浙江理工大学学报，2022，47(3)：405-412
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.03.016

# 基于高分辨率网络和自注意力机制的歌声分离算法

倪　欣,任　佳
(浙江理工大学机械与自动控制学院,杭州 310018)

**摘　要**：针对现有歌声分离算法分离精度不高的问题,提出了一种基于高分辨率网络和自注意力机制的歌声分离算法。该算法构建了基于频域模型的深度神经网络,将高分辨率网络作为主干网络,以此保证分离精度,并在网络中融入自注意力机制来捕获歌曲中的重复旋律。在歌声分离算法中,首先通过短时傅里叶变换对音乐信号进行时频转换,得到幅值谱;其次通过构建的神经网络将歌曲幅值谱进行分离,得到人声和伴奏的幅值谱;最后结合原歌曲的相位谱,通过短时傅里叶逆变换得到人声和伴奏的时域信号。结果表明:在MUSDB18数据集上,分离得到的人声和伴奏信号偏差比指标分别为7.68 dB和12.85 dB,相比于基准模型分别提高了21.52%和1.26%。该算法可以增强神经网络特征表达能力,有效提升歌声分离效果。

**关键词**：歌声分离;高分辨率网络;自注意力机制;深度神经网络;频域模型
**中图分类号**：TN912.3　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 05-0405-08

# Singing voice separation algorithm based on high resolution network and self-attention mechanism

*NI Xin*, *REN Jia*
(Faculty of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou 310018)

**Abstract**: To address the problem of low separation accuracy of the existing singing voice separation algorithms, a singing voice separation algorithm based on high-resolution network and self-attention mechanism was proposed, which constructed a deep neural network based on the frequency-domain model, used high-resolution network as the backbone network to ensure the separation accuracy, and integrated the self-attention mechanism into the network to capture the repeated melody in the song. The process of singing voice separation algorithm is as follows: Firstly, the short-time Fourier transform was used for the time-frequency transformation of music signal to get the amplitude spectrogram; second, the amplitude spectrum of song was separated by the established neural network to obtain the amplitude spectrogram of the singing voice and accompaniment; finally, the time domain signals of singing voice and accompaniment were obtained by short-time inverse Fourier transform according to the phase spectrogram of the original song. The experimental results show that: on the MUSDB18 dataset, the signal-to-deviation ratio index of singing voice and accompaniment is 7.68 db and 12.85 db respectively, an increase of 21.52% and 1.26% than the benchmark model, indicating that the algorithm proposed in this study can strengthen the feature expression ability of neural network, and effectively improve the effect of singing voice separation.

**Key words**: singing voice separation; high-resolution network; self-attention mechanism; deep neural network; frequency-domain model

收稿日期:2021－07－06　　网络出版日期:2021－12－17
基金项目:浙江省公益技术研究项目(LGG20F030007)
作者简介:倪　欣(1993－　),男,江苏常州人,硕士研究生,主要从事智能信号处理方面的研究。
通信作者:任　佳,E-mail:jren@zstu.edu.cn

## 0 引 言

随着多媒体技术的快速发展,音乐数量大量增加,人们对音乐信号处理技术的需求日益增加。歌声分离是指将人声和伴奏从歌曲中分离,是歌词识别、歌手识别和旋律提取[1]等任务的重要预处理环节,具有重要的应用价值。

传统歌声分离算法,如非负矩阵分解法[2]、鲁棒主成分析法[3]、反复结构提取法[4]等,都是针对特定类型的音乐,根据先验知识来设计算法。然而,实际生活中的音乐风格各异、种类繁多,所以这些算法的泛化性能不佳,难以满足当下需求。近年来,随着深度学习在图像、自然语言和语音等领域的广泛应用,歌声分离领域也涌现出一系列基于数据驱动且分离效果优于传统分离算法的深度学习算法[5]。这些算法主要可以分为基于时域模型[6-8]的算法和基于频域模型[9-11]的算法两大类。基于时域模型的算法采用端到端的训练方式,即直接将时域混合信号输入神经网络,网络输出分离后的人声和伴奏的时域信号,如 Wave-U-Net[6] 和 Demucs[8];而基于频域模型的算法则是先将时域信号进行短时傅里叶变换(Short-time Fourier transform, STFT),得到幅值谱图,然后将其输入神经网络,得到目标信号的幅值谱图,最后通过短时傅里叶逆变换(Inverse short-time Fourier transform, ISTFT)得到时域目标信号。

