文章编号:1671-4598(2023)10-0174-08 DOI:10.16526/j.cnki.11—4762/tp.2023.10.027 中图分类号:TN912 文献标识码:A

基于嵌入式注意机制的目标语音提取算法

郭志楷，杨明堃，蒋国峰，陶 祁，刘欢欢，马红强
（空军工程大学航空机务士官学校 航空电子工程系，河南 信阳 464099）

摘要：针对说话人语音提取问题，提出了一种基于深度神经网络多任务学习的嵌入式注意机制单声道说话人语音提取方法；该算法将语音分离和语音提取统一到单个框架中，向频谱映射分离模型中嵌入说话人注意机制，并在引入说话人辅助信息的注意机制中得到时变注意权重，利用时变注意权重分离出目标说话人的内部嵌入向量，随后采用提取模型对目标说话人的嵌入向量进行非线性处理运算，估计出目标说话人对应的掩蔽，进而提取出目标说话人语音；同时借助 TIMIT 数据集，进行了语音提取实验；实验结果验证了所提算法的可行性和有效性，并在说话人语音提取的性能上有明显的优越性。

关键词：深度神经网络；单声道说话人语音提取；多任务学习；嵌入式注意机制

Target Speech Extraction Algorithm Based on Embedded Attention Mechanism

GUO Zhikai, YANG Mingkun, JIANG Guofeng, TAO Qi, LIU Huanhuan, MA Hongqiang
(Aircraft Maintenance NCO Academy of Air Force Engineering University, Xinyang 464099, China)

Abstract: Aiming at the problem of speaker speech extraction, a monophonic speaker speech extraction method based on deep neural network multi-task learning embedded attention mechanism is proposed. The algorithm unifies the speech separation and speech extraction into single framework, embeds the speaker attention mechanism in the spectrum mapping separation network, obtains the time-varying attention weight in the attention mechanism of the speaker auxiliary information, utilizes the time-varying attention weight to separate the internal embedded vector of the target speaker, and then adopts the extraction model to perform nonlinear processing operations on the embedded vector of the target speaker, estimates the mask corresponding to the target speaker, and then extracts the target speaker's voice. At the same time, by means of the TIMIT dataset, the speech extraction experiments are carried out. The experimental results show the feasibility and effectiveness of the proposed algorithm, and it has obvious superiority in the performance of speaker speech extraction.

Keywords: deep neural network; monophonic speaker speech extraction; multi-task learning; embedded attention mechanism

0 引言

单声道语音分离是将说话人语音信号从混合语音中分离出来，也被称为鸡尾酒会问题[1]。人类的听觉系统可以很容易的从混合语音中分离感兴趣的源信号，但是这对于计算机识别系统来讲并不容易，尤其是在单声道情况下，提取目标语音非常困难。因而关于语音信号处理的大多数研究都集中在单声道语音分离（SCSS，single channel speech separation）[2-4]。非负矩阵分解（NMF，nonnegative matrix factorization）[5]和计算听觉场景分析（CASA，computational auditory scene analysis）[6]都是 SCSS 的常用方法。在文献［5］中，NMF 为每个源都训练一个非负基的集合，以此来进行语音分离。在文献［6］中，CASA 由语音的客观质量评估（OQAS，objective quality assessment of speech）指导，解决了语音质量与分离过程相结合的问题。但是对于多个说话人混合的语音，NMF 和 CASA 取得的分离效果有限。

近几年，深度学习技术在很多领域都得到了很好的应用。随着深度学习技术的发展，研究学者们已经提出了很多基于深度学习的语音分离方法[7-10]，SCSS 技术取得了很大的进步。基于深度神经网络（DNN，deep neural network）的语音分离通常在以下 3 种情况下应用：1）声音与乐器之间的分离；2）多个说话者的分离；3）嘈杂语音的分离。基于 DNN 的单声道语音大体可分为两种主要形式，第一种是将混合信号的特征直接通过 DNN 映射到源信号的特征[11]，第二种是将混合信号映射到各种频谱掩蔽，以解释混合信号中每个源的贡献。众多研究表明，二进制掩蔽相比较比例掩蔽分离性能低，比例掩蔽表示混合信号中源信号所占的真实能量比[12]。大多数关于混合语音的分离研究，都是针对所有源信号的分离。然而在实际情况下，例如，单个扬声器向个人移动设备发出语音查询，或者自动语音识别设备对说话人的语音识别，在这些场景下更倾向于恢复单个目标扬声器，同时降低噪声和干扰扬声器的影响，这个问题被定义为目标说话人提取[13-14]。与语音分离相比，提取目标说话者可以有效解决置换不变训练（PIT，permutation invariant training）[15]、说话者数量未知的说话人跟踪等问题。当网络仅专注于目标说话者语音提取时，

收稿日期:2023-04-24； **修回日期**:2023-06-02。

作者简介:郭志楷(1993-),男,硕士，助教。

引用格式:郭志楷,杨明堃,蒋国峰,等. 基于嵌入式注意机制的目标语音提取算法[J]. 计算机测量与控制,2023,31(10):174-181.

总体分离性能可能会更好。

大多数针对提取目标说话人的研究，都是在目标说话人语音基础上只训练一个神经网络，以此建立专门用来提取说话人的模型[16-18]。在这些提取模型的训练过程中，目标说话者和干扰者的语音都被使用，而训练的目的只是为了估计目标说话人的掩蔽，单一的网络难以充分考虑语音样本的深度特征。

Zhao 等人[19]发现频谱映射在去混响中比时频掩蔽更有效，而掩蔽在去噪和分离方面比频谱映射更好。因此构造了两个阶段的 DNN，其中第一阶段执行掩蔽去噪，第二阶段执行频谱映射去混响。受此启发，利用这两种方法的优点，可以开发一个包含频谱特征映射分离和掩蔽提取功能的框架，可在目标说话人提取过程中同时融入这两种方法的优势[20]。与单一网络相比，联合网络识别目标语音的精度更高[21]。

本文着重进行了目标说话人语音提取研究，提出了一个包含语音分离和提取相结合的注意机制模型，基于语音数据的迭代训练过程，仿真了模型训练的收敛性，利用训练好的网络模型进行了目标说话人语音提取实验，并给出部分实验的处理结果。

