别的准确率和鲁棒性，使得模型同时捕捉输入序列中的全局特征和局部特征，从而增强模型的表现能力。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

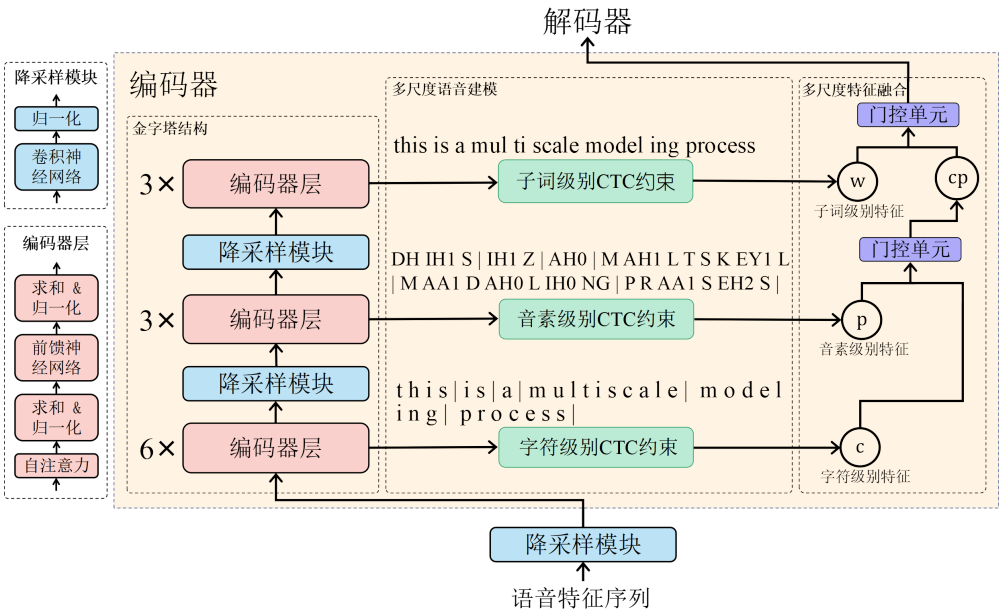


图3.多尺度语音识别建模方法

3.4 多尺度特征融合

为进一步探讨多尺度特征对端到端语音识别任务的促进作用，本文将不同尺度文本特征序列根据一定方式进行融合，通过引入多元化特征改善压缩过度导致的信息丢失问题。在特征融合上，本文主要考虑从长度以及维度两个角度融合多尺度特征。长度融合的问题在于不同尺度的特征长度并不一致，本文依旧采用了高效的卷积网络将其他尺度的特征长度对齐到子词级别的特征。

虽然长度不一致问题能够通过卷积进行处理，但是也难以保证得到不同尺度的特征之间是能够相互帮助，而不是给其他级别特征引入噪音进而降低了特征质量。因此在特征维度融合的角度，考虑到不同尺度特征在训练过程中贡献并不完全相同，可通过引入门控机制来解决这个问题。门控循环单元(Gated Recurrent Unit, GRU)(Hochreiter and Schmidhuber, 2016) 是一种常用的循环神经网络模型，它通过门控单元来控制信息的流动和捕捉长期依赖关系，帮助网络动态的调整每个特征的权重，从而更灵活地融合特征。基于此本文提出了一种基于门控的多尺度特征融合方法，将字符、音素和子词级别的特征进行融合，使得各种尺度之间的特征相互帮助。如图3中，通过逐级地两两融合特征，将其作为**多尺度特征融合**方法。

具体而言，首先将字符级别的特征F_c和音素级别的特征F_p通过一个门控模块进行融合，其中门控机制使用了一些可学习的参数，这些参数通过$sigmoid$函数进行非线性变换，从而控制门的打开程度，并通过预设融合比例实现特征融合的目的，得到融合后的特征F_{cp}，如下所示：

$$F_{cp} = W_{cp} \odot F_c + (1 - W_{cp}) \odot F_p \tag{7}$$

$$W_{cp} = a \cdot \sigma(arg_{cp}) + b \tag{8}$$

其中W_{cp}为字符级别特征与音素级别特征融合时的门控模块，σ表示sigmoid激活函数，arg_{cp}为定义的字符音素特征融合可学习参数，共同参与神经网络模型训练。a为门控参数的sigmoid函数的偏置，b则是为了保证门控参数的范围，以避免过于偏向某个特征而导致融合结果出现偏差，二者均为可调节的超参数用于控制不同特征的融合比例。