目前,歌声分离算法大多采用了图像分割领域常用的 U 型神经网络结构。Hennequin 等[12]将 U-Net 模型引入到歌声分离任务中,将音乐信号的幅值谱视作图像,先通过多次下采样来降低分辨率,提取深层特征,再通过多次上采样来提高分辨率,实现多尺度特征融合。Takahashi 等[10]将 U-Net 中的双层卷积模块替换成密集卷积模块来增强网络学习能力。Park 等[13]将多个 U 型网络串行连接,形成堆叠沙漏网络,以此增强网络多尺度特征融合能力。上述采用 U 型网络的算法在多次下采样过程中会产生细节信息丢失的问题,而多次上采样则会影响网络输出的精度。Sun 等[14]在人体姿态估计任务中提出了高分辨率网络(High resolution network, HRNet),该网络将多个不同分辨率的子网分支并行连接,同时子网络间特征互相交换融合,这使得网络可以输出高精度的特征图,因此 HRNet 可以很好地解决上述 U 型网络存在的问题,提升歌声分离效果。

目前,大多数基于深度神经网络的歌声分离算法只是简单地将图像领域模型迁移至歌声分离任务,并未将音乐信号的特征和网络进行有机融合。众所周知,音乐中通常会出现一些重复的旋律,例如打击乐中常出现有规律的击鼓声。张天琪等[4]针对该特征提出了基于多反复结构模型的歌声分离算法,结果表明该特征可以有效提升算法的分离效果。因此,本文将用于捕获音乐中重复旋律的网络结构融入主干网络,以此增强网络特征学习能力。然而,由于音乐中的重复旋律通常间隔时间不确定且较长,因此其在时间维度上呈现长期依赖关系,例如每隔 4 s 以上贝斯可能才会弹奏同一个旋律。普通的卷积神经网络由于卷积核大小的限制,导致其对局部特征更为敏感,而难以发现长期依赖关系。因此,本文采用自注意力机制对歌曲中的重复旋律建模。自注意力机制[15]通过计算一条序列中任意两个位置数据的特征相似度得到相似度矩阵,无论序列中两个数据位置间隔有多远,它们的特征相似度都是直接计算得到的,路径长度始终为 1,因此自注意力机制可以很好地捕获长期依赖关系。

现有基于 U 型网络结构的歌声分离算法由于采用串行的上下采样结构,导致高分辨率特征图预测不精确,难以满足高品质歌声分离的需求。同时,现有歌声分离算法只简单借鉴了图像领域的网络模型,未将音乐信号独有的特征和神经网络进行有机融合。因此,本文提出了一种基于高分辨率网络和自注意力机制(High resolution network with self-attention mechanism, AHRNet)的歌声分离算法。该算法使用适用于歌声分离的高分辨率网络代替传统 U 型网络,利用其高精度特性,保证分离后人声和伴奏信号的质量。考虑到音乐中通常存在重复旋律这一特性,在高分辨率网络中融入自注意力机制,用以捕捉重复旋律的依赖关系,以进一步提升歌声分离质量。

## 1 AHRNet 歌声分离算法

### 1.1 算法整体框架

本文提出的歌声分离算法分别对人声和伴奏建立模型,主要分为训练和测试两个阶段。人声信号分离算法的整体框架如图 1 所示,伴奏信号与之类似。在训练阶段,输入时域形式的混合了人声和伴奏的歌曲信号以及纯净的人声信号,使用 STFT 进行时频转换,得到其幅值谱和相位谱,由于人耳对相位变化的感知不敏感,所以训练阶段算法只使用幅值谱。歌曲的幅值谱经过本文提出的网络模型

AHRNet 后得到掩膜，将掩膜和歌曲幅值谱点乘，得到预测的人声幅值谱，以预测幅值谱和真实幅值谱均方误差最小为目标来训练模型。在测试阶段，将混合歌曲的幅值谱输入训练好的 AHRNet 模型中，得到预测幅值谱，再结合混合歌曲的相位谱，通过 ISTFT 得到预测的时域人声信号。