1　语音提取问题描述

对于单声道语音提取问题，可理解为从线性混合的单声道语音 $y(t)$ 中提取目标说话人语音 $s_0(t)$ 的过程。混合信号为：

$$y(t) = s_0(t) + \sum_{i=1}^{I} s_i(t) \tag{1}$$

式中，$s_i(t)$ 为任何数量的干扰者语音或者是噪声（在实验中考虑了干扰者）；$i = 1,2,\cdots,I$ 为干扰说话人或者是噪声的索引。

通过短时傅里叶变换（STFT，short time fourier transform）将混合信号 $y(t)$ 转化为 $\boldsymbol{Y}(t,f)$：

$$\boldsymbol{Y}(t,f) = \boldsymbol{S}_0(t,f) + \sum_{i=1}^{I} \boldsymbol{S}_i(t,f) \tag{2}$$

式中，t 和 f 分别为时间和频率索引；$\boldsymbol{Y}(t,f)$、$\boldsymbol{S}_0(t,f)$ 和 $\boldsymbol{S}_i(t,f)$ 分别为 $y(t)$、$s_0(t)$ 和 $s_i(t)$ 在时频域的表示。

在语音增强[22-23]和语音分离[24-25]的研究中表明，对 DNN 训练时，采用信号幅度谱近似（SA，signal approximation）损失收敛方法比理想比例掩蔽（IRM，ideal ratio mask）和估计的幅值谱掩蔽（SMM，spectral magnitude mask）之间的近似损失收敛方法性能更好。

提取的目标说话人语音的幅度谱 $|\hat{\boldsymbol{S}}_0(t,f)|$ 如下：

$$|\hat{\boldsymbol{S}}_0(t,f)| = \hat{\boldsymbol{M}}(t,f) \odot |\boldsymbol{Y}(t,f)| \tag{3}$$

其中：$\odot$ 为矩阵元素相乘计算方式；$\hat{\boldsymbol{M}}(t,f)$ 为目标语音对应的估计掩蔽；$|\boldsymbol{Y}(t,f)|$ 为混合信号的幅度谱。

$$\hat{\boldsymbol{S}}_0(t,f) = |\hat{\boldsymbol{S}}_0(t,f)| \times e^{j\angle Y(t,f)} \tag{4}$$

其中：$\hat{\boldsymbol{S}}_0(t,f)$ 为重构的目标语音信号频谱；$j\angle\boldsymbol{Y}(t,f)$ 为混合语音信号的相位信息。最后对重构的目标语音频谱进行短时傅里叶逆变换，即可得到目标语音的时域信号。

2　目标值

在基于 DNN 的监督语音分离系统中，语音的分离工作通常分两阶段进行，首先是模型的训练阶段，其次是测试分离阶段。我们要讲的是在训练阶段中目标的获取，目标的选取一般都是基于干净的目标语音和背景干扰得到的，合适有效的目标对于模型的学习能力和系统的分离性能起着重要的作用。目前使用的目标主要分为两类：基于时频掩蔽的目标和基于语音幅度谱估计的目标。这里简单介绍下主要的四种分离目标。

2.1　理想二值掩蔽

理想二值掩蔽（IBM，ideal binary mask）经常作为深度神经网络模型学习的目标，该目标是一个二值函数（0 或 1），该二值掩蔽的取值是根据语音信号时频谱的每个时频单元中语音能量和噪声能量的大小关系决定。首先设定一个阈值，如果一个时频单元中局部信噪比大于阈值，则对应的单元掩蔽值设为 1，反之为 0。IBM 表示为：

$$IBM(t,f) = \begin{cases} 1, & \text{若 } SNR(t,f) > LC \\ 0, & \text{其他} \end{cases} \tag{5}$$

其中：$SNR(t,f)$ 表示语音信号时频单元的局部信噪比，$IBM(t,f)$ 表示理想二值掩蔽，LC 是设置的阈值。

2.2　理想比例掩蔽

Wang 等人首先提出了理想比例掩蔽（IRM，ideal ratio mask），IRM 是一种软函数类型的目标[12]。该目标计算公式如下：

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta} = \left(\frac{SNR(t,f)}{SNR(t,f)+1}\right)^{\beta} \tag{6}$$

其中：$IRM(t,f)$ 是在时间 t 和频率为 f 的时频单元掩蔽值，$S^2(t,f)$ 和 $N^2(t,f)$ 分别表示语音能量和噪声能量，β 是一个可调节参数，而 Wang 等人已经通过实验证明，β 为 0.5 时，模型的训练结果是最好的。IRM 的值在 [0，1] 之间是连续的，这样在分离语音的时候可以提高目标语音能量谱完整性。

2.3　幅度谱掩蔽

幅度谱掩蔽（SMM，spectral magnitude mask）由目标语音和带噪语音的 STFT 谱计算得到，表示如下：

$$SMM(t,f) = \frac{|S(t,f)|}{|M(t,f)|} \tag{7}$$

$|S(t,f)|$ 和 $|M(t,f)|$ 分别表示目标语音和带噪语音幅度谱，利用两者的比值得到 SMM 目标。由于 SMM 目标用来估计目标语音的幅度谱，所以在信号的重构时需要结合带噪语音信号或目标语音信号的相位，经过 STFT 得到重构的目标语音的时域信号。

2.4　信号近似估计

信号近似估计（SA，signal approximation）的思想就是最小化目标语音和估计输出的语音幅度之间的误差，当误差逐渐收敛时，默认为此时的模型参数最优，损失函数如下：

$$SA(t,f) = (RM(t,f)|Y(t,f)| - |S(t,f)|)^2 \quad (8)$$

其中：$RM(t,f)$ 是网络模型的输出，可直接认为是估计的掩蔽，也可以通过用 SMM 目标估计 $RM(t,f)$ 来训练模型参数，然后通过上述目标函数最小化对模型参数进行微调得到最优解。

3 频谱映射分离网络

3.1 神经网络结构

DNN 是模仿人类神经系统而设计的信息分析处理结构，由神经元作为基本单元组成。一组输入经过加权进入神经单元，然后对加权后的输入进行激活计算，最后产生某种输出。其结构如图 1 所示。

