第 46 卷 第 1 期 计 算 机 工 程 2020 年 1 月
Vol. 46 No. 1 Computer Engineering January 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)01-0302-07 文献标志码：A 中图分类号：TP391

基于生成对抗网络的语音信号分离

刘 航，李 扬，袁浩期，王俊影

（广东工业大学 机电工程学院，广州 510006）

摘 要： 基于深度学习的单声道语音分离需要计算时频掩蔽，但现有语音分离方法中时频掩蔽不可学习，也未将其封装到深度学习中进行优化，通常依赖于维纳滤波法进行后续处理。为此，提出一种基于生成对抗网络的语音信号分离方法。在语音生成阶段引入递归推导算法和稀疏编码器来改进时频掩蔽生成结果，并将生成的语音输入至判别器进行分类，以降低信号源之间的扰动。实验结果表明，与基于深度神经网络的语音信号分离方法相比，该方法的 SDR、SIR 分离指标分别提高 6.2 dB 和 5.0 dB。

关键词： 单声道语音分离；生成对抗网络；时频掩蔽；递归推导；稀疏编码器

开放科学（资源服务）标志码（OSID）：

中文引用格式： 刘航，李扬，袁浩期，等. 基于生成对抗网络的语音信号分离[J]. 计算机工程，2020，46(1)：302-308.

英文引用格式： LIU Hang, LI Yang, YUAN Haoqi, et al. Speech signal separation based on generative adversarial networks[J]. Computer Engineering, 2020, 46(1): 302-308.

Speech Signal Separation Based on Generative Adversarial Networks

LIU Hang, LI Yang, YUAN Haoqi, WANG Junying

(School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

【Abstract】 The single-channel speech separation based on deep learning needs to calculate the time-frequency masking, which, however, cannot be learnt in the existing methods. Moreover, the time-frequency masking is not encapsulated in in-depth learning for optimization, so it relies on Wiener filtering for subsequent processing. Therefore, this paper proposes a speech signal separation method based on Generative Adversarial Networks (GAN). In the speech generation stage, the recursive derivation algorithm and sparse encoder are introduced to improve the time-frequency generation results. Then, the generated speach is eatered into the discriminator for classification, so as to reduce the disturbance between signal sources. The experimental results show that compared with other speech signal separation methods, such as the codec-based method and the recurrent neural network-based method, the SDR and SIR separation indexes of the proposed method increase by 6.2 dB and 5.0 dB respectively.

【Key words】 single-channel speech separation; Generative Adversarial Networks (GAN); time-frequency masking; recursive derivation; sparse encoder

DOI： 10.19678/j.issn.1000-3428.0053446

0 概述

语音信号分离是指在不知源信号和传输信道任何先验知识的情况下，仅根据输入语音源信号的统计特征（时域和频域信息），通过观察信号来恢复出各独立源信号的过程[1]。目前，语音信号分离主要应用于自动语音识别（Automatic Speech Recognition, ASR）、助听器设计和移动语音通信等方面。一个性能良好的前端语音分离模块可以提高自动识别的目标语音质量[2]。然而，在非平稳噪声和单声道条件下，语音分离较为困难，语音分离技术的性能仍有较大的提升空间。

已有的语音信号分离方法大致可以分为两类。一类是基于信号滤波（包括估计噪声的功率谱或理想维纳滤波器[3]）的方法，如谱减法和维纳滤波法。其中，维纳滤波法是在极小均方误差计算函数下分离纯净语音的最优滤波器。基于信号滤波的方法在假定语音与噪声相互独立的条件下，将语音信号与噪声信

基金项目： 广东省科技计划项目（2013B011304008，2013B090600031）；佛山市产学研专项资金项目（2012HC100195）。

作者简介： 刘 航（1994—），男，硕士研究生，主研方向为机器学习、语音信号处理；李 扬，教授、博士；袁浩期、王俊影，硕士研究生。

收稿日期： 2018-12-20 **修回日期：** 2019-01-26 **E-mail：** 15521331910@163.com

计算机工程 www.ecice06.com

号叠加为混合语音信号,对混合语音信号进行短时傅里叶变换(STFT),将其转化为二维的时频信号,在此基础上,推断出语音的频谱系数。在噪声为平稳或者慢变的情况下[4],这些信号滤波方法能够取得较好的分离效果。但是在现实情况中,上述条件通常难以满足,特别是在低信噪比条件下,分离方法的性能会大幅下降。另一类是基于统计建模的方法。其中,基于浅层模型的非负矩阵分解[5]方法利用混合前的纯净信号分别构建语音和噪声模型,其能挖掘非负数据中的局部表示,但是处理高维数据的能力有限,很难通过拓展上下文时间帧来挖掘语音信号中的时频相关性,且非负矩阵分解的推断过程较费时[6],严重依赖事先训练的语音和噪声模型。