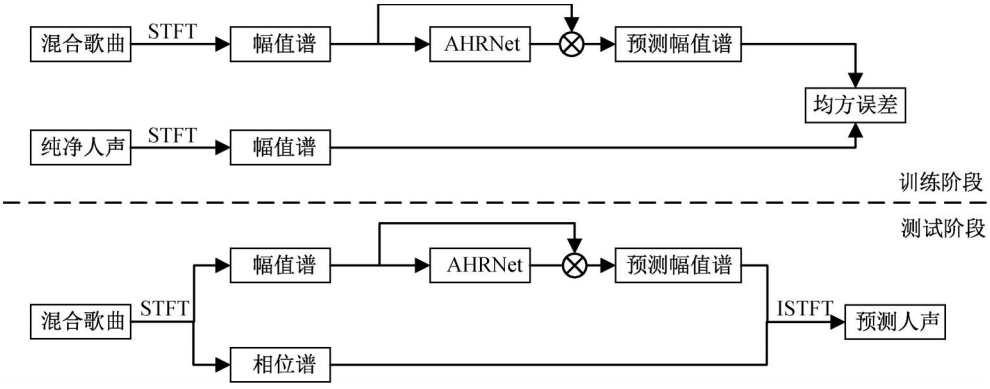



**图 1　人声分离算法整体框架**

## 1.2　高分辨率网络 HRNet

本文将 Sun 等[14]提出的高分辨率网络加以改进应用于歌声分离中，其结构如图 2 所示。歌曲经过 STFT 后得到的幅值谱图作为网络输入，本文中其特征图尺寸为 2×512×256，其中 2 为通道数，表示歌曲为双声道，512 为频率轴频带数，256 为时间轴帧数。

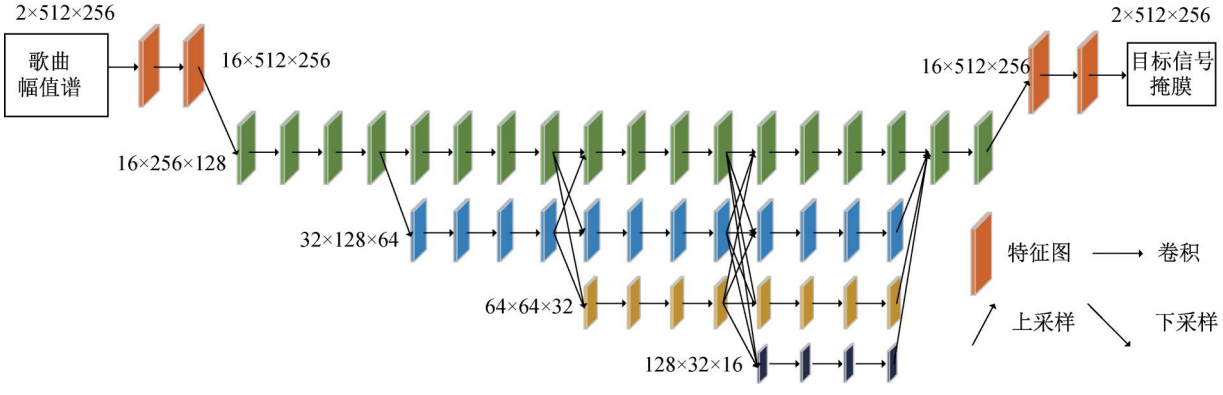


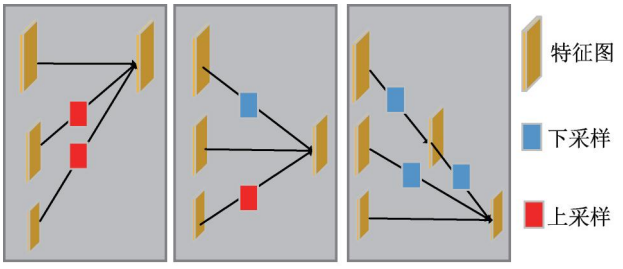

**图 2　高分辨率网络 HRNet 结构**

本文提出的 HRNet 网络主要由三部分组成：

a)输入层：通过两次使用 3×3 的卷积将通道数由 2 增加到 16，同时保持分辨率不变。之后为减小计算量和显存占用，下采样至原尺寸的一半，特征图尺寸变为 16×256×128。与原始 HRNet 相比减少一次下采样过程，细节信息得以更好地保留。

b)多尺度融合层：构建 4 条并行的子网分支，每条分支保持分辨率不变，每经过 3 次 Bottleneck 卷积[14]运算后就向下采样构建新分支，新分支的特征图分辨率尺寸减半，通道数翻倍。如图 2 所示，本文 4 条分支的特征图分别以不同颜色进行表示，尺寸分别为 16×256×128、32×128×64、64×64×32 和 128×32×16，每条分支的特征图尺寸始终保持不变，保证不同分辨率特征图都具有高精度。每次构建新子网分支时，将不同分支的不同分辨率特征图融合，使得输出的每个分辨率特征图都包含了输入的所有分辨率特征图的信息，从而形成多尺度特征融合。图 3 展示了 3 种不同输出分辨率特征图融合的过程。假设有 $s$ 个不同分辨率的输入特征图：$\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_s\}$，$s$ 个不同分辨率的输出特征图 $\{\boldsymbol{N}_1,\boldsymbol{N}_2,\cdots,\boldsymbol{N}_s\}$。每个分辨率的输出特征图都是融合所有分辨率的输入特征图，即 $\boldsymbol{N}_k=\sum_{i=1}^{s}f(\boldsymbol{M}_i,k),k\in[1,s]$。其中 $f(M_i,k)$ 表示将分辨率为 $i$ 的输入特征图映射为分辨率为 $k$ 的输出特征图，映射方式包括恒等映射、上采样和下采样。当输入和输出特征图分辨率相等时，使用恒等映射，即输出等于输入；当前者分辨率小于后者时，先用 1×1 的卷积降低通道数，再使用双线性插值法上采样提高分辨率；当前者分辨率大于后者时，使用步长为 2 的 3×3 的卷积进行下采样，使分辨率降为输入的一半。