接着，将融合后的特征和子词级别特征以相同的方式再次融合，得到最终的特征表示F，如下所示：

$$F = W_{cpw} \odot F_{cp} + (1 - W_{cpw}) \odot F_w \tag{9}$$

$$W_{cpw} = a \cdot \sigma(arg_{cpw}) + b \tag{10}$$

其中，W_{cpw}为字符音素融合特征以及子词级别特征融合时的门控模块，a与b与公式8一致，arg_{cpw}为定义的多特征融合可学习参数。

本文使用门控模型对不同级别的特征进行加权融合，以避免低质量特征的干扰。在每次融合前，利用卷积神经网络对低层特征实现两倍压缩，将待融合的两个特征序列统一到相同的长度，对融合后的特征进行层标准化操作，以缓解梯度消失和梯度爆炸的问题。

4 实验

4.1 数据处理

本文实验主要基于LibriSpeech数据集(Panayotov et al., 2015)100小时子集、960小时完整数据集以及TEDLIUM2数据集(Rousseau et al., 2014)，在相应数据集上训练模型并使用标准验证集和测试集对模型训练结果进行评分。

特征提取 音频信号中存在着许多不同的特征，机器通过分析语音找到特征对应的特征参数，从语音中提取出能够有效反映关键特征参数的特征向量序列的过程就是特征提取，目前常用的几种特征参数包括：Mel 频率倒数系数(Mel-Frequency Cepstral Coefficients, MFCC)、基于滤波器库的Fbank (Log-Mel-Filter Bank)特征、线性预测分析(Linear Prediction Coefficient, LPC)等。基于人耳结构特点，Stevens and S. (1936)提出符合人耳听觉特性的Mel尺度，在Mel尺度下，对三角滤波器组输出取对数即可得到Fbank特征，Fbank特征具有二维结构，可以通过引入卷积神经网络进行处理，本文主要采用Fbank 特征进行实验。

尺度处理 本文为训练建立多尺度信息时，主要根据目标文本（本文以英语为例）建立了字符级别与音素级别文本，字符级别旨在将目标文本中所有单词均变为以单个字符为基本单元的形式，例如“speech”则对应为“s p e e c h”；在处理音素级别时，本文利用CMU发音词典[1](Carnegie Mellon University Pronouncing Dictionary)，将单词转为相应的发音音素格式，例如“speech”对应为“S P IY1 CH”，CMU词典是一本面向北美英语的开源机器可读发音词典，该词典音素集主要包含39个音素以及超过134000个单词及其发音，具有从单词到发音的映射，并依然在不断更新，在语音识别和合成领域发挥了巨大作用。而部分未出现在CMU词典中的单词，由于不能转换为对应的音素形式，本文利用谷歌开发的开源工具sentencepiece[2]先对词级别目标文本构建词典以及模型，利用该模型对音素级训练语料中仍然存在的单词级别文本进行进一步切分，将单词切分为更细粒度的语义单元。在模型训练时，本文将三个尺度文本信息通过统计词频的方式统一到大小为10000的同一张词表，以此降低训练代价。

速度扰动 由于端到端模型常常需要大规模数据进行训练，在小规模数据集时表现往往受到限制，因此数据增强相关方法也被广泛研究。速度扰动(Speed Perturb)(Liberman and Mattingly, 1985)是一种数据增广方法，主要通过对音频播放速度的调整，得到更快语速或更慢语速的语音数据，从不同语速的语音数据中分别提取特征序列，以达到训练数据的扩充。本文为LibriSpeech-100h以及TEDLIUM2数据集语音分别设置0.9、1.0、1.1倍语速，将训练数据扩充为原数据集三倍，弥补小规模数据集可能导致的训练不充分问题。