基于深度学习的语音信号分离方法通常需要结合计算听觉场景分析(Computational Auditory Scene Analysis,CASA)[7]。CASA 通过模拟人耳对声音的处理机制以及人耳的听觉掩蔽功能来实现语音信号分离[8],其能够极大地提高语音的可懂度[9]。然而,以听觉掩蔽即时频掩蔽的计算作为神经网络的直接或间接计算目标,均不是最优的,如通过混合信号直接预测时频掩蔽,在语音信号与噪声信号幅度谱相加的计算中需要人为选取一个 α 值,或通过混合信号直接估算目标信号源[11-13],并根据估值计算间接得到时频掩蔽,但在这个过程中,时频掩蔽的质量严重依赖于神经网络映射的输出,且计算过程需要一个明确的时频掩蔽函数,常见的时频掩蔽计算函数是理想二值掩蔽和理想浮值掩蔽[14],其时频掩蔽的计算函数并非神经网络中的一部分(仅目标语音信号),故时频掩蔽的生成是不可学习的。

本文基于生成对抗网络(Generative Adversarial Networks,GAN)[15]提出一种可学习的时频掩蔽生成方法,在语音生成阶段引入一种递归推导算法和稀疏编码器以改进时频掩蔽的生成结果,将生成的语音信号以及真实语音信号输入至判别器进行分类,在此基础上,依据收敛的时频掩蔽获得最终的目标语音信号。

1　设计原理

本文提出一种基于 GAN 的语音信号分离算法。GAN 在计算机视觉领域得到广泛应用,其能生成逼真的图像用于语义分割与数据增强等[16-18]。本文利用时频分解技术提取时频单元级别的特征并作为输入[19],将混合信号与目标语音信号看作 2 种不同概率分布的信号并进行非线性映射。基于 GAN 的语音信号分离算法结构如图 1 所示。

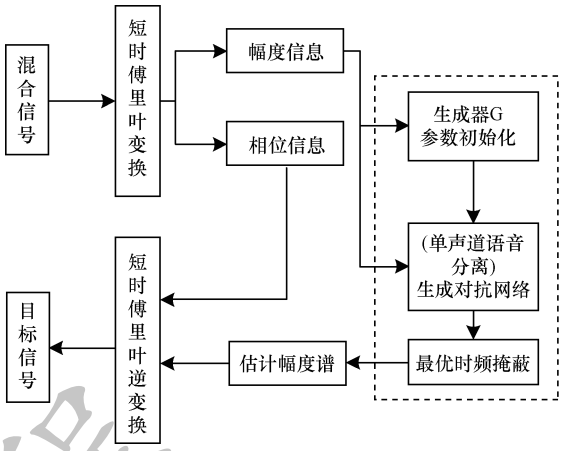


图 1　基于 GAN 的语音信号分离算法结构

Fig. 1　Structure of speech signal separation algorithm based on GAN

1.1　时频掩蔽计算

语音信号的频谱 $\hat{\boldsymbol{Y}}^j \in \mathbb{R}^{M\times N}$ 计算如下:

$$\hat{\boldsymbol{Y}}^j = \boldsymbol{Y} \otimes \boldsymbol{M}^j \tag{1}$$

其中,$\otimes$ 表示矩阵元素积,$\boldsymbol{M}^j \in \mathbb{R}^{M\times N}$ 为时频掩蔽值,其表征一个混合信号中不同信号源(如语音分离中的目标说话人和干扰说话人)之间的相互关系,时频掩蔽值 $\boldsymbol{M}^j$ 的估计采用 α-功率幅度谱的维纳滤波法[20],具体如下:

$$\boldsymbol{M}^j = \frac{|\hat{\boldsymbol{Y}}^j|^{\alpha}}{\sum_j |\hat{\boldsymbol{Y}}^j|^{\alpha}} \tag{2}$$

其中,$|\cdot|$ 表示矩阵的绝对值,α 是根据假定语音的概率分布选取的指数。

1.2　生成对抗网络原理

受博弈论中二元零和博弈的启发,GAN 模型中包含一对相互对抗的模型:生成模型 G 和判别模型 D。判别器 D 尽可能正确地判断输入的数据是来自真实样本还是来自伪样本,生成器 G 则尽量去学习真实数据样本的数据分布。GAN 模型结构如图 2 所示。

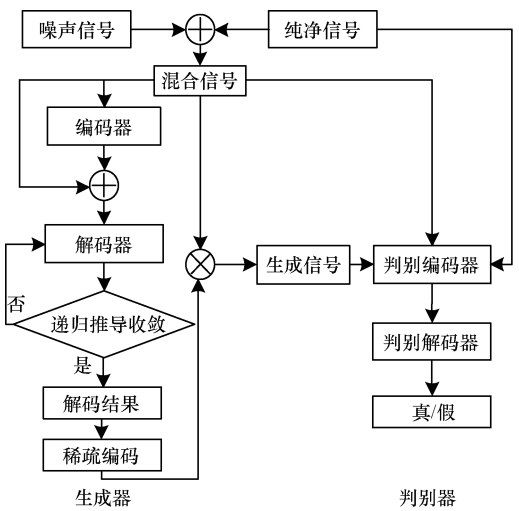


图 2　GAN 模型结构

Fig. 2　Structure of GAN model

www.ecse06.com

为在博弈中胜出,判别器的判别过程与生成器的学习过程相互对抗并迭代优化,使得判别器和生成器的性能不断提升,优化的目标是寻找两者间的纳什均衡。在上述极小-极大的优化过程中,生成器和判别器的优化目标函数为:

$$\min_G \max_D V_{\mathrm{CGAN}}(G,D)=E_{x\sim P_{\mathrm{data}(x)}}[\lg D(x)]+E_{z\sim P_{G(z)}}[\lg(1-D(G(z)))] \tag{3}$$