c)输出层：将上一步得到的特征图上采样恢复至原尺寸，再使用两次卷积将通道数降为 2，得到的目标信号掩膜和输入的歌曲幅值谱尺寸相同。




(a)分辨率最大 (b)分辨率中等 (c)分辨率最小

**图 3 不同分辨率下的输出特征图**

歌曲幅值谱经过 HRNet 后输出目标信号掩膜,再将两者点乘得到预测的目标信号幅值谱,通过优化预测幅值谱和真实幅值谱间的误差来训练网络。损失函数采用均方误差函数,其定义为:

$$L = \| \boldsymbol{Y} - \boldsymbol{X} \odot \boldsymbol{M} \|_2 \tag{1}$$

其中:$\boldsymbol{X}$ 为歌曲幅值谱,$\boldsymbol{M}$ 为预测的目标信号掩膜,$\boldsymbol{Y}$ 为真实的目标信号幅值谱。

### 1.3 自注意力模块

本文提出的自注意力模块结构如图 4 所示。记输入特征图尺寸为 $C \times F \times T$,其中:$C$ 表示通道数,$F$ 表示频率,$T$ 代表时间。首先将特征图尺寸转化为 $(C \times F) \times T$,然后通过三个全连接矩阵映射得到查询矩阵(Query, $Q$)、键矩阵(Key, $K$)和值矩阵(Value, $V$),最后根据这三个矩阵得到输出矩阵。计算过程可用式(2)表示:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = V \cdot \mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{D}}\right) \tag{2}$$

其中:$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbb{R}^{D \times T}$;$\boldsymbol{Q}^{\mathrm{T}}$ 表示矩阵转置;$D$ 为键矩阵中每个列向量的维度,起调节作用,防止内积过大。最后将式(2)的结果通过全连接矩阵恢复至与输入特征图相同的尺寸。

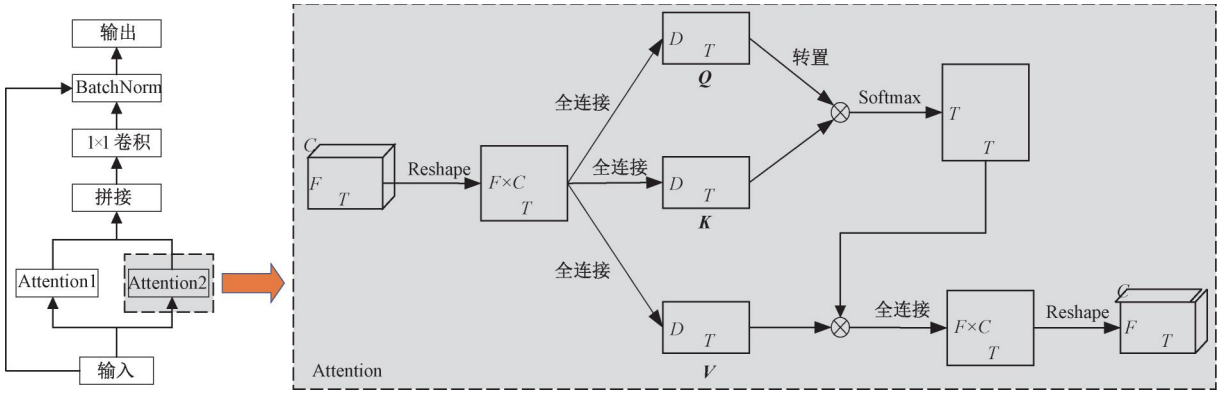



**图 4 自注意力模块结构**

自注意力机制的本质在于模型通过计算向量间相似度,过滤不重要信息,更多地关注关联性高的重要信息。本文通过计算查询矩阵 $Q$ 和键矩阵 $K$ 列向量的点积为其相似度"打分",得到相应的权重值,再经过 Softmax 函数归一化后与值矩阵 $V$ 相乘,得到最后的注意力矩阵。由于查询矩阵、键矩阵和值矩阵三者都是通过同一个输入矩阵映射得到,因此称之为自注意力机制。当歌曲中出现重复旋律,则相应位置处的权重值较大,网络将更关注此处的特征信息。