4.2 实验设置

本文实验主要基于Fairseq的S2T框架[3]，并在该框架基础上完成本文提出的多尺度建模方法。模型配置上，除了前文介绍的12层编码器，6层解码器架构外，各层隐层变量维度均为256，前馈网络维度均为2048，多头注意力头数为4，dropout为0.1。学习率最大阈值为2e-7，学习率预热迭代次数为10000，并采用inverse sqrt对学习率动态调整。使用Adam优化算法，其中两次估计指数衰减率分别为β_1=0.9，β_2=0.98，使用标签平滑率为0.1的交叉熵损失作为目标函数。在引入金字塔模型时，本文在第六层和第九层加入降采样模块。每次降采样均采用步长为2，卷积核大小为5的卷积层实现。在训练时，利用CTC损失函数辅助模型训练，对于本文中涉及到多重CTC损失（即L_w、L_p、L_c）权重α设为0.2，在特征融合时，特征融合比例参数a=0.2，b=0.4。训练结束时，对训练最后十轮模型参数进行平均，采用大小为5的束搜索算法(Jelinek, 1980)进行解码。本文通过词错误率(Word Error Rate, WER)来评价自动语音识别的效果。

[1]http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2]https://github.com/google/sentencepiece
[3]https://github.com/pytorch/fairseq

4.3 实验结果

在本文的主要实验中，以LibriSpeech以及TEDLIUM2数据集实验得到的结果如表2所示。本文使用fairseq的端到端语音到文本模型(Wang et al., 2020)作为基线模型，并在此基础上引入了所提出的多尺度语音建模以及多尺度特征融合方法，本文模型参数量大小均为30M。此外，本文还对比了Higuchi et al. (2021)、Andrusenko et al. (2022)在相同数据集上的相关工作。实验结果表明，本文基线系统相比先前工作已经有了一定提升。然而，当直接采用金字塔结构时，性能下降较为明显，验证了前文介绍的金字塔结构容易对语音信息造成过度压缩的问题。本文提出的多尺度语音建模方法在LibriSpeech100小时测试集上的WER相对于基线分别降低了0.66和1.38，在TEDLIUM2测试集降低了0.68，在LibriSpeech960小时测试集other上降低了0.36，从而证明了多尺度语音建模方法的有效性，而在clean数据集上，由于语音信号的质量较高且干扰较少，传统的单尺度模型已经能够有效地识别语音，使用多尺度模型可能反而增加模型的复杂度，对于更加复杂的数据集，多尺度模型可能会比单尺度模型更有效。此外，本文引入了多个尺度特征共同参与训练，并采用基于多尺度特征融合的方法来进一步提升识别性能。实验结果表明，在LibriSpeech-100h测试集上WER表现比基线分别低1.37和2.03，在LibriSpeech-960h测试集上比基线低0.24和0.66，在TEDLIUM2测试集比基线低0.76，WER相对于多尺度建模方法降低了一定的程度，这证明了多尺度特征融合对于多尺度建模过程的进一步推动作用。

方法	LibriSpeech-100h				LibriSpeech-960h				TED LIUM2
	Dev		Test		Dev		Test		
	clean	other	clean	other	clean	other	clean	other	Test
Higuchi	11.50	24.80	11.80	25.50	4.20	10.00	4.50	9.90	10.70
Andrusenko	10.40	27.10	10.70	27.10	3.70	10.10	3.70	9.90	-
本文基线	9.68	23.37	10.46	23.28	4.17	9.66	4.41	9.22	10.58
金字塔结构	10.28	22.52	11.97	24.63	4.20	9.39	5.16	9.35	10.71
多尺度语音建模	8.96	21.26	9.80	21.90	4.13	8.92	4.72	8.86	9.90
多尺度特征融合	**8.39**	**21.23**	**9.09**	**21.25**	**3.86**	**8.69**	**4.17**	**8.56**	**9.82**