其中,x 是服从 $P_{\mathrm{data}(x)}$ 的真实数据样本,$G(z)$ 代表 $P_{G(z)}$ 分布的数据样本,在本文中其具体表示真实语音信号和由生成器生成的语音信号,V_{CGAN} 表示待优化的目标函数。

假定有一个额外的约束条件向量 $\boldsymbol{y}$ 作为辅助信息,生成器 $G(z,\boldsymbol{y})$ 在 $\boldsymbol{y}$ 的约束下生成语音信号,判别器 $D(x,\boldsymbol{y})$ 也在 $\boldsymbol{y}$ 的约束下对真假语音信号进行判别[21],则目标函数转换为:

$$\min_G \max_D V_{\mathrm{CGAN}}(G,D)=E_{x,\boldsymbol{y}\sim P_{\mathrm{data}(x,\boldsymbol{y})}}[\lg D(x,\boldsymbol{y})]+E_{z\sim P_{G(z)},\boldsymbol{y}\sim P_{\mathrm{data}(\boldsymbol{y})}}[\lg(1-D(G(z,\boldsymbol{y}),\boldsymbol{y}))] \tag{4}$$

2 算法设计

2.1 生成对抗网络设计

GAN 模型结构由生成器 G 和判别器 D 2 个部分组成。本文提出一种可学习的时频掩蔽生成器,该生成器引入了一个具有神经网络结构的递归推导算法和一个用于生成时频掩蔽 $\boldsymbol{M}^j$ 的稀疏编码层[22-23]。其中,生成器由多层递归神经网络(RNN)和稀疏编码层组成,RNN 输出到稀疏编码层,稀疏编码层的输出为相应的时频掩蔽 $\boldsymbol{M}^j$。该方法省去了后续的信号滤波等处理过程,也无需人工定义神经网络的层数。

如图 2 所示,生成器将一层双向递归神经网络作为编码器(R_{enc})、一层递归神经网络作为解码器(R_{dec})及一层前馈神经网络作为稀疏编码层 F_{FFN}。稀疏编码层的输出(即时频掩蔽 $\boldsymbol{M}^j$)再与混合信号进行矩阵元素相乘,即得到目标语音信号。判别器由一层前馈神经网络(作为编码器)和另外一层前馈神经网络(作为解码器)组成,并输出[0,1]区间的值。生成器与判别器经过迭代优化,得到最优时频掩蔽 $\boldsymbol{M}^j$,并以此估算目标语音信号的幅度谱,然后结合混合信号的相位谱以短时傅里叶逆变换(ISTFT)重建时域信号。

2.2 输入语音处理

以 44.1 kHz 为采样频率进行采样,以 0 dB 为信噪比,设混合得到的语音时域信号为 Z,Z_t 通过短时傅里叶变换,将一维的时域信号转化为二维的时频信号 $\boldsymbol{Y}\in\mathbb{R}^{M\times N}$。短时傅里叶变换按照交叠分段的方法进行分帧加窗,窗函数采用汉明窗,帧长设定为 23 ms,帧移为 6 ms,即每帧包括 $N=1\ 024$ 个采样点,相邻时间帧之间存在 256 个采样点的重叠部分。转换后,按时间段 T 将时频信号 $\boldsymbol{Y}$ 分割为子频带群 B,剩余帧数以 0 值填充使时间维度扩充至 T。为保持语音片段衔接处的相关性,后一帧子频带与前一帧子频带重叠 $L\times 2$ 的时间段。将 $\boldsymbol{Y}$ 中的每一份子频带 b 的幅度谱 $|\boldsymbol{Y}_{\mathrm{in}}|\in\mathbb{R}^{T\times N}$ 作为生成器的输入,但考虑到声音的高频部分能量小且相对人耳听觉不敏感,本文在训练阶段忽略信号中大于频率 F 的高频部分,以 $|\boldsymbol{Y}_{\mathrm{filter}}|\in\mathbb{R}^{T'\times F}$ 作为输入,以此减少训练参数量并保留语音的最主要信息。

2.3 可学习时频掩蔽生成器

在经过输入语音处理后,以 $|\boldsymbol{Y}_{\mathrm{filter}}|$ 作为编码器 R_{enc} 的输入,R_{enc} 采用双向 RNN(Bi-GRU),每一时间帧的输出 h_t 随时间帧 t 的迭代而更新,并采用残差网络进行叠加,叠加过程为[24]:

$$\boldsymbol{h}_{\mathrm{enc}_t}=\boldsymbol{h}_t+|\boldsymbol{y}_{\mathrm{filter}_t}|$$
$$|\boldsymbol{Y}_{\mathrm{filter}}|=[|\boldsymbol{y}_{\mathrm{filter}_T}|,|\boldsymbol{y}_{\mathrm{filter}_{T-1}}|,\cdots,|\boldsymbol{y}_{\mathrm{filter}_t}|,\cdots,|\boldsymbol{y}_{\mathrm{filter}_1}|],|\boldsymbol{y}_{\mathrm{filter}_t}|\in\mathbb{R}^F \tag{5}$$