为了表示更丰富的特征信息,本文参照 Vaswani 等[15]提出的算法,引入多头自注意力机制。首先,并行构建多个 Attention,将其输出进行拼接,得到表达能力更丰富的特征。然后,利用 1×1 卷积将输出特征图尺寸和原始输入保持一致。最后,将两者相加,经过批归一化(BatchNorm)得到最终输出特征图。

在 1.2 节所述高分辨率网络 HRNet 基础上,本文将自注意力模块插入在不同分支特征融合之后,得到 AHRNet 网络,其结构如图 5 所示。HRNet 的每条分支主要由 Bottleneck 卷积模块[14]和特征融合模块交替组成,将自注意力模块插入在特征融合模块之后,可以使网络捕获不同分辨率特征图中重复旋律的特征信息。然后通过 Bottleneck 卷积模块提取特征图中的深层信息。

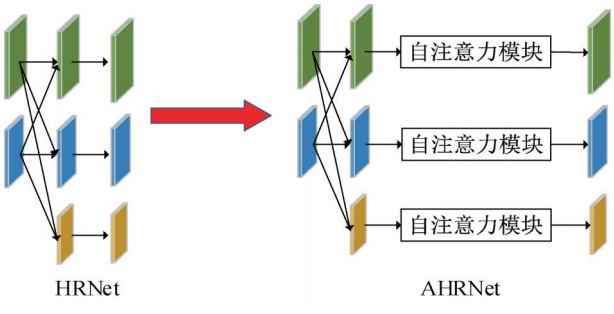



**图 5 融合自注意力模块的 AHRNet 结构**

## 2　实验结果与分析

### 2.1　实验设置及数据集

本文采用 MUSDB18 数据集进行实验，该数据集是 2018 年 SiSEC 竞赛音乐分离子任务的标准数据集，训练集和测试集分别有 100 首和 50 首歌曲。其中每首歌曲均为双通道立体声，采样率为 44100 Hz，包含人声、鼓、贝斯和其他伴奏四个音轨。由于本文实验仅关注歌声分离，因此将鼓、贝斯和其他伴奏合并成一条伴奏音轨。为加快模型训练速度，减少模型占用显存，预先将所有歌曲降采样至 16000 Hz。短时傅里叶变换选择汉宁窗函数，帧长为 1024，帧移为 512。采用 Uhlich 等[17]提出的数据增强方法，即随机交换立体声通道、混合不同歌曲的人声和伴奏，增强后数据集扩充至原来的 64 倍。

本文实验所用计算机配置如下：CPU 为 Intel（R） Core（TM） i7-4790 K，GPU 为 NVIDIA 1080Ti，内存为 16 GiB，系统为 Windows10 64 位操作系统。采用 Python3.7 和 Pytorch1.6.0 编程。训练时损失函数为均方误差，优化函数为 Adam，Batch size 为 8，学习率为 0.001。

### 2.2　评价指标

歌声分离的客观评价标准一般使用盲源分离工具包[18]来评估，其原理是将源信号 $s(t)$ 的估计信号 $\hat{s}(t)$ 分解为四部分：

$$\hat{s}(t)=s_{\text{target}}(t)+e_{\text{interf}}(t)+e_{\text{noise}}(t)+e_{\text{artif}}(t) \tag{3}$$

其中：$s_{\text{target}}(t)$ 表示估计信号中属于源信号的部分，$e_{\text{interf}}(t)$ 表示由其他信号源所引起的估计误差，$e_{\text{noise}}(t)$ 表示观测信号中包含的噪声干扰误差，$e_{\text{artif}}(t)$ 表示由于算法本身产生的系统噪声误差。

由于大多数情况下可以忽略噪声的影响，所以 $e_{\text{noise}}(t)$ 可忽略。本文采用的客观评价指标为信号偏差比 $R_{\text{SD}}$、信号干扰比 $R_{\text{SI}}$ 和系统误差比 $R_{\text{SA}}$，定义分别为：

$$R_{\text{SD}}=10\lg\left(\frac{\|s_{\text{target}}(t)\|^2}{\|e_{\text{interf}}(t)+e_{\text{artif}}(t)\|^2}\right) \tag{4}$$

$$R_{\text{SI}}=10\lg\left(\frac{\|s_{\text{target}}(t)\|^2}{\|e_{\text{interf}}(t)\|^2}\right) \tag{5}$$

$$R_{\text{SA}}=10\lg\left(\frac{\|s_{\text{target}}(t)+e_{\text{interf}}(t)\|^2}{\|e_{\text{artif}}(t)\|^2}\right) \tag{6}$$