表2.主实验结果

值得注意的是，本文对比了LibriSpeech小规模和大规模数据集的识别精度，实验结果表明，在小规模数据集上，本文方法比传统方法取得了更好的效果。根据分析，一个可能的原因可能是由于小规模数据集的数据量较少，模型容易出现过拟合的情况，同时数据中的噪声和变化也较大，这会影响模型的泛化能力和识别精度。而本文方法能够在不同尺度下提取不同粒度的语音特征，并通过多层CTC对齐实现特征序列到字符、音素和子词三级对齐，从而更好地捕捉语音信号的结构和特征，提高模型的鲁棒性和泛化能力，因此在小规模数据集上取得更好的效果。而在大规模数据上，传统方法往往已经能够很好地处理数据的分布和噪声，同时具有更高的样本覆盖率和更多的数据多样性，导致了本文方法效果提升更微弱。

与其他人工作对比，也不难发现本文方法在多项指标上均有优势，说明多尺度建模方法利用多个尺度的特征信息，能够更全面地捕捉语音信号的特征，并通过层次化的处理逐渐融合和传递上下文信息，可以更好地捕捉语音信号中的上下文关系，以及同时处理多个尺度的特征，对噪声和干扰具有更好的抗干扰能力，具备更出色的语音建模能力。而在LibriSpeech960小时clean数据集上，我们方法略差于Andrusenko et al. (2022)，主要考虑为该工作中使用的Conformer结构在语音任务中更好地处理时间关系、全局上下文和位置关系，因此相对于传统的Transformer结构，更适用于语音识别任务，在后续工作中我们也将在Conformer结构上展开进一步实验。

5 实验分析

5.1 多尺度语音建模消融实验

为了验证不同尺度CTC都能对语音建模起到帮助作用，本文在LibriSpeech100小时数据集上展开实验（见表3），通过分别去掉字符级别、音素级别、子词级别CTC约束进行实验，可发现任意一种CTC约束的去除都会对语音识别的性能产生不同程度的影响。这说明三个级别

的CTC都能够在该任务中提供有价值的信息，验证了本文提出的三种级别CTC存在各自的尺度优势，可以帮助模型逐级学习语音和文本之间的映射，从而缓解模态鸿沟问题。同时，不难发现去掉子词级别CTC会对实验结果产生最大的影响，这是因为子词级别CTC最接近于语音识别的输出，可以更好地对应识别结果的最终目标，使得模型能够更准确地预测完整的单词，在语音识别任务中起着更为重要的作用。

方法	LibriSpeech-100h			
	Dev		Test	
	clean	other	clean	other
多尺度语音建模	8.96	21.26	9.80	21.90
-字符级别CTC约束	9.02	22.53	9.93	22.24
-音素级别CTC约束	9.68	22.27	10.21	22.29
-子词级别CTC约束	9.59	21.92	10.44	22.77

表3.多尺度语音建模消融实验（“-”表示在原始方法上进行处理）

5.2 多尺度特征融合消融实验

为了验证本文在处理特征融合的合理性，本文在LibriSpeech100小时数据集上展开实验（见表4），当不使用门控单元而将三者在维度上直接融合送入解码器时，可以看到WER在测试集中平均高出1.24，说明这种简单的融合方式不仅难以有效将不同特征进行互相补充，反而给原本特征引入了噪声，导致识别效果下降。同时，为了证明本文提出的多元化特征能够改善压缩导致的信息丢失问题，本文分别去掉字符特征、音素特征进行实验，发现当仅使用两种特征用门控方式融合时，WER在测试集上的表现也都有不同程度的上涨，这也证明了三个特征融合的合理性与必要性，融入多个特征能够帮助模型对于语义信息的理解。一个有意思的现象是，特征融合在other数据集中的表现并不如clean数据集，甚至在验证集other上融合后的特征表现反而更为劣势，分析可知，LibriSpeech other数据集比clean数据集更加复杂和多样化，包含更多的噪声和变化。当融合多尺度特征时，这些噪声和变化的影响可能被强化，导致在验证集上表现反而不如未融合的特征。相比之下，clean数据集相对更简单，噪声和变化较少，融合特征能够更好地捕捉到语音的变化，利于提高性能。

方法	LibriSpeech-100h			
	Dev		Test	
	clean	other	clean	other
多尺度特征融合	8.39	21.23	9.09	21.25
-门控单元	9.36	21.82	10.47	22.35
-字符特征	8.70	20.92	9.37	21.32
-音素特征	8.83	21.13	9.36	21.28