其中,$\boldsymbol{h}_{\mathrm{enc}_t}$ 表示每一时间帧 t 的输出 $\boldsymbol{h}_t$ 叠加 $|\boldsymbol{y}_{\mathrm{filter}_t}|$ 后的幅度谱向量,残差网络有利于加快训练速度[24]。

合并时间段 T 中每一时间帧 $t\in T$ 的 $\boldsymbol{h}_{\mathrm{enc}_t}$ 为 $\boldsymbol{H}_{\mathrm{enc}}\in\mathbb{R}^{T\times(2\times F)}$,并减去 2.2 节中所述重叠时间段 $L\times 2$(L表示子频带重叠的时间段),得到输出 $\tilde{\boldsymbol{H}}_{\mathrm{enc}}\in\mathbb{R}^{T'\times(2\times F)}$,其中,$T'=T-(L\times 2)$。具体为:

$$\tilde{\boldsymbol{H}}_{\mathrm{enc}}=[\boldsymbol{h}_{\mathrm{enc}_{1+L}},\boldsymbol{h}_{\mathrm{enc}_{2+L}},\boldsymbol{h}_{\mathrm{enc}_{T-L}}] \tag{6}$$

本文引入的递归推导算法通过编码器 R_{dec} 不断递归生成临时变量 $\boldsymbol{H}^j_{\mathrm{dec}}$,直到满足如下的收敛判别准则:生成相邻的临时变量 $\boldsymbol{H}^j_{\mathrm{dec}}$ 的均方误差 L_{MSE}[22]。设最大迭代次数为 N_{iter},阈值为 τ_{term},f^j_{dec} 表示解码器 R_{dec} 的训练函数,递归推导算法伪代码如下:

算法 1 递归推导算法

$S^j_0\leftarrow f^j_{\mathrm{dec}}(\tilde{H}_{\mathrm{enc}})$
for $i\in\{1,2,\cdots,N_{\mathrm{iter}}\}$ do
$H^j_{\mathrm{dec}}\leftarrow f^j_{\mathrm{dec}}(S^j_{i-1})$
if $L_{\mathrm{MSE}}(S^j_{i-1},H^j_{\mathrm{dec}})<\tau_{\mathrm{term}}$ then
break
$S^j_{i-1}\leftarrow H^j_{\mathrm{dec}}$
Return H^j_{dec}

在 $\boldsymbol{H}^j_{\mathrm{dec}}$ 解码收敛后,传递至生成时频掩蔽 $\boldsymbol{M}^j$ 的稀疏编码层,每一时间段 T 共享稀疏编码层的权重参数,如下:

$$\tilde{\boldsymbol{M}}^j=R_{\mathrm{ReLU}}(\boldsymbol{H}^j_{\mathrm{dec}}\boldsymbol{W}_{\mathrm{mask}}+b_{\mathrm{mask}}) \tag{7}$$

其中,修正线性单元(ReLU)函数的定义如下:

$$R_{\mathrm{ReLU}}(x)=\begin{cases}x, & x>0\\ 0, & x\leqslant 0\end{cases} \tag{8}$$

在式(7)、式(8)中,R_{ReLU} 为分段线性函数,其将

计算机工程 www.ecice06.com

所有的负值设为 0,而正值不变,这种设置称为单侧抑制,能够使神经元具有稀疏激活性[18]。稀疏化处理是为了改善干扰抑制,同时恢复频率维度至目标语音信号的频率维度 N。$\boldsymbol{W}_{\text{mask}} \in \mathbb{R}^{(2\times F)\times N}$ 是前馈神经网络的权重系数矩阵,$b_{\text{mask}} \in \mathbb{R}^{N}$ 为相应的偏差。

通过前文定义的编码器和解码器,得到目标语音信号的幅度谱 $|\hat{\boldsymbol{Y}}^{j}_{\text{filter}}| \in \mathbb{R}^{T'\times N}$ 如下:

$$\begin{cases} |\hat{\boldsymbol{Y}}^{j}_{\text{filter}}| = |\boldsymbol{Y}_{\text{filter}}| \otimes \tilde{\boldsymbol{M}}^{j} \\ |\boldsymbol{Y}_{\text{filter}}| = [\,|\boldsymbol{y}_{\text{in}_L}|, |\boldsymbol{y}_{\text{in}_{L-1}}|, \cdots, |\boldsymbol{y}_{\text{in}_{T-L}}|\,] \end{cases} \tag{9}$$