其中：$R_{\text{SD}}$ 表示算法的整体分离效果，$R_{\text{SI}}$ 表示算法的分离纯净度，$R_{\text{SA}}$ 表示算法的鲁棒性。三个指标的值越高，表示算法分离效果越好。

本文采用 SiSEC 竞赛提供的 museval 工具包[19]来计算三个评价指标，和 SiSEC 竞赛评判标准一样，使用 50 首测试歌曲的指标中位数作为最后测试结果。

### 2.3　实验结果

为直观展示不同算法的分离效果，将本文算法和 UMX[9]、TAK1[11]分离的人声和伴奏幅值谱图进行可视化对比。其中：UMX 和 TAK1 分别为 SiSEC 竞赛音乐分离任务的基准模型和最优模型，TAK1 采用了 U 型网络结构，用来对比验证高分辨率网络相对于 U 型网络的优越性。从测试数据集中选择歌曲 Punkdisco-Oral Hygiene 进行歌声分离，图 6 展示了 2:00－2:05 时间段内（时长 5 s）纯净人声和伴奏的幅值谱图，以及不同算法分离得到的幅值谱图。观察纯净人声的幅值谱图可以发现，第 2 秒到第 4 秒内无人声只有伴奏，而 UMX 出现了明显的伴奏信号，本文提出的 AHRNet 则较好地将伴奏信号去除了。此外，通过比较第 4 秒到第 5 秒的人声幅值谱图可以发现，相比于其他算法，采用 AHRNet 分离得到的人声幅值谱图与纯净人声幅值谱图差距最小，细节保留程度最好，说明其在分离时未引入较大的失真，保证了分离后的人声质量。观察伴奏的幅值谱图可以发现，不同算法间的差距不如人声明显，但第 1 秒到第 2 秒 2000 Hz 到 4000 Hz 处可以明显发现，AHRNet 的细节保留程度最好，分离效果最优。

表 1 为本文所提算法与其他分离算法在测试数据集上的客观评价指标比较结果。本文提出的 AHRNet 在除伴奏 $R_{\text{SI}}$ 外的其他指标上都取得了最高值，人声和伴奏的 $R_{\text{SD}}$ 值相比于基准模型 UMX 分别提升了 21.52% 和 1.26%，说明本文算法的分离性能较好。伴奏分离方面，和上文幅值谱图对比结果一致，四种算法的分离效果相差不大，$R_{\text{SD}}$ 值均大于 12 dB，较好地完成了伴奏分离任务。而在人声分离方面，HRNet 相比于 U 型网络 TAK1 分离效果提高了 9.39%，说明高分辨率网络相比于 U 型网络分离更加精准，可以更好地保留细节信息。将自注意力机制融入 HRNet 后，六个指标值均得到了提升，其中 AHRNet 的 $R_{\text{SD}}$ 值达到了 7.68 dB，相比于 HRNet 提升了 6.37%，说明自注意力机制利用音乐旋律在时间维度上的相关性增强了网络的特征表达能力，进一步提升了网络的歌声分离能力。