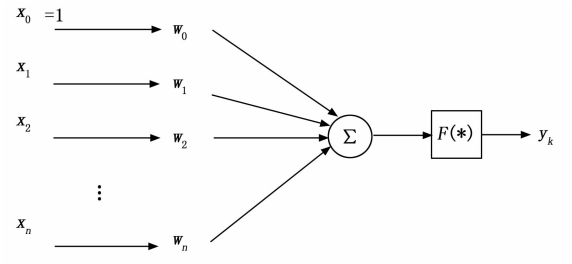

图 1 神经单元结构

基本神经单元中含有多个输入、一组权重、一个加法器、一个激活函数和一个输出，其计算原理为：

$$y_k = F\left(\sum_{i=1}^{n} w_i x_i + w_0\right) \quad (9)$$

其中：x_i 表示输入数据，w_i 表示权重和偏置（$i = 0$），F 表示激活函数，y_k 表示第 k 层神经单元的输出。

激活函数 F 有多种表达式，常用的激活函数有：线性函数、双曲正切函数（Tanh）、Sigmoid 函数、线性整流函数（ReLU，rectified linear units）。

1）线性函数：

$$F(x) = x \quad (10)$$

2）Tanh 函数：

$$F(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (11)$$

3）Sigmoid 函数：

$$F(x) = \frac{1}{1 + e^{-x}} \quad (12)$$

4）ReLU 函数：

$$F(x) = max(0,x) \quad (13)$$

除了上述四种激活函数外，还有阈值函数等。激活函数是影响神经网络功能的重要因素之一，不同的激活函数实现的功能是不一样的，例如 Tanh 函数在特征相差明显时效果会更好，ReLU 函数的稀疏性可解决网络训练时的梯度消失现象。连续平滑的 Sigmoid 函数和具有稀疏性的 ReLU 函数常用于语音分离任务中。

深度神经网络又包含三种属性层，即输入层、隐藏层、输出层，深度的大小取决于神经网络的隐藏层个数。图 2 展示了一个三层的神经网络。

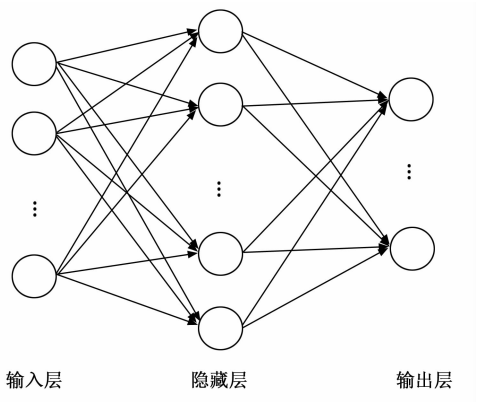

图 2 三层神经网络结构

3.2 附加掩蔽层的频谱映射网络

频谱映射分离网络主要由单个 DNN 体系结构组成，其中每个扬声器对应一个输出层，而利用谱映射分离后的两个语音幅度谱之和不等于混合语音的幅度谱，表明直接映射分离语音幅度谱是有缺陷的。因此，一个掩蔽层被添加到网络输出端，很好地解决了这个问题，其网络结构如图 3 所示。

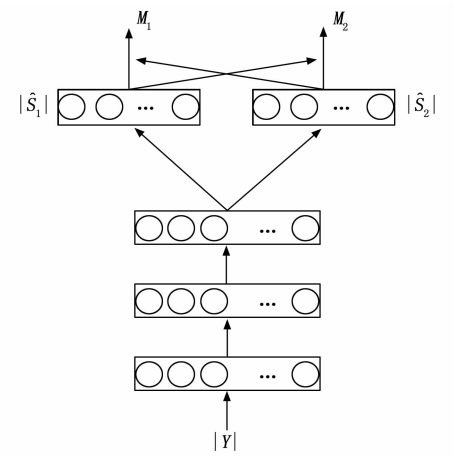

图 3 附加掩蔽层的频谱映射分离网络

式中，$|\boldsymbol{Y}|$ 为混合语音幅度特征；$|\hat{\boldsymbol{S}}_1|$ 和 $|\hat{\boldsymbol{S}}_2|$ 为该网络估计出两个说话人语音的幅度特征。

Huang 等人[26]认为两个估计的源信号 $|\hat{\boldsymbol{S}}_1|$ 和 $|\hat{\boldsymbol{S}}_2|$ 之和与 $|\boldsymbol{Y}|$ 不相等，因此将一个掩蔽层添加到网络中。$\hat{\boldsymbol{M}}_1$ 和 $\hat{\boldsymbol{M}}_2$ 分别为对应源信号的掩蔽，计算如下：

$$\hat{\boldsymbol{M}}_1 = \frac{|\hat{\boldsymbol{S}}_1|}{|\hat{\boldsymbol{S}}_1| + |\hat{\boldsymbol{S}}_2|} \quad (14)$$

$$\hat{\boldsymbol{M}}_2 = \frac{|\hat{\boldsymbol{S}}_2|}{|\hat{\boldsymbol{S}}_1| + |\hat{\boldsymbol{S}}_2|} \quad (15)$$

频谱映射分离网络将说话人选择机制包括在其分离框架中，在输出层之后进行说话人语音的选择，然而目前还不清楚这是否会提供最佳的说话人语音。因此本文将基于频谱映射的分离解释为内部分离机制的频谱映射，如图 4 所示。