其中,$|\boldsymbol{Y}_{\text{filter}}|$ 为生成器的真实输入。

2.4　判别器

生成器生成的时频掩蔽包含来自噪声信号的扰动,判别器通过对真假语音信号进行判定,使得生成的信号 $|\hat{\boldsymbol{Y}}^{j}_{\text{filter}}|$ 不断地逼近目标语音信号,从而起到降噪的作用。判别器由前馈神经网络 F_{enc} 和 F_{dec} 的编解码组成,输入分为 2 种,一种是生成器生成的语音信号 $\hat{\boldsymbol{Y}}^{j}_{\text{filter}}$ 和混合信号 $\boldsymbol{Y}_{\text{in}}$,另一种是真实语音信号 $\boldsymbol{Y}^{j}$ 和混合信号 $\boldsymbol{Y}_{\text{in}}$,输入合并为 $\boldsymbol{Y}^{j}_{\text{concat}}$,$F_{\text{enc}}$ 和 F_{dec} 通过时间段 T 共享权重参数。判别器的输出为:

$$S = R_{\text{ReLU}}(R_{\text{ReLU}}(\boldsymbol{Y}^{j}_{\text{concat}}\boldsymbol{W}_{\text{enc}} + b_{\text{enc}})\boldsymbol{W}_{\text{dec}} + b_{\text{dec}}) \tag{10}$$

其中,$\boldsymbol{W}_{\text{enc}} \in \mathbb{R}^{2N\times(N/2)}$、$\boldsymbol{W}_{\text{dec}} \in \mathbb{R}^{(N/2)\times 1}$ 分别表示前馈神经网络 F_{enc}、F_{dec} 的权重系数矩阵,$b_{\text{enc}} \in \mathbb{R}^{N/2}$、$b_{\text{dec}} \in \mathbb{R}^{1}$ 为相应的偏差,S 的结果为 0、1。

2.5　训练目标

根据生成器与判别器的输入,本文调整目标函数为:

$$\min_{G}\max_{D} V_{\text{CGAN}}(G,D) = E[\lg D(\boldsymbol{Y}^{j}\boldsymbol{Y}_{\text{in}})] + E[\lg D(1 - D(G(\boldsymbol{Y}_{\text{in}}),\boldsymbol{Y}_{\text{in}}))] \tag{11}$$

其中,$\boldsymbol{Y}^{j}$ 是真实语音信号,$\boldsymbol{Y}_{\text{in}}$ 为输入混合信号,G 为生成语音信号。

根据式(2)可知,判别器的输入除原语音信号 $\boldsymbol{Y}^{j}$ 以及生成器生成的信号 $\hat{\boldsymbol{Y}}^{j}_{\text{filter}}$ 之外,还有额外的混合信号 $\boldsymbol{Z}_t$ 通过短时傅里叶转化得到的时频信号 $\boldsymbol{Y}_{\text{in}}$,且 $\boldsymbol{Y}_{\text{in}}$ 约束生成器的生成方向。GAN 网络使生成语音信号不仅逼近了目标语音信号的概率分布,还学习了此环境下音频信号的频谱结构。

3　实验结果与分析

3.1　实验设计

本文实验环境为 Ubuntu 14.04,使用 Python2.7 以及 Tensorflow1.1 进行编程,并采用 MIR-1K 作为数据集[25],输入设置如 2.2 节所描述,每一段语音的采样频率为 44.1 kHz 并维持在4 s~13 s范围内。这些语音片段是 110 首分别由男声和女声演唱的歌曲,本文随机选择 80% 的语音片段用于训练,10% 用于验证,10% 用于测试。设生成器输入的语音序列初始长度 $T=50$,实验验证大约为 0.5 s,重叠时间段 $L=10$,所有参数使用 Adam 算法作为梯度下降方法[26],学习率为 $1\times e^{-4}$,批量训练数据 $B=16$,整个训练集的 10% 用作交叉验证,最大迭代次数为 10 000。

为评估系统的信号分离性能,根据 BSS-EVAL 度量标准[27],本文测量了 3 个定量值:信伪比(Signal to Artifact Ratio, SAR)、信干比(Signal to Interferece Ratio, SIR)以及信失比(Singal to Distortion Ratio,SDR),这 3 个指标的核心思想是将预测的信号 $\hat{\boldsymbol{Y}}^{j}_{\text{filter}}$ 分解成目标语音信号 $\boldsymbol{S}_{\text{target}}$、干扰信号(即噪声)$\boldsymbol{S}_{\text{interf}}$ 和剩余信号 $\boldsymbol{S}_{\text{artif}}$ 3 个部分,如下:

$$\hat{\boldsymbol{Y}}^{j} = \boldsymbol{S}_{\text{target}} + \boldsymbol{S}_{\text{interf}} + \boldsymbol{S}_{\text{artif}} \tag{12}$$

根据式(12)对预测信号的分解,定义如下 3 个度量标准:

$$\begin{cases} S_{\text{SDR}} = 10\lg\dfrac{\|\boldsymbol{S}_{\text{target}}\|^2}{\|\boldsymbol{S}_{\text{interf}} + \boldsymbol{S}_{\text{artif}}\|^2} \\ S_{\text{SIR}} = 10\lg\dfrac{\|\boldsymbol{S}_{\text{target}}\|^2}{\|\boldsymbol{S}_{\text{interf}}\|^2} \\ S_{\text{SAR}} = 10\lg\dfrac{\|\boldsymbol{S}_{\text{target}} + \boldsymbol{S}_{\text{interf}}\|^2}{\|\boldsymbol{S}_{\text{artif}}\|^2} \end{cases} \tag{13}$$

