基于多尺度特征融合的单通道语音分离网络设计

杨海龙　胡中刚　薛特
广东开放大学
DOI:10.32629/acair.v3i4.17932

[摘　要] 将源信号从混合信号中分离出来并保持较高的语音清晰度一直是混合语音分离难点。近年来语音分离技术取得了较大进展,但是分离出来的语音往往不具有较好的清晰度。针对此问题,本文提出了一种多尺度特征融合的单通道语音分离网络。首先将时域的语音信号映射到多维特征空间,接着通过不同大小的卷积核提取不同尺度特征,最后利用通道注意力机制实现特征加权与融合。实验结果表明,该模型可以从混合语音中分离出清晰度较高的源信号,与其他方法比较,本文方法取得了更好的分离效果。
[关键词] 单通道；语音分离；卷积核
中图分类号：TN711　文献标识码：A

Design of a single-channel speech separation network based on multi-scale feature fusion

Hailong Yang　Zhonggang Hu　Te Xue
Guangdong Open University

[Abstract] Separating the source signal from the mixed signal and maintaining high speech clarity has always been a challenge in mixed speech separation. Although significant progress has been made in speech separation technology in recent years, the separated speech often lacks good speech clarity. To address this issue, this paper proposes a single-channel speech separation network based on multi-scale feature fusion. Firstly, the time-domain speech signal is mapped to a multi-dimensional feature space. Then, different scale features are extracted using convolutional kernels of varying sizes. Finally, an attention mechanism is utilized to achieve feature weighting and fusion. Experimental results show that this model can separate speech signals with high clarity from mixed speech, ultimately achieving better separation effects.
[Key words] single channel; speech separation; convolution kernel

引言

语音分离源于“鸡尾酒会问题”,即从多个说话人的混合语音中分离出源信号,目前深度神经网络在该领域得到广泛的应用。语音是典型的时序信号,前后帧存在非常强的关联性,利用循环神经网络可以捕获语音信号中的依赖关系。例如,LOU[1]等人提出双路径循环神经网络,该网络可以对语音序列进行建模,对混合语音有较好的分离效果。LI[2]等人提出基于循环神经网络的语音分离系统,该系统能够实现在线混合语音分离。但基于循环神经网络的混合语音分离方法存在一些缺点,首先循环神经网络每个时间步的输出仅依赖前一时刻的隐藏状态,无法直接获取后续或全局的上下文信息,导致全局特征的捕捉能力较为薄弱。其次,分离极长的混合语音序列时,隐藏状态会随时间步增加而逐渐丢失早期关键特征,导致分离性能下降。

基于卷积神经网络的语音分离模型则具有较强的特征提取能力。时间卷积网络是一种专为处理时序数据设计的卷积神经网络,它将多个膨胀卷积层堆叠且膨胀率按指数级递增,使得网络的总感受野随层数呈指数扩大。因此,它可以捕捉语音的长时序依赖关系且保持并行计算能力。例如,Yuan[3]等人提出多分辨率卷积神经网络,它使用各种大小的池化算子来提取多分辨率特征。LUO[4]等人提出端到端语音分离网络,该网络利用卷积块对时域语音波形直接提取特征,同时通过膨胀卷积扩大感受野捕获长距离依赖。SHI[5]等人提出端对端语音分离网络,该网络采用多个卷积块堆叠类似于金字塔结构,通过多尺度特征动态加权提高说话人分离性能。

尽管上述方法能够实现语音分离,但分离后语音的可懂度与感知质量还有进一步的提升空间。语音特征分为单元级别的特征和帧级别的特征,不同语音特征粒度能反映不同的语音细节。例如单元级别的特征能够反映语音的瞬时动态变化,帧级别的特征能够反映语音的音色。因此网络提取语音特征的能力会直接影响模型的分离性能,通过对多种特征提取机制的结构化整合,可以有效的提升网络分离性能。

受此启发,本文提出一种多尺度特征融合的单通道语音分

离网络。该网络采用编码-解码的框架，引入不同大小卷积核并通过通道注意力模块进一步提升特征提取能力，从而获得更好的分离性能。

1 模型

语音分离网络采用Conv-TasNet[4]架构，该架构主要由编码器、分离模块和解码器三个部分组成。混合语音信号首先输入编码器提取关键语音特征，随后这些特征被传递至分离模块，该模块为每个声源生成对应的掩码。最终解码器通过掩码重建出各个目标语音信号。下面将概述编码器和解码器，并详细描述分离模块。