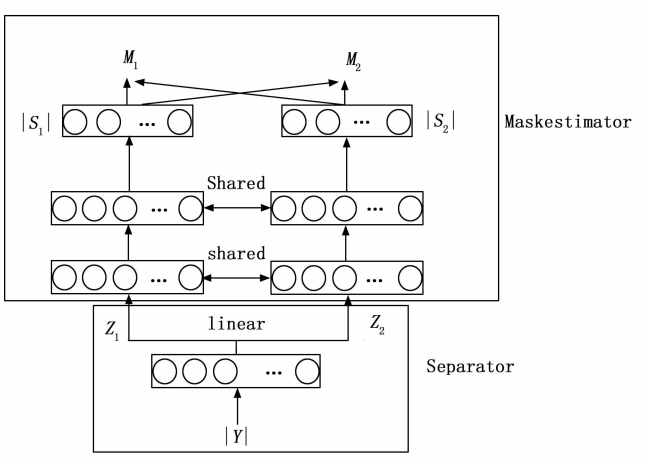

图 4　谱映射分离系统的内部分离机制

由此，可以认为分离机制存在于两个模块中，其中一个分离模块生成了对应每个源信号的内部嵌入向量 $\boldsymbol{Z}_i$，另一个掩蔽估计模块生成来自内部嵌入向量的时频掩蔽 $\boldsymbol{M}_i$，如式（16）、(17) 函数所示：

$$\{\boldsymbol{Z}\}_{i=1}^{I} = \text{Separator}(|\boldsymbol{Y}|) \tag{16}$$

$$\boldsymbol{M}_i = \text{MaskEstimator}(\boldsymbol{Z}_i) \quad (i = 1,\ldots I) \tag{17}$$

其中：Separator（·）为内部嵌入向量分离器；i 为 Separator（·）源信号对应的嵌入向量的索引；MaskEstimator（·）为基于嵌入向量的掩蔽估计器。假设 I 个源共用 MaskEstimator（·），并且其中 shared 表示参数和网络层激活函数共享，linear 是 DNN 中的线性运算。

4　多任务学习的嵌入式注意机制模型

对于人耳听力来讲，在一个多人说话的环境中只关注自己感兴趣的语音是很容易的。然而这对于人机交互的语音识别设备来说是很困难的，因此为了更好地识别感兴趣的说话人，就需要提取目标说话人的语音信息而忽略其他人声音。为了解决这个问题，本文提出的基于注意机制的多任务学习语音提取算法，它成功地提取出了目标说话人信息，同时辅助信息的利用更好地提高了说话人语音质量。

4.1　分离和提取相结合的嵌入式注意机制

本文提出的分离系统可以扩展到更多源信号混合的分离工作，为了简化说明，只考虑两个源信号混合的分离提取工作（目标语音 s_1，干扰语音 s_2）。

基于分离和提取相结合的嵌入式注意机制模型如图 5 所示，意在实现一个分离和提取双重标准下的语音处理系统。该模型由分离器、注意机制模块和掩膜估计器三部分串联而成，分离器分离出不同说话人的嵌入向量 $\{\boldsymbol{Z}_i\}$ I i =1，在注意机制模块中与说话人辅助语音谱特征相结合运算，提取出目标说话人的嵌入向量 $\boldsymbol{Z}_{\text{tar}}$，进而在掩蔽估计器中得出目标说话人对应的时频掩蔽 $\boldsymbol{M}_{\text{tar}}$。

该模型通过在分离器和掩蔽估计器之间添加说话人注意机制模块，该模块可以有针对性的选择对应目标说话人的嵌入向量，从而集成了说话人感知提取功能。下列功能函数可表示基于嵌入式注意机制的分离和提取进程：

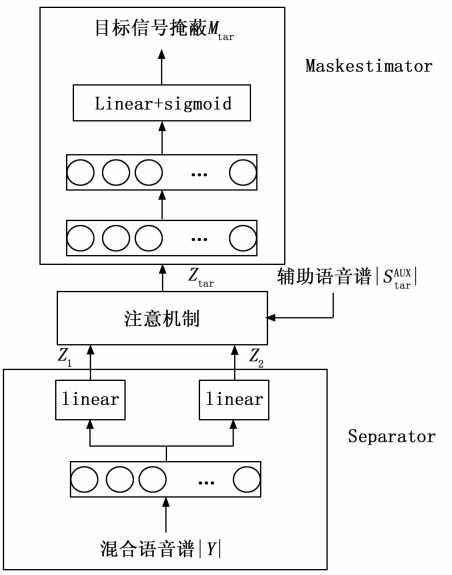

图 5　分离和提取相结合的嵌入式注意机制模型

$$\{\boldsymbol{Z}_i\}_{i=1}^{I} = Separator(|\boldsymbol{Y}|) \tag{18}$$

$$\boldsymbol{z}_{\text{tar}}^{t} = \sum_{i=1}^{I} \beta_{\text{tar},i}^{t} \boldsymbol{z}_i^{t} \quad (t = 1,\ldots,T) \tag{19}$$

$$\boldsymbol{M}_{\text{tar}} = MaskEstimator(\boldsymbol{Z}_{\text{tar}}) \tag{20}$$

其中：Separator（·）将混合语音幅度谱 $|Y|$ 转化成分离后的内部嵌入向量 $\{\boldsymbol{Z}_i\}_{i=1}^{I}$，其中 $\boldsymbol{Z}_i = \{\boldsymbol{z}_i^t\}_{t=1}^{T}$。$\{\boldsymbol{\beta}_{\text{tar},i}^t\}_{i=1}^{I}$是对应于目标说话人语音在第 t 时间帧的注意权重向量。说话人注意机制通过修改 I 个源信号中每帧内部的嵌入向量 $\{\boldsymbol{z}_{ti}\}_{i=1}^{I}$来构造目标人嵌入向量 $\boldsymbol{Z}_{\text{tar}} = \{\boldsymbol{z}_{\text{tar}}^t\}_{t=1}^{T}$。接下来 MaskEstimator（·）利用目标说话人的嵌入向量 $\boldsymbol{Z}_{\text{tar}}$估计时频掩蔽 $\boldsymbol{M}_{\text{tar}}$。

式（19）概括了注意机制的运行机理，在提出的注意机制中，目标说话人语音注意权重向量 $\{\boldsymbol{\beta}_{\text{tar},i}^t\}_{i=1}^{I}$由分离的内部嵌入向量 $\{\boldsymbol{Z}_i\}_{i=1}^{I}$和目标人的额外辅助说话人语音幅值特征 $\boldsymbol{S}_{\text{tar}}^{\text{AUX}}$经过一定的计算得到，具体的计算方式如下：

$$\boldsymbol{X}_i^{\Gamma} = \text{MEAN}(\text{MLP}^{\Gamma}(\boldsymbol{Z}_i)) \tag{21}$$

$$\boldsymbol{X}_{\text{tar}}^{\text{AUX}} = \text{MEAN}(\text{MLP}^{\text{AUX}}(\boldsymbol{S}_{\text{tar}}^{\text{AUX}})) \tag{22}$$