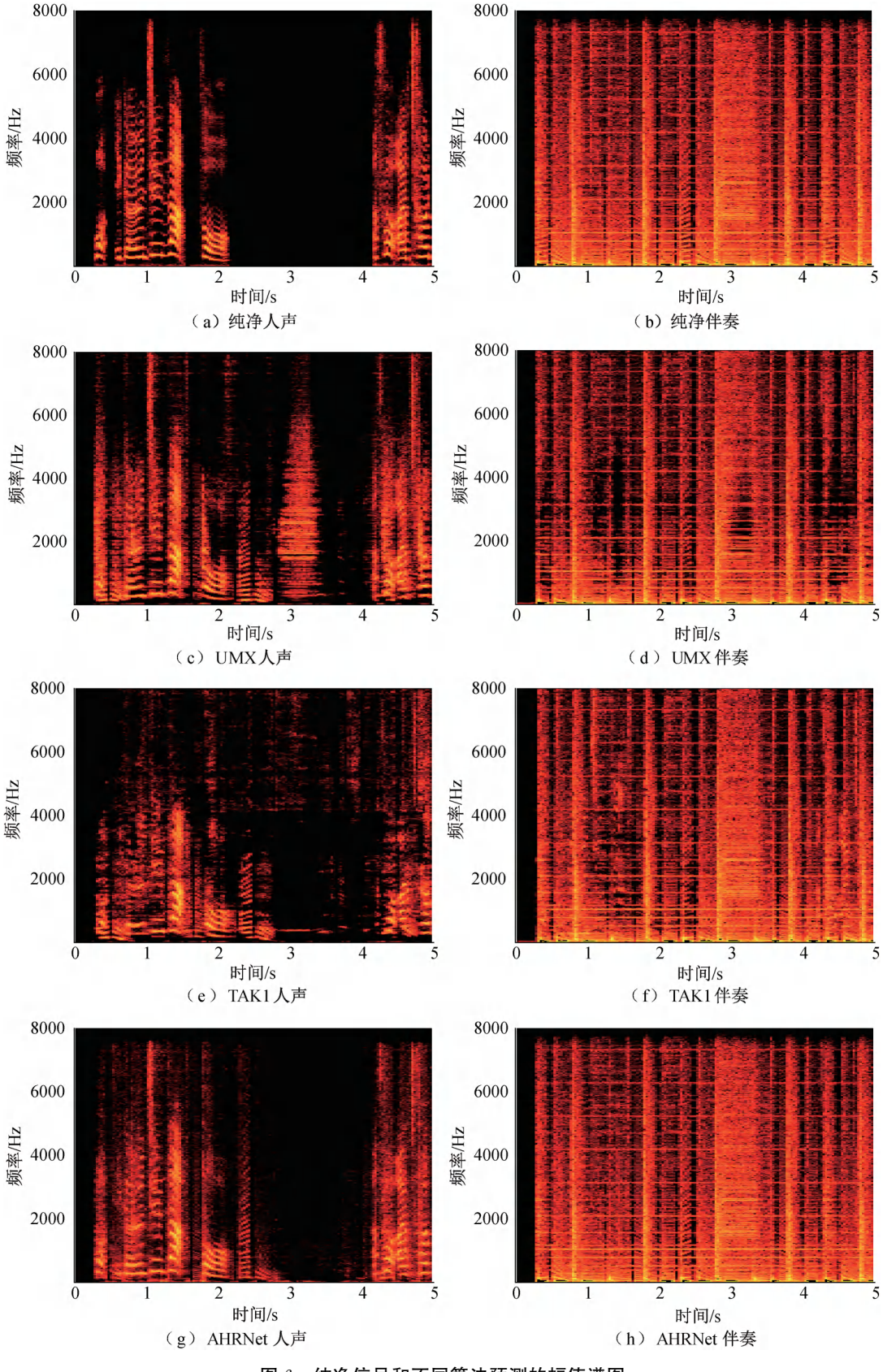



**图 6 纯净信号和不同算法预测的幅值谱图**

**表 1　不同歌声分离算法性能比较**　dB

| 算法 | 人声 | | | 伴奏 | | |
|---|---|---|---|---|---|---|
| | $R_{SD}$ | $R_{SI}$ | $R_{SA}$ | $R_{SD}$ | $R_{SI}$ | $R_{SA}$ |
| UMX[9] | 6.32 | 13.33 | 6.52 | 12.69 | **18.14** | 13.65 |
| TAK1[11] | 6.60 | 14.37 | 6.37 | 12.83 | 16.69 | 14.08 |
| HRNet | 7.22 | 13.97 | 7.73 | 12.52 | 15.48 | 15.53 |
| AHRNet | **7.68** | **14.93** | **8.34** | **12.85** | 15.73 | **15.91** |

为进一步探究自注意力子模块对网络性能的影响，对 AHRNet 四条子网分支的注意力子模块进行消融性实验，分别移除某一分支的注意力子模块，研究其对分离效果的影响。按照特征图分辨率由高到低排序，四条子网分支分别为第一分支、第二分支、第三分支和第四分支。考虑到伴奏分离效果差距不明显，只对人声进行消融性实验，移除不同分支自注意力子模块后的人声分离效果如表 2 所示。将原 AHRNet 第一分支的自注意力子模块移除后，分离效果下降最明显，三个指标值均为最低，而将第四分支的自注意力子模块移除后，分离效果和原网络最接近。特征图从第一分支到第四分支，分辨率由高变低，而特征信息的维度由低变高，因此高分辨率特征图含有浅层特征信息，低分辨率特征图含有深层特征信息。音乐中的重复旋律更多地体现在浅层，原始幅值谱中可以明显观察到重复旋律，而在低分辨率的深层特征图中这一现象不明显。因此，第一分支的自注意力子模块对网络性能影响最大，是因为该分支特征图分辨率最大，含有最浅层最丰富的特征信息，更易捕获到重复旋律特征信息。