表4.多尺度特征融合消融实验（“-”表示在原始方法上进行处理）

5.3 基于不同建模方式预测结果实验

为了更直观体现出本文方法对于语音识别结果促进作用，本文在表5中展示了不同方法对于同一条语句的识别输出结果。在基线模型中，语音识别错误两个词语，WER为4.16，而在金字塔结构中，正如前文分析的存在过度压缩导致信息丢失的问题，从而识别出的信息也存在缺失，并且由于信息缺失进而加剧了预测错误率。在本文提出的多尺度语音建模方法中可以看到，在识别单词“word”时并没有像基线任务中错误预测成“world”，而二者在发音规则中极为接近，这说明多个尺度的文本能够帮助模型丰富语料信息，进一步理解语义，提升预测效果，而通过多尺度特征融合，更是帮助该语句预测准确率达到100%，说明融入的音素特征使得模型加强了对声学信号的理解，解决了该句基线方法中的所有预测错误。

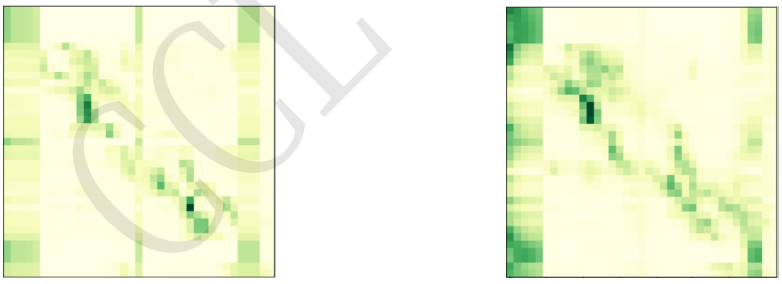
第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

原文	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriotism but this was impossible and if the phantom of his father was there in the gloom.	
方法	语音识别结果	WER
本文基线	who were a mere handful against an army should he be untrue at once to his love to country to his world(错误) should he give to his cowardice the pretext of patriot ism but this was impossible and if the fathom(错误) of his father was there in the gloom.	4.16
金字塔结构	who were a mere handful against an army should(缺失) he had and(错误) untrue at once to his love to country to his world(错误) should he give to his cowardice the pretext of(缺失) patriot ism but this was impossible and if the phantom of his father was there in the gloom.	14.6
多尺度语音建模	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriot ism but this was impossible and if the fathom(错误) of his father was there in the gloom.	2.08
多尺度特征融合	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriot ism but this was impossible and if the phantom of his father was there in the gloom.	0

表5.比较基于不同建模方式语音识别结果

5.4 多尺度建模注意力可视化分析

本文从注意力机制的角度进一步分析本文方法的有效性，通过图4比较基线模型和本文方法模型在注意力可视化方面的表现，可以发现本文提出的多尺度模型在这一方面表现更优异。在基线模型的注意力可视化结果中，颜色较浅，表示模型关注的区域较少，可能存在遗漏的信息，而多尺度模型的注意力可视化结果则更为深色，表明模型能够更充分地关注输入信号的不同部分，能够更好地捕捉语音信号中的重要特征，从而提高语音识别的性能。并且在多尺度方法图中一些较远距离单元色块颜色也更为深色，这说明多尺度方法能够更好地捕捉到长距离之间的关系，从而理解上下文信息，提高语音识别模型的性能。



(a) 基线模型　　(b) 本文方法

图4.注意力权重分布比较

6 结论

本文提出一种多尺度语音识别的建模方法，通过构建不同尺度下的文本信息，并利用CTC实现逐级对齐预测，从低层次的细粒度序列最终预测出完整文本序列，从而有效缓解了模态鸿沟问题。同时，本文还融合了不同尺度下的特征，加强了训练语料的丰富性与完整性，进一步提高了模型的推理能力。本文在LibriSpeech小规模和大规模数据以及TEDLIUM2数据集上进行了实验，结果显示本文方法相比基线系统，词错误率平均约降低1.7、0.45和0.76。后续的实验分析表明，多尺度语音建模和多尺度特征融合都促进了模型性能的提升。