其中：$\boldsymbol{X}_i^{\Gamma}$ 和 $X_{\text{tar}}^{\text{AUX}}$是对应于 $\boldsymbol{Z}_i$ 和 $\boldsymbol{S}_{\text{tar}}^{\text{AUX}}$的嵌入向量，MLP（·）是多层感知器网络，MEAN（·）是以时间为轴的平均操作。

在嵌入注意机制中多层感知器的输出端使用了双曲正切函数，该函数在特征相差明显时效果会很好，循环过程中不断扩大特征效果。其计算如下：

$$e_{\text{tar},i}^{t} = w\tanh(\boldsymbol{W}^{\Gamma}\boldsymbol{X}_i^{\Gamma} + \boldsymbol{W}^{\text{AUX}}\boldsymbol{X}_{\text{tar}}^{\text{AUX}} + \boldsymbol{b}) \tag{23}$$

$$\beta_{\text{tar},i} = \frac{\gamma\exp(e_{\text{tar},1}^{t})}{\sum_{i=1}^{I}\exp(e_{\text{tar},i}^{t})} \tag{24}$$

式中，$\boldsymbol{w}$,$\boldsymbol{W}^{\Gamma}$,$\boldsymbol{W}^{\text{AUX}}$为网络可训练的权重；$\boldsymbol{b}$ 为网络模型偏置参数；γ 为设置的超参数。

4.2 多任务学习进程

本文提出的基于分离和提取相结合的注意机制框架，为确保能同时优化分离和提取功能，采用了多任务学习的目标函数 $\boldsymbol{L}_{\mathrm{MTL}}$。在网络模型训练过程中混合语音幅度谱 $|\boldsymbol{Y}|$、混合语音中的目标说话人语音幅度谱 $|\boldsymbol{S}_{tar=1}|$、干扰说话人语音幅度谱 $|\boldsymbol{S}_2|$ 和目标说话人辅助语音幅度谱 $|\boldsymbol{S}_{\mathrm{tar}}^{\mathrm{AUX}}|$ 都是可利用的。则包含分离损失和提取损失的多任务学习目标函数 L_{MTL} 为：

$$L_{\mathrm{MTL}} = \alpha L_{\mathrm{SEPA}} + (1-\alpha) L_{\mathrm{EXTR}} \tag{25}$$

$$L_{\mathrm{SEPA}} = min \frac{1}{T}\sum_{t=1}^{T} (\left\|\hat{\boldsymbol{M}}_1 \otimes |\boldsymbol{Y}| - |\boldsymbol{S}_1|\right\|^2 + \left\|\hat{\boldsymbol{M}}_2 \otimes |\boldsymbol{Y}| - |\boldsymbol{S}_2|\right\|^2) \tag{26}$$

$$L_{\mathrm{EXTR}} = min \frac{1}{T}\sum_{t=1}^{T} \left\|\boldsymbol{M}_{\mathrm{tar}} \otimes |\boldsymbol{Y}| - |\boldsymbol{S}_{\mathrm{tar}}|\right\|^2 \tag{27}$$

式中，$\hat{\boldsymbol{M}}_1$ 和 $\hat{\boldsymbol{M}}_1$ 为分离阶段估计的掩蔽；$\boldsymbol{M}_{\mathrm{tar}}$ 为目标语音提取阶段估计的掩蔽；L_{SEPA} 和 L_{EXTR} 分别为分离和提取的网络训练损失函数，且函数遵从均方误差准则；$\alpha \in [0,1]$ 为多任务训练中衡量损失函数所占比重的参数，调节该参数来增强训练的灵活度和自适应性。

5 实验及结果分析

为了验证目标语音提取算法的有效性和优越性，设计了两组实验。第一组实验证明了本算法的有效性，同时探讨了说话人性别对目标语音提取的影响。第二组实验分别使用不同的训练目标作为目标语音提取的对比试验，验证了算法的优越性。

5.1 实验数据

实验所用语音数据由 TIMIT[27] 数据库提供，分别从 TIMIT 数据库中选取两个不同性别的说话人语音片段，针对每个说话人截取了 40 秒时长的语音，前 8 秒作为测试样本，中间 16 秒作为训练样本，最后 16 秒作为辅助语音样本。然而为了研究说话人性别和语种影响，采集了两段相同时长的母语为汉语的说话人语音数据。根据采集得到的数据，利用 Matlab 软件对信号进行处理分析，将两说话人语音进行混合，混合的信噪比（SNR，signal-to-noise ratio）从 0～5 dB 均匀分布。采样频率为 16 000 Hz。

5.2 实验设置

本实验分离和提取的统一网络采用五层结构的 DNN，一个输入层，三个隐藏层和一个输出层，其每层网络的单元数为 [513 1024 1024 1024 513]。

预训练：掩蔽估计网络采用玻尔兹曼机（RBM，restricted boltzmann machine）进行预训练，训练迭代次数为 20，语音数据最小批次大小为 256（帧数），学习率为 0.003。通过 RBM 预训练，得到网络的初始权重和偏置。

实验使用 RBM 预训练方法初始化网络参数，将前一层的输出作为下一层的输入以这种数据传递方式训练 RBM 模型，其模型如图 6 所示。

RBM 是一种无方向的两层神经网络，严格意义上并不算深层网络。在图 6 中，下面一层神经元组成了可见层

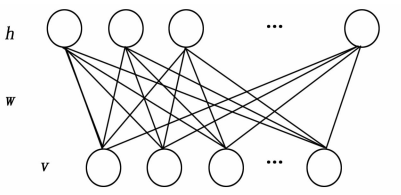

图 6 玻尔兹曼机模型

（输入层），用 v 表示可见层的神经单元值，上面一层神经元组成了隐藏层（输出层），用 h 表示隐藏层的神经单元值。可见层和隐藏层是全连接的，两层之间的权重由 w 表示。RBM 工作时，首先获取一个训练样本 v，计算隐藏层节点概率，然后在这基础上获取隐藏层激活的样本 h，计算 v 和 h 的外积作为“正梯度”。反过来从 h 中获取重构的可见层激活向量样本 v'，然后从 v' 再次获得隐藏层激活向量 h'，计算 v' 和 h' 的外积作为“负梯度”。利用正负梯度差乘上学习率更新权重 w。