SIR、SAR 和 SDR 分别反映分离算法对干扰信号的抑制能力、分离算法对引入噪声的抑制能力及分离算法的综合分离效果。三者的数值越大,信号的分离效果越好[27]。

3.2　性能对比分析

本文分别对利用深度神经网络(DNN)来预测目标语音信号对应的浮值掩蔽方法[10](记为方法 1)、基于编解码器结构预测目标语音信号方法[13](记为方法 2)、基于递归神经网络(RNN)并引入递归推导算法的时频掩蔽生成器(记为方法 3)以及本文加入判别器的生成对抗网络(记为方法 4)进行实验验证,其中,输入特征均保持一致且仅采用幅度谱作为特征。首先,对方法 1 进行实验分析,其首层隐藏单元个数为输入的维度,中间隐藏层单元个数为 512,神经网络层数分别采用 2 层、3 层和 4 层,当神经网络层数为 3 时方法 1 取得较优效果,而层数达到 4 时,SDR、SIR、SAR 都为下降的趋势,说明过多的神经元增加了方法 1 的网络复杂度,导致了过拟合现象;在方法 3 中,递归推导算法在最大迭代次数 $N_{\text{iter}}=3$、阈值 $\tau_{\text{term}}=1\times e^{-2}$ 和最大迭代次数 $N_{\text{iter}}=10$、阈值 $\tau_{\text{term}}=1\times e^{-3}$ 时(分别记为方法 3-1、方法 3-2),SIR 和 SDR 均取得最高值。各方法针对人声与背景音的 SDR、SIR 和 SAR 对比结果如表 1、表 2 所示。

计算机工程 www.ecice06.com

表 1 各方法针对人声的 SDR、SIR 和 SAR 结果

Table 1 SDR, SIR and SAR results of each method for human voice

方法	SDR/dB	SIR/dB	SAR/dB
方法 1	6.03	11.52	9.65
方法 2	6.54	12.13	8.41
方法 3-1	9.74	14.66	9.98
方法 3-2	10.58	13.84	9.22
方法 4	12.28	16.53	18.88

表 2 各方法针对背景音的 SDR、SIR 和 SAR 结果

Table 2 SDR, SIR and SAR results of each method for background sound

方法	SDR/dB	SIR/dB	SAR/dB
方法 1	5.96	13.18	9.12
方法 2	6.82	12.19	8.30
方法 3-1	9.36	15.28	9.44
方法 3-2	10.95	13.29	9.63
方法 4	11.35	18.27	16.33

从表 1、表 2 可以看出，方法 1 与方法 2 在人声分离效果上相差不大，方法 1 的 SAR 值要高于方法 2，说明以时频掩蔽为计算目标的分离在噪声抑制上效果较优。方法 3 相比方法 1 和方法 2 在 SIR 和 SDR 上有显著提升，可验证语音信号在时域和频域上具有很强的相关性，且根据时序进行建模能体现语音信号中时序的相关性。在方法 3 的基础上引入判别器进行分类，即为方法 4。将生成器生成的目标语音信号与真实语音信号在真实带噪环境下进行判别分类，可以看出，方法 4 的 SIR 和 SDR 高于方法 3-1、方法 3-2，且在 SAR 上取得了较好效果，说明生成对抗网络对引入噪声的抑制能力较强。

针对方法 3（人声），分析不同的输入时间步长 T 对分离效果的影响，输入步长分别取 $T=50$、$T=100$ 和 $T=500$，其余设置保持不变，结果如表 3 所示。从表 3 可以看出，当 $T=100$ 时，SIR、SDR 及 SAR 均较高，此时实际实验时间为 5 s～6 s 左右，与实验中输入的语音片段更接近，说明时间步长需要与输入的语音片段长度保持一致，过大的时间步长会提高时间复杂度。

表 3 方法 3 中不同时间步长对应的指标结果

Table 3 SDR, SIR and SAR results corresponding to different time steps in method 3

输入时间步长	SDR/dB	SIR/dB	SAR/dB
50	5.21	10.73	6.10
100	9.01	13.28	8.35
500	7.15	11.41	7.22

将通过方法 4 分离的目标语音信号与原纯净语音信号同时转化为时域波形，图 3 为真实人声与分离人声时域波形图，图 4 为真实背景音与分离背景音的时域波形图。从图 3、图 4 可以看出，人声的分布较密集集中，而背景音分布较均匀且变换波动大，分离语音包络线与原语音包络线形状逼近，且人声与背景音分离的平均误差仅为 0.14 dB。