在未来的工作中，我们将专注于如何在训练过程中更准确地评估不同尺度特征对模型训练的贡献，采用更加灵活的方法获取不同尺度约束比例和特征融合方法，进一步提高多尺度语音识别系统的在不同数据集上的表现，并探索多尺度建模在各种语音相关任务中的潜力。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

致谢

感谢国家自然科学基金（62276056）；国家重点研发计划项目；科技部科技创新2030—“新一代人工智能”重大项目（2020AAA0107904）；辽宁省自然科学基金（2022-KF-16-01）；云南省科技厅科技计划项目（202103AA080015）；中央高校基本科研业务费项目（N2216016、N2216001、N2216002）；111引智基地（B16009）的资助。

参考文献

Andrei Andrusenko, Rauf Nasretdinov, and Aleksei Romanenko. 2022. Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition. *CoRR*, abs/2208.07657.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

P. J. Burt and E. H. Adelson. 1987. The laplacian pyramid as a compact image code. *Readings in Computer Vision*, 31(4):671–679.

W. Chan, N. Jaitly, Q. Le, and O. Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *Eprint Arxiv*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Neural Information Processing Systems*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. 2013. Recent advances in deep learning for speech research at microsoft. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8604–8608. IEEE.

L. Dong, X. Shuang, and X. Bo. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

H. Fan, B. Xiong, K. Mangalam, Y. Li, and C. Feichtenhofer. 2021. Multiscale vision transformers.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

A. Graves, S Fernández, and F. Gomez. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *ACM*.

P. Guo, F. Boyer, X. Chang, T. Hayashi, and Y. Zhang. 2020. Recent developments on espnet toolkit boosted by conformer.

Y. Higuchi, K. Karube, T. Ogawa, and T. Kobayashi. 2021. Hierarchical conditional end-to-end asr with ctc and multi-granular subword units. *arXiv e-prints*.

Sepp Hochreiter and Jürgen Schmidhuber. 2016. Learning to forget: Continual prediction with lstm. *Neural Computation*, 28(10):2451–2471.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Jelinek and F. 1976. Continuous speech recognition by statistical methods. *Proc IEEE*, 64(4):532–556.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proceeding of the Workshop on Pattern Recognition in Practice*, pages 381–397.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Alvin M Liberman and Ignatius G Mattingly. 1985. The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Tsung Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. *IEEE Computer Society*.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Kanishka Rao, Hasim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

A. P. Rosenfeld and M. Thurston. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on Computers*.

Anthony Rousseau, Paul Deléglise, Yannick Esteve, et al. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, pages 3935–3939.

Stevens and S. S. 1936. A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5):405–416.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv*.

C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq.

Chen Xu, Xiaoqian Liu, Xiaowen Liu, Qingxuan Sun, Yuhao Zhang, Murun Yang, Qianqian Dong, Tom Ko, Mingxuan Wang, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2023a. Ctc-based non-autoregressive speech translation. *CoRR*, abs/2305.17358.

Chen Xu, Yuhao Zhang, Chengbo Jiao, Xiaoqian Liu, Chi Hu, Xin Zeng, Tong Xiao, Anxiang Ma, Huizhen Wang, and Jingbo Zhu. 2023b. Bridging the granularity gap for acoustic modeling. *CoRR*, abs/2305.17356.

S. Yoon, S. Byun, and K. Jung. 2018. Multimodal speech emotion recognition using audio and text. *IEEE*.

S. Yoon, S. Byun, S. Dey, and K. Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. *IEEE*.

J. Zhang, W. Yan, and Y. Zhang. 2022a. A new speech feature fusion method with cross gate parallel cnn for speaker recognition. *arXiv e-prints*.

Yuhao Zhang, Chen Xu, Bojie Hu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2022b. Improving end-to-end speech translation by leveraging auxiliary speech and text data. *CoRR*, abs/2212.01778.

Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu. 2018. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. *Springer, Cham*.

袁文浩, 时云龙, 胡少东, and 娄迎曦. 2021. 一种基于时频域特征融合的语音增强方法. 计算机工程, 047(010):75–81.