精调：预训练得到初始化网络参数，在此基础上利用反向传播算法有监督的训练神经网络，使用随机梯度下降法更新权重，并且在训练过程中引入了可变动量项，训练的前十次动量项为 0.5，后续的迭代过程中动量项为 0.9 的可变化学习率，其值在区间 [0.08，0.004] 中均匀减小，自适应学习率改善了固定学习率在学习权重时精确性差的问题。精调阶段的训练次数为 180，隐藏层和输出层的激活函数分别是 ReLU 函数和 Sigmoid 函数。

$$ReLU(x) = max(0,x) \tag{28}$$

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{29}$$

Sigmoid 函数的连续光滑性质，使网络输出在一定范围内，数据在传递过程中不易发散，ReLU 函数的稀疏性可解决网络训练时的梯度消失现象。

掩蔽估计网络的目标函数为 L_{MTL}，参数 α 设为 0.5，多次试验表明 $\gamma = 2$ 时分离性能最优，其收敛曲线如图 7 所示，曲线逐渐趋于收敛，这表示网络的训练是有效的。

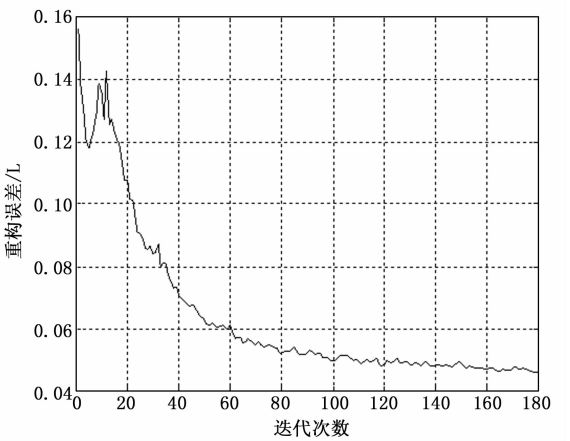

图 7 多任务学习的损失曲线

为了评估说话人语音的提取性能，实验采用了 BSS _

EVAL 工具箱中的三个评估指标：源信号失真比（SDR，source to distortion ratio）、源信号伪影比（SAR，source to artifacts ratio）、源信号干扰比（SIR，source to interference ratio）。SDR 反映了综合分离效果，SAR 反映算法对产生噪声的抑制能力，SIR 反映算法对干扰信号的抑制能力。三者数值越大就说明分离提取性能越高。

5.3　实验结果

首先对算法的有效性进行了实验评估，实验结果以波形图和语谱图的形式展示，如图 8 和图 9 所示。

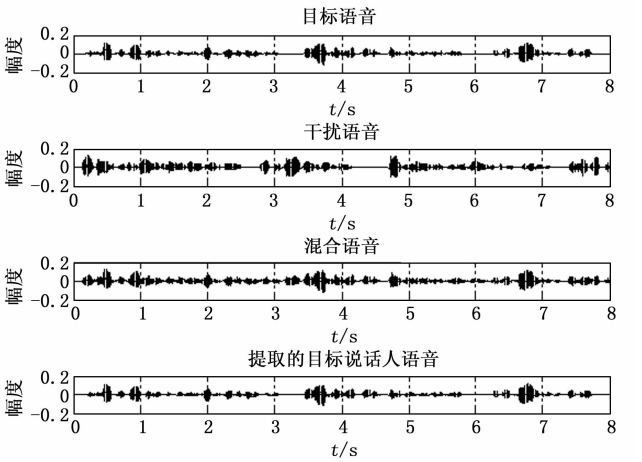

图 8　语音时域信号波形图

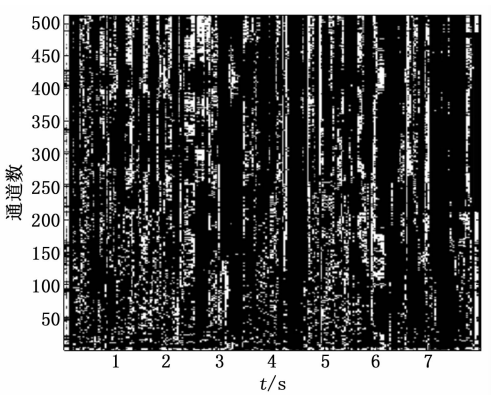

图 9　目标说话人的估计掩

图 8 分别展示目标语音、干扰语音、混合语音和算法提取的目标说话人语音的波形图。波形图的横轴表示时间，纵轴表示波形的幅值大小。通过对比提取的目标说话人波形和混合语音的波形，可以看出算法具有提取目标人语音的功能，提取的目标说话人波形与目标源语音波形的相似程度体现了算法模型对目标说话人语音提取性能的优劣。

图 9 和图 10 分别表示目标说话人的估计掩蔽插图和语谱图。掩蔽插图横坐标为时间帧，纵坐标为网络输出通道数。该掩蔽插图由掩蔽值归一化后描绘而成，其图上的白色部分是有值的，在 0～1 之间取值。黑色背景代表很小的值，接近于 0。注意下列图右上角的矩形框区域，在掩蔽插图和目标语音语谱图框内黑色占主导，说明此区域的谱值绝大多数很小或为 0，而对应的干扰语音和混合语音矩形框内具有不同颜色值，说明此区域的谱值大于 0，最终提取的目标说话人语音语谱图在相应位置也是黑色占主导，这在时频域里体现了掩蔽提取目标说话人的本质。