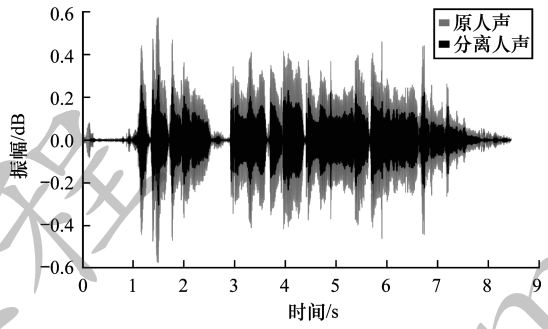


图 3 人声分离时域波形图

Fig. 3 Time domain waveform of human voice separation

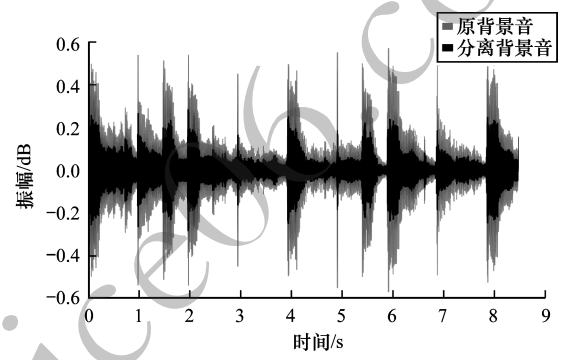


图 4 背景音分离时域波形图

Fig. 4 Time domain waveform of background sound separation

将时域波形图转化为二维时频功率谱并进行对比分析，结果如图 5～图 9 所示。从图 5、图 6 可以看出，真实人声比分离人声略多出细微毛刺，形状极为相似，且图 6 结果可验证在训练阶段忽略大于频率 F 的高频部分以减少训练参数量完全可行，原因是高频部分信号的能量值小（近乎为 0）且人耳听觉对其不敏感。

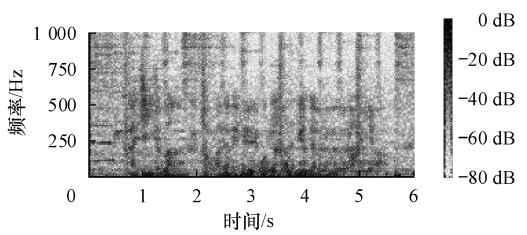


图 5 混合语音信号二维时频功率谱

Fig. 5 Two dimensional time-frequency power spectrum of mixed speech signal

www.ecice06.com

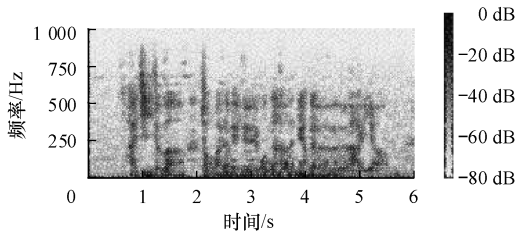


图 6　真实人声二维时频功率谱

Fig. 6　Two dimensional time-frequency power spectrum of real human voice

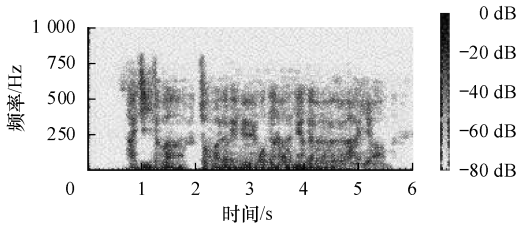


图 7　真实分离人声二维时频功率谱

Fig. 7　Two dimensional time-frequency power spectrum of real separated human voice

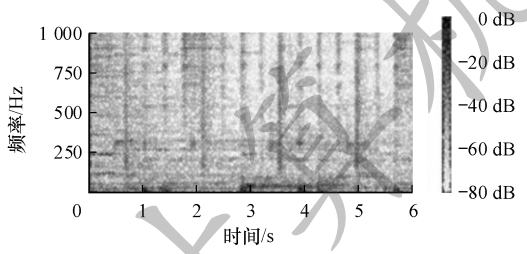


图 8　背景音二维时频功率谱

Fig. 8　Two dimensional time-frequency power spectrum of background sound

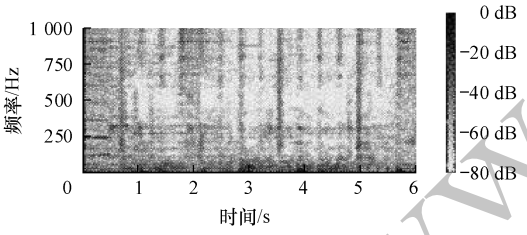


图 9　分离背景音二维时频功率谱

Fig. 9　Two dimensional time-frequency power spectrum of separated background sound

4　结束语

本文提出一种应用于单声道语音信号分离的生成对抗网络深度学习方法。将时频掩蔽与目标语音信号的幅度谱融入到神经网络中,作为生成器中的学习目标,并将生成的语音信号输入至判别器进行分类,以降低信号源之间的扰动。实验结果表明,相对传统非负矩阵分解以及深度神经网络等方法,该方法具有更好的信号分离性能。在实际应用中,测试环境与训练环境匹配率较低,因此,下一步将利用半监督与聚类相结合的方式来解决信号源不对称等问题。

参考文献

[1] WOODRUFF J. Integrating monaural and binaural cues for sound localization and segregation in reverberant environments[EB/OL]. [2018-12-02]. http://pdfs.semanticscholar.org/6eda/6a32fa17c982dd0e28e66e6bb44530833fab.pdf.

[2] LIU Jing. Speech signal recognition based on blind source separation and noise suppression[J]. Computer Measurement and Control, 2018, 26(12): 140-144. (in Chinese)
刘晶. 基于盲源分离和噪声抑制的语音信号识别[J]. 计算机测量与控制, 2018, 26(12): 140-144.