1.1编码器/解码器。单通道语音分离是仅利用一个麦克风采集的混合语音信号，通过算法模型分离出每个目标说话人的纯净语音，如图1所示。

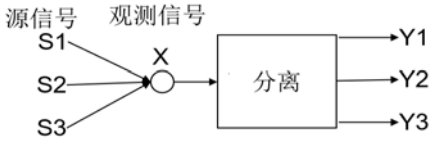

图1 单通道语音分离

原始语音信号是连续的时域波形，其直接输入分离网络存在两个问题：原始信号冗余度高、缺乏结构化特征，使得分离网络难以直接区分不同说话人。编码器通过1D卷积层直接对原始波形进行滑动窗口卷积，将一维信号转换到多维空间，公式如下：

$$\mathrm{X}=\mathrm{ReLU}(\mathrm{x}*\mathrm{W}) \tag{1}$$

其中 x 为混合信号，ReLU 为整流线性激活函数，X 为输出的映射值，W 为编码器的基函数。这样做的好处是将原始波形转换为多维特征序列，既保留了时间细节，又通过高维度增强了区分度。使不同说话人的特征在高维空间中距离最大化，从而降低分离难度。

解码器则将分离后的特征序列降维恢复成时域波形，解码器采用转置卷积模块并重构分离的语音信号：

$$\hat{X}=\mathrm{XV} \tag{2}$$

V 为转置卷积的参数，解码器的结构和功能与编码器是对称的。编码器的卷积核参数需与解码器的转置卷积参数共享，减少重构误差。

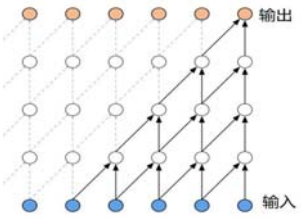

图2 时间卷积网络

1.2卷积分离模块。卷积分离模块采用的是时间卷积网络，如图2所示。该网络采用金字塔的形式，通过膨胀卷积可以在不增加核大小和网络深度的前提下，使感受野随层数呈指数级增长。这种网络结构具有较强的长时序建模能力，可直接捕捉长序列中远距离的关键特征。

为了增强网络提取多尺度时序特征的能力，如细粒度的辅音动态特征、粗粒度的元音长时特征。卷积分离模块采用不同大小的卷积核对多维空间进行卷积操作，例如将多维空间划分为3组，每组采用不同大小的卷积核提取不同粒度的特征。卷积公式如下：

$$\mathrm{Y}=\sum X\bullet W \tag{3}$$

其中，X 为输入张量，W 为卷积核。将 X 切分为 g 组，每组采用不同大小的卷积核进行特征提取，其中第 t 组的输出为：

$$Y_t=\sum x_t\bullet w_t \tag{4}$$

最后把所有组的输出结果拼接在一起，输出为：

$$\mathrm{Y}=\mathrm{Concat}(\mathrm{Y}_1,\mathrm{Y}_2,\cdots,\mathrm{Y}_t) \tag{5}$$

提取到不同尺度的语音特征后，通过通道注意力模块对每个通道的时序特征进行统计。通过学习通道权重，评估每个通道对分离语音的重要程度，增强重要特征并抑制不重要特征。如语音分离中，某通道对应说话人A的基频特征，该通道权重会被调高。通过强化重要通道、压制无关通道，输出优化后的特征，该过程如图3所示。

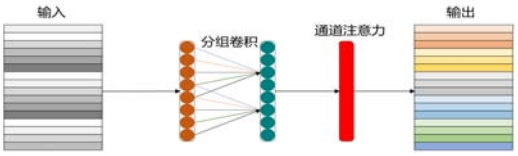

图3 分组特征提取

语音分离的常用目标有理想二值掩蔽(IBM)和理想浮值掩蔽(IRM)，它们能显著地提高分离语音的可懂度和感知质量，公式如下：

$$\mathrm{IBM}=\begin{cases}1,|S_i|>|S_{i\neq j}|\\0,\text{其他}\end{cases} \tag{6}$$

$$\mathrm{IRM}=\frac{|S_i|^2}{\sum_{j=1}^{n}|S_j|^2} \tag{7}$$

其中 S_i 为第 i 个估计源信号的值，n 为源信号总数，$\mathrm{i}=1,\cdots,\mathrm{n}$ 。得到的掩蔽矩阵值后，可估计出源语音：

$$\mathrm{d}=\mathrm{X}\bullet\mathrm{M} \tag{8}$$

最后通过解码器转换为时域波形：

$$\hat{\mathrm{X}}=\mathrm{d}\bullet\mathrm{V} \tag{9}$$

本文采用UPIT[4]损失函数：

$$\vartheta_{pit}=\frac{1}{T\times n}\sum_{i=1}^{n}\|\hat{X}_i-X_i\|^2 \tag{10}$$