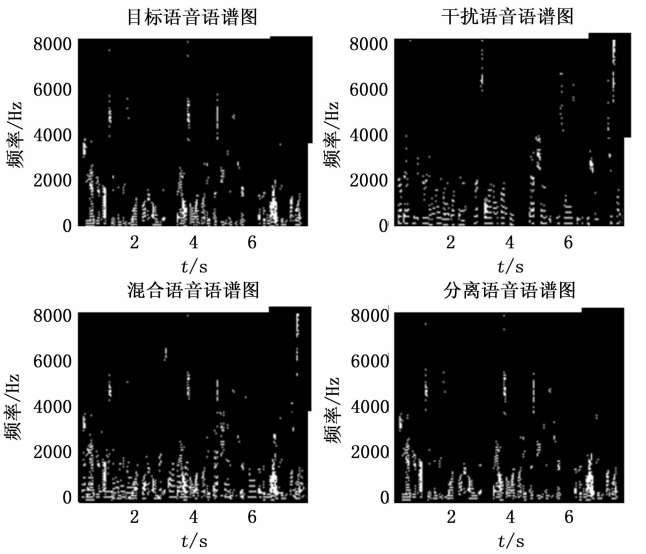

图 10　语音频域信号语谱

除了验证所提算法的有效性，同时也在相同信噪比混合情况下，探讨了说话人的性别对提取算法的影响，实验结果如表 1 所示。

表 1　混合信噪比为 0 dB 下的男 1 目标语音提取性能　dB

指标	男 1＋女 1	男 1＋女 2	男 1＋男 2
SDR	12.362 3	10.543 4	10.556 8
SAR	18.367 8	16.450 2	16.507 9
SIR	18.361 1	16.442 5	16.507 8

表 2　混合信噪比为 0 dB 下的女 1 目标语音提取性能　dB

指标	女 1＋男 1	女 1＋男 2	女 1＋女 2
SDR	12.145 2	11.197 9	11.201 6
SAR	18.246 3	17.222 5	17.830 2
SIR	18.178 2	17.202 5	17.437 9

利用这 4 个人的语音分别得到了以上 5 种组合方式，其中男 1 和女 1 为不同性别的目标说话人，选自是 TIMIT 数据集中的说话人语音。男 2 和女 2 是干扰说话人，为课题组录制的说话人语音。通过分析表 1 和表 2 指标，可以发现，相比较同性别混合语音，不同性别混合语音中的提取效果更好。在同性别混合语音中，女声的提取效果由于男声的提取效果，这可能与说话人的音质和音色有一定的关系。除了说话人性别对语音的提取有影响以外，干扰说话人语音的说话内容和语种对目标语音提取性能也有关系。同语种的混合的说话人提取效果要比不同语种混合的说话人提取效果好。这表明由同语种混合语音训练的网络模型，对

本语种语音信号的提取更有效。而对于不同语种的语音来讲，特征可能相差较大，无法在同一特征水平上进行很好的分离提取。

表 3 不同信噪比下的语音提取性能 dB

指标	0 dB	1 dB	2 dB	3 dB	4 dB	5 dB
BSDR	12.362	12.980	13.512	14.024	14.530	15.040
SAR	18.367	19.025	19.569	20.105	20.592	21.162
SIR	18.361	18.998	19.603	20.183	20.604	21.273

为了探究混合信噪比对提取性能的影响，因此在不同混合语音信噪比下进行了语音提取性能测试，由表 3 分析可知，随着干扰混合信噪比的增大，语音提取性能也不断提高。这表明在目标语音信号功率越大时，提取性能越高。

为了验证所提算法的优越性，分别使用幅度谱掩蔽（SMM，spectral magnitude mask）和信号近似估计（SA，signal approximation）目标方法进行了对比实验，结果如表 4 所示。

表 4 混合信噪比为 0 dB 下不同方法的目标语音提取性能 dB

提取方法	SDR	SAR	SIR
多任务学习嵌入注意机制	11.513 9	17.967 3	18.135 1
SA	10.636 8	16.581 4	16.581 5
SMM	8.223 5	14.126 9	14.127 1

根据表 4 的实验结果表明，相比较 SA 和 SMM 这两种方法，本文提出的基于多任务学习的嵌入式注意机制语音提取算法在 SDR 分别取得了 0.877 1 dB 和 3.290 4 dB 的提高。对于 SAR 和 SIR 指标，本文算法也均优于其它两种方法。

6 结束语

在这篇文章中，针对目标说话人语音的提取，我们提出了一种基于分离和提取多任务学习的嵌入式注意机制目标语音提取算法。本文的算法模型主要分为分离模块、嵌入式注意机制模开、语音提取模块三部分，在分离和提取的多任务优化标准下，充分利用了说话人辅助信息，更加集中地对目标说话人语音进行提取。实验结果表明，本文提出的算法利用较少的训练数据集，可实现相对较高的提取性能。

本文的不足之处在于使用的数据集单一，下一步努力方向是扩大数据集总类，保证语音信号质量的前提下，提高模型的普适性。同时可探究在其他各种噪声环境下目标说话人语音的提取性能。

参考文献：

[1] CHERRY C E. Some further experiments upon the recognition of speech, with one and with two ears [J]. The Journal of the Acoustical Society of America, 1954, 26 (4): 554.

[2] ZHU B, LI W, LI R, et al. Multi-stage non-negative matrix factorization for monaural singing voice separation [J]. IEEE Transactions on Audio Speech & Language Processing, 2013, 21 (10): 2096 - 2107.

[3] HUANG L, CHENG G F, ZHANG P Y, et al. Utterance-level permutation invariant training with latency-controlled BLSTM for single-channel multi-talker speech separation [C] //2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, 2019: 1256 - 1261.

[4] RADFAR M H, DANSEREAU R M. Single-channel speech separation using soft mask filtering [J]. IEEE Transactions on Audio Speech & Language Processing, 2007, 15 (8): 2299 - 2310.

[5] BRAIN K, CEDRIC F, PARIS S. Optimal cost function and magnitude power for NMF-based speech separation and music interpolation [C] // IEEE International Workshop on Machine Learning for Signal Processing, IEEE, 2012: 1 - 6.

[6] LI P, GUAN Y, XU B, et al. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech [J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14 (6): 2014 - 2023.