[3] WANG Fangjie, JIN Yun. Speech enhancement based on wiener filteringin digital hearing aids[J]. Chinese Journal of Electron Devices, 2017, 40(4): 1021-1025. (in Chinese)
王方杰, 金赟. 基于维纳滤波的数字助听器语音增强算法[J]. 电子器件, 2017, 40(4): 1021-1025.

[4] LIANG Shan, LIU Wenju, JIANG Wei. A new Bayesian method incorporating with local correlation for IBM estimation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(3): 476-487.

[5] MOHAMMADIHA N, SMARAGDIS P, LEIJON A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(10): 2140-2151.

[6] JIANG Maosong, WANG Dongxia, NIU Fanglin, et al. Speech enhancement method based on sparsity-regularized non-negative matrix factorization[J]. Journal of Computer Applications, 2018, 38(4): 1176-1180. (in Chinese)
蒋茂松, 王冬霞, 牛芳琳, 等. 稀疏正则非负矩阵分解的语音增强算法[J]. 计算机应用, 2018, 38(4): 1176-1180.

[7] WANG D L, BROWN G. Computational auditory scene analysis: principles, algorithms, and applications [M]. Washington D.C., USA: IEEE Press, 2006.

[8] QU Junling, LI Hongyan. Research on speech separation based on computational auditory scene analysis[J]. Application Research of Computers, 2014, 31(12): 3822-3824. (in Chinese)
屈俊玲, 李鸿燕. 基于计算听觉场景分析的混合语音信号分离算法研究[J]. 计算机应用研究, 2014, 31(12): 3822-3824.

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

计算机工程 www.ecice06.com

[10] GRAIS E M, SEN M U. Deep neural networks for single channel source separation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2014: 3734-3738.

[11] UHLICH S, GIRON F. Deep neural network based instrument extraction from music[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2015: 2135-2139.

[12] LIANG Yao, ZHU Jie, MA Zhixian. A monaural speech separation algorithm based on deep neural networks[J]. Information Technology, 2018, 42(7): 24-27. (in Chinese)
梁尧, 朱杰, 马志贤. 基于深度神经网络的单通道语音分离算法[J]. 信息技术, 2018, 42(7): 24-27.

[13] CHANDNA P, MIRON M, JANER J, et al. Monoaural audio source separation using deep convolutional neural networks[C]//Proceedings of the 13th International Conference on Latent Variable Analysis and Signal Separation. Washington D. C., USA: IEEE Press, 2017: 258-266.

[14] XIA Shasha, ZHANG Xueliang, LIANG Shan. Supervised speech separation using optimal ratio mask[J]. Acta Automatica Sinica, 2018, 44(10): 1876-1887. (in Chinese)
夏莎莎, 张学良, 梁山. 基于优化浮值掩蔽的监督性语音分离[J]. 自动化学报, 2018, 44(10): 1876-1887.

[15] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2014: 2672-2680.

[16] CAO Yangjie, JIA Lili, CHEN Yongxia, et al. Review of computer vision based on generative adversarial networks[J]. Journal of Image and Graphics, 2018, 23(10): 1433-1449. (in Chinese)
曹仰杰, 贾丽丽, 陈永霞, 等. 生成式对抗网络及其计算机视觉应用研究综述[J]. 中国图象图形学报, 2018, 23(10): 1433-1449.

[17] DONAHUE C, LI B, PRABHAVALKAR R. Exploring speech enhancement with generative adversarial networks for robust speech recognition[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 5024-5028.

[18] YU Liang, WU Haijun, JIANG Weikang. Multi-channel speech enhancement based on beamforming and GAN network[J]. Noise and Vibration Control, 2018, 38(S2): 591-596. (in Chinese)
余亮, 吴海军, 蒋伟康. 结合波束形成和 GAN 网络的多通道语音增强研究[J]. 噪声与振动控制, 2018, 38(S2): 591-596.

[19] HUANG Jianjun, ZHANG Xiongwei, ZHANG Yafei. Single channel speech enhancement via time-frequency dictionary learning[J]. Acta Acustica, 2012, 37(5): 539-547. (in Chinese)
黄建军, 张雄伟, 张亚非. 时频字典学习的单通道语音增强算法[J]. 声学学报, 2012, 37(5): 539-547.

[20] ODUGU K, RAO B M S S. New speech enhancement using Gamma tone filters and perceptual wiener filtering based on sub banding[C]//Proceedings of International Conference on Signal Processing and Communication. Washington D. C., USA: IEEE Press, 2014: 236-241.

[21] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 5967-5976.

[22] HUANG Gao, SUN Yu, LIU Zhuang, et al. Deep networks with stochastic depth[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 646-661.

[23] VARDAN P, ROMANO Y, ELAD M. Convolutional neural networks analyzed via convolutional sparse coding[J]. Journal of Machine Learning Research, 2016, 18: 2887-2938.

[24] HE Kaiming, ZHANG Xiangru, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[25] HSU C L, JANG J S R. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(2): 310-319.

[26] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2018-12-02]. https://arxiv.org/pdf/1412.6980.pdf.

[27] VINCENT E G. Performance measurement in blind audio source separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1462-1469.

编辑 吴云芳

计算机工程 www.ecice06.com