**表 2　移除不同分支自注意力模块后人声分离效果对比**　dB

| 移除模块所在分支 | $R_{SD}$ | $R_{SI}$ | $R_{SA}$ |
|---|---|---|---|
| 第一分支 | 7.41 | 14.46 | 7.95 |
| 第二分支 | 7.53 | 14.60 | 8.11 |
| 第三分支 | 7.58 | 14.65 | 8.25 |
| 第四分支 | 7.65 | 14.91 | 8.31 |

## 3　结　论

本文提出了一种基于高分辨率网络和自注意力机制的歌声分离算法，通过使用改进的高分辨率网络在一定程度上解决了普通 U 型网络信息丢失的问题，提升了人声和伴奏幅值谱图的精度；同时，利用自注意力机制使网络学习音乐中重复旋律这一特征，增强了网络分离性能。在 MUSDB18 数据集上的实验结果表明，分离得到的人声和伴奏信号偏差比指标分别达到了 7.68 dB 和 12.85 dB，与目前多个主流算法相比，本文提出的算法分离效果最优，可以有效地分离出高质量人声和伴奏。

**参考文献：**


[1] 李伟，李子晋，高永伟. 理解数字音乐：音乐信息检索技术综述[J]. 复旦学报(自然科学版)，2018，57(3)：271-313.

[2] 熊梅，张天骐，张婷，等. 结合 HPSS 的非负矩阵音乐分离方法[J]. 计算机工程与设计，2018，39(4)：1089-1094.

[3] Huang P S，Chen S D，Smaragdis P，et al. Singing-voice separation from monaural recordings using robust principal component analysis [C] // 2012 IEEE International Conference on Acoustics，Speech and Signal Processing. Kyoto：IEEE，2012：57-60.

[4] 张天骐，徐昕，吴旺军，等. 多反复结构模型的精确音乐分离方法[J]. 声学学报，2016，41(1)：135-142.

[5] Rafii Z，Liutkus A，Stöter F R，et al. An overview of lead and accompaniment separation in music[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2018，26(8)：1307-1335.

[6] Stoller D，Ewert S，Dixon S. Wave-U-net：A multi-scale neural network for end-to-end audio source separation [C] // 19th International Society for Music Information Retrieval Conference. Paris：ISMIR，2018：334-340.

[7] Samuel D，Ganeshan A，Naradowsky J. Meta-learning extractors for music source separation[C] // 2020 IEEE International Conference on Acoustics，Speech and Signal Processing. Barcelona：IEEE，2020：816-820.

[8] Défossez A，Usunier N，Bottou L，et al. Music source separation in the waveform domain [EB/OL]. (2019-11-27) [2021-07－10]. https://arxiv.org/abs/1911.13254.

[9] Stöter F R，Uhlich S，Liutkus A，et al. Open-unmix：A reference implementation for music source separation [J]. Journal of Open Source Software，2019，4(41)：1667.

[10] Takahashi N，Mitsufuji Y. Multi-scale multi-band densenets for audio source separation[C] // 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York：IEEE，2017：21-25.

[11] Takahashi N，Goswami N，Mitsufuji Y. Mmdenselstm：An efficient combination of convolutional and recurrent neural networks for audio source separation[C] // 2018 16th International Workshop on Acoustic Signal Enhancement. New York：IEEE，2018：106-110.

[12] Hennequin R，Khlif A，Voituret F，et al. Spleeter：A fast and efficient music source separation tool with pre-trained models[J]. Journal of Open Source Software，2020，5(50)：2154.

[13] Park S，Kim T，Lee K，et al. Music source separation

using stacked hourglass networks[C]// 19th International Society for Music Information Retrieval Conference. Paris: ISMIR, 2018:289-296.

[14] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piacataway: IEEE, 2019: 5686-5696.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017: 5998-6008.

[16] Rafii Z, Pardo B. Repeating pattern extraction technique (REPET): A simple method for music/voice separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(1): 73-84.

[17] Uhlich S, Porcu M, Giron F, et al. Improving music source separation based on deep neural networks through data augmentation and network blending[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans: IEEE, 2017: 261-265.

[18] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1462-1469.

[19] St öter F R, Liutkus A, Ito N. The 2018 signal separation evaluation campaign[C]// International Conference on Latent Variable Analysis and Signal Separation. Guildford: Springer, 2018: 293-305.

(责任编辑:康 锋)