[7] TU Y H, DU J, LEE C H. A speaker-dependent approach to single-channel joint speech separation and acoustic modeling based on deep neural networks for robust recognition of multi-talker speech [J]. Journal of Signal Processing Systems for Signal, Image, and Video Technology, 2018, 90 (7): 963 - 973.

[8] SAMUI S, CHAKRABARTI I, GHOSH S K. Deep recurrent neural network based monaural speech separation using recurrent temporal restricted Boltzmann machines [C] // Interspeech 2017, 2017: 3622 - 3626.

[9] ZHAN G, HUANG Z, YING D, et al. Improvement of mask-based speech source separation using DNN [C] // International Symposium on Chinese Spoken Language Processing, Tianjin, 2016: 1 - 5.

[10] CHEN J, WANG D L. Long short-term memory for speaker generalization in supervised speech separation [J]. Journal of the Acoustical Society of America, 2017, 141 (6): 4705 - 4714.

[11] WANG D L, CHEN J. Supervised speech separation based on deep learning: an overview [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26 (10): 1702 - 1726.

[12] WANG Y, NARAYANAN A, WANG D L. On training targets for supervised speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22 (12): 1849 - 1858.

[13] DELCROIX M, ZMOLIKOVA K, KINOSHITA K, et al. Single channel target speaker extraction and recognition with speaker beam [C] // IEEE International Conference on Acoustics, Calgary, 2018: 5554 - 5558.

[14] ZMOLIKOVA K, DELCROIX M, KINOSHITA K, et al.

SpeakerBeam: speaker aware neural network for target speaker extraction in speech mixtures [J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13 (4): 800 - 814.

[15] YU D, KOLBK M, TAN Z H, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation [C] // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, 2017: 241 - 245.

[16] DU J, TU Y, XU Y, et al. Speech separation of a target speaker based on deep neural networks [C] // 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, 2014: 473 - 477.

[17] ZHANG X L, WANG D L. A deep ensemble learning method for monaural speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24 (5): 1.

[18] DU J, TU Y, DAI L R, et al. A regression approach to single-channel speech separation via high-resolution deep neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24 (8): 1424 - 1437.

[19] ZHAO Y, WANG Z Q, WANG D L. A two-stage algorithm for noisy and reverberant speech enhancement [C] // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017: 5580 - 5584.

[20] 郭志楷. 基于深度学习的目标语音信号提取算法 [D]. 南昌：南昌大学，2020.

[21] 任晨曦，王黎明，韩星程，等. 基于联合神经网络的水声目标识别方法 [J]. 舰船科学技术，2022，44 (1)：136 - 141.

[22] ERDOGAN H, HERSHEY J R, WATANABE S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks [C] //2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015: 78 - 712.

[23] DU J, LI R. An experimental study on speech enhancement based on deep neural networks [J]. IEEE Signal Processing Letters, 2014, 21 (1): 65 - 68.

[24] NARAYANAN A, WANG D L. Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23 (1): 92 - 101.

[25] XIA S, LI H, ZHANG X. Using optimal ratio mask as training target for supervised speech separation [C] //2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017: 163 - 166.

[26] HUANG P S, KIM M, HASEGAWA-JOHNSON M, et al. Deep learning for monaural speech separation [C] //2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014: 1562 - 1566.

[27] SUN D X, LI D. Analysis of acoustic-phonetic variations in fluent speech using TIMIT [C] //1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 1995: 201 - 204.

(上接第 152 页)

[7] HONGYANG L, QIEGEN L, MINGHUI Z. Gradient - based low rank method and its application in image inpainting [J]. Multimedia Tools and Applications, 2018, 77 (5): 5969 - 5993.

[8] WEIDONG D, HONGJING P. Second - order total generalized variational based on tight frame image inpainting model [J]. Computer Engineering and Applications, 2018, 54 (11): 178 - 184.

[9] 郑成松，李　琦. 基于改进优先权的对称相似图像修复算法 [J]. 信息技术与网络安全，2018，37 (12)：14 - 17.

[10] YING H, KAI L, MING Y. An improved image inpainting algorithm based on image segmentation [J]. Procedia Computer Science, 2017, 107 (1): 796 - 801.

[11] 兰小丽，刘洪星，姚寒冰. 基于纹理块与梯度特征的图像修复改进算法 [J]. 计算机工程与应用，2018，54 (20)：172 - 177.

[12] TONGDI H, ZONGXI C. A remote sensing image fusion method based on manifold learning [J]. Revista de la Facultad de Ingenieria, 2017, 32 (13): 33 - 39.

[13] HE K, NIU J, SHEN C. Image inpainting algorithm with adaptive patch using SSIM [J]. Journal of Tianjin University Science and Technology, 2018, 51 (7): 763 - 767.

[14] MENG F, YANG X, ZHOU C. A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery [J]. Sensors, 2017, 17 (9): 1 - 16.

[15] 张桂梅，陈兵兵，徐　可，等. 结合分数阶微分和图像局部信息的 CV 模型 [J]. 中国图象图形学报，2018，23 (8)：1131 - 1143.

[16] MA G. Inpainting technology in art pictures based on edge and texture feature [J]. Journal of Interdisciplinary Mathematics, 2018, 21 (5): 1199 - 1203.

[17] 郑玉婷，吴　谨. 基于蚁群算法的 Criminisi 图像修复 [J]. 红外技术，2017，39 (3)：221 - 225.

[18] CASTILLO S, CUNNINGHAM D W, WINGER C. Morphological Amoeba-based patches for exemplar-based inpainting [J]. Journal of WSCG, 2018, 26 (2): 112 - 121.

[19] HE K, GAO J, LU W. Image inpainting algorithm based on improved confidence function and matching criterion [J]. Journal of Tianjin University Science and Technology, 2017, 50 (4): 399 - 404.

[20] JANARDHANA RAO B, CHAKRAPANI Y, SRINIVAS K S. Image Inpainting Method with Improved Patch Priority and Patch Selection [J]. IETE Journal of Education, 2018, 59 (1): 26 - 34

[21] RUIFANG W. The image inpainting algorithm based on pruning samples by referring to four-domains [J]. The Imaging Science Journal, 2019, 67 (4): 179 - 186.