T、n分别为帧数、声源个数，$\| \hat{X}_i - X_i \|^2$ 表示第 i 个估计源与干净源差值的平方值。

1.3评价指标。本实验采用尺度不变的信噪比(SI-SNR)与SDR评价分离结果。SDR公式定义如下：

$$\mathrm{SDR} = 10\log_{10}\frac{\| X \|^2}{\| \hat{X} - X \|^2} \tag{11}$$

$\hat{X}$ 为估计语音，X 为估计语音中的干净分量，$\hat{X} - X$ 为噪声分量。

SI-SNR公式定义如下：

$$\begin{cases} s_{target} = \dfrac{\langle \hat{x}, x \rangle x}{\| x \|^2} \\ e_{noise} = \hat{x} - s_{target} \\ SI - SNR = 10\log_{10}\dfrac{\| s_{target} \|^2}{\| e_{noise} \|^2} \end{cases} \tag{12}$$

其中 $\hat{x}$ 、x 分别是估计信号与原始干净信号，$\|x\|^2 = \langle x, x \rangle$ 表示信号功率。

2 实验

2.1数据集。本实验选取《华尔街日报》(Wall Street Journal,WSJ0)数据集开展相关评估，核心采用其中的WSJ0-2mix子数据集，该数据集专门用于模拟双说话人语音混叠场景。WSJ0-2mix的数据集构成如下：训练集规模为30小时，验证集为10小时；训练数据的构建方式为，从WSJ0原始数据中随机抽取49名男性说话人和51名女性说话人的语音片段，通过随机组合进行混叠处理，且混合语音的信噪比(SNR)严格控制在0-5dB区间内，呈均匀分布。测试集则基于WSJ0数据集中si_dt_05和si_et_05子集的16名未参与训练的说话人语音构建，采用与训练集一致的混叠方式，最终生成5小时的测试数据。

表1 WSJ0-2mix分离结果比较

方法	模型大小	SI-SNR(dB)	SDR(dB)
DPCL++	13.6M	10.8	-
DANet	9.1M	10.5	-
ADANet	9.1M	10.4	10.8
LSTM-TasNet	32.0M	10.8	11.2
MA-Tasnet	5.3M	11.2	12.5

2.2实验参数设置。实验所用语音数据的单段时长设定为4秒，模型训练过程共迭代100个epoch。学习率初始值为 $1e^{-3}$，同时采用自适应调整策略：若验证集准确率在连续3个epoch内未出现提升，则将当前学习率下调至原有数值的50%。优化器方面，本次训练选用 Adam 优化器以优化模型参数更新过程。

表1展示了对比方法与本文方法在数据集WSJ0-2mix上的分离性能，对比方法的分离结果摘自原文献，DPCL++与DANet的原文献中并没有给出SDR值。

本文提出的MA-Tasnet模型在SI-SNR与SDR两项核心分离指标上，相较于当前最优分离方法分别实现了0.4dB与1.3dB的性能提升。尽管该提升幅度处于小幅区间，但模型在尺寸控制上展现出显著优势，进而大幅降低了存储阶段的内存占用成本。

3 结语

本文提出了一种多尺度特征融合的单通道语音分离网络，由于不同大小的卷积核感受野不同，因此能够捕捉到不同尺度的语音特征，例如感受野相对较小的擅长捕捉语音信号中的局部细节特征，然后利用通道注意力机制为不同的通道特征分配权重。实验结果表明，相比于已有方法，本文方法有效提高了语音分离的准确性和效果。

[参考文献]

[1]LUO Y,CHEN Z,YOSHIOKA T.Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation[C]//IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2020.

[2]LI K and LUO Y.On The Design and Training Strategies for Rnn-Based Online Neural Speech Separation Systems.IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),Rhodes Island,Greece,2023,1-5.

[3]YUAN W,DONG B,WANG S,UNOKI M and WANG W.Evolving Multi-Resolution Pooling CNN for Monaural Singing Voice Separation.in IEEE/ACM Transactions on Audio, Speech, and Language Processing,vol.29,pp.807-822,2021.

[4]LUO Y,MESGARANI N.Conv-tasnet:Surpassing ideal time-frequency magnitude masking for speech separation[J].IEEE/ACM transactions on audio,speech,and language processing,2019,27(8):1256-1266.

[5]SHI Z,LIN H,LIU L,et al.Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation[E/OL].[2021-09-23].https://www.isca-speech.org/archive/pdfs/interspeech_2019/shi19b_interspeech.pdf.

作者简介：

杨海龙(1992--),男,湖南省醴陵市人,博士,研究方向为语音分离。

胡中刚(1995--),男,江苏省宿迁市人,硕士,研究方向为深度学习。

薛特(1997--),男,河南省洛阳市人,硕士,研究方向为图像识别。