**COMPLEX RATIO MASKING FOR SINGING VOICE SEPARATION**

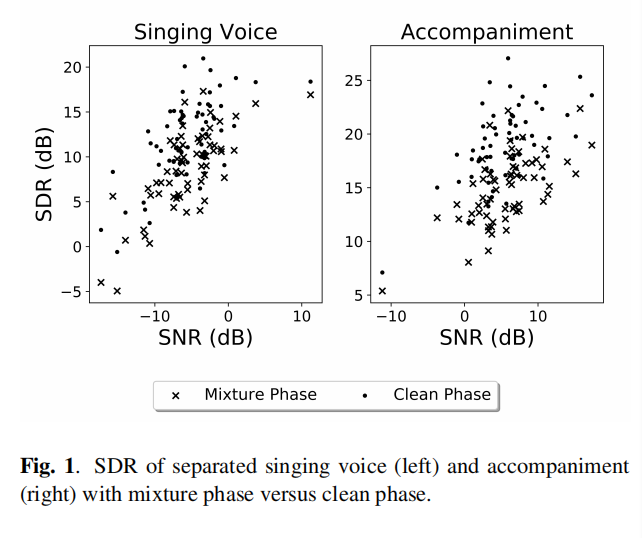
针对歌声分离的复比率掩膜

## 1、研究动机

音乐源分离对于卡拉ok和重新混音等应用程序很重要。以往的研究大多集中在估计短时间傅里叶变换(STFT)的大小和丢弃相位信息上。然而对于歌唱语音分离，相位可以显著提高分离质量。本文提出了一种用于声音和伴奏分离的complex ratio masking方法。该方法采用自注意的DenseUNet来估计各声源STFT的实分量和虚分量。还引入了一种简单的集成技术来进一步提高分离性能。评价结果表明，该方法在分离的声音和伴奏方面都优于最近的最先进的模型。

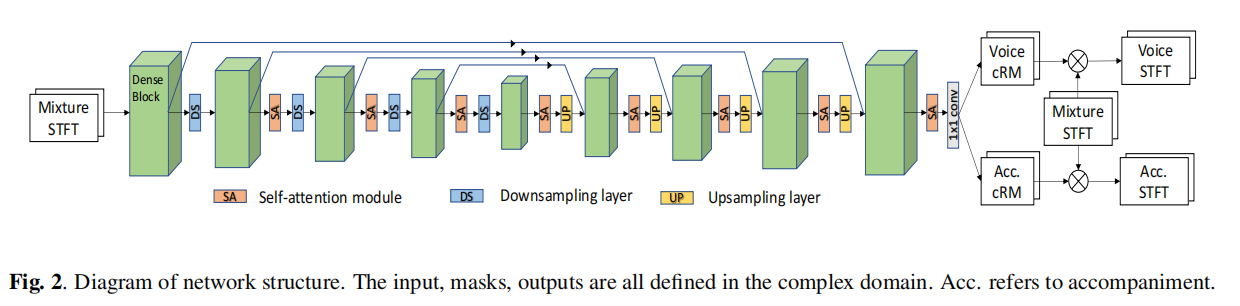
## 2、研究方法

### ①研究相位对歌声分离的重要性



实验首先使用SA-DenseUNet网络估计63首不同信噪比的歌曲歌声和伴奏的频谱图，然后将重新合成的混合相和纯净的目标的相位的SDR作比较，可以看出纯净相的SDR比估计出的混合相高出4到5dB左右。由此可以看出相位的重要性。

### ②网络结构



本研究扩展了SA-DenseUNet[11]，以估计目标源STFT的实分量和虚分量。输入实分量和虚分量作为两个独立的通道。该网络采用了一个由一个编码器和一个相应的解码器组成的DenseUNet结构。

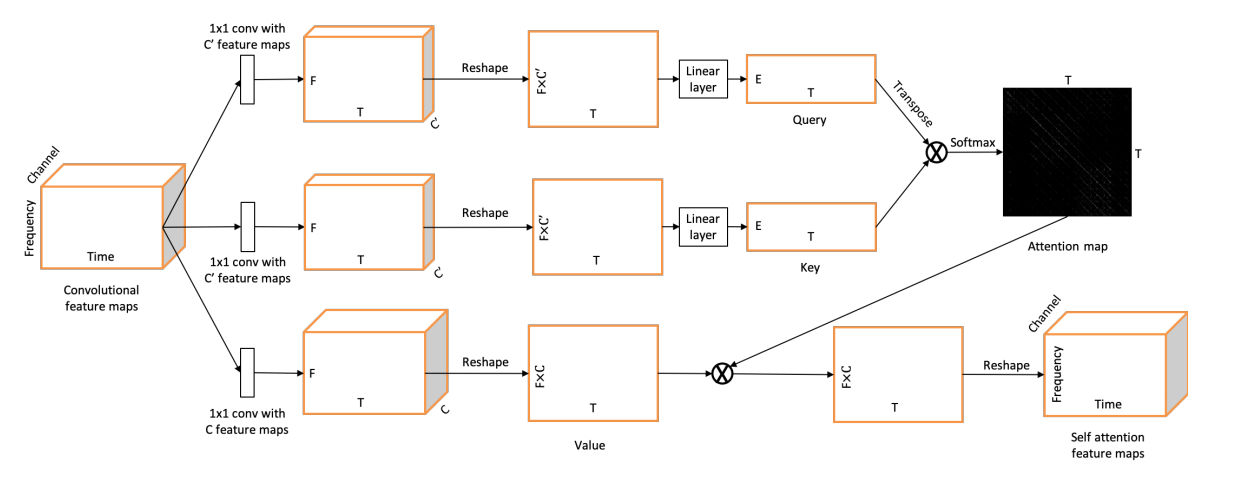
**·编码器**包括一系列密集连接的卷积层Dense Block、自注意模块和下采样层。**解码器**由Dense Block、自注意模块和上采样层组成。在编码器和解码器之间加入了残差连接。

**·Dense Block** ，每一个密集块包含多个卷积层，每一层的输入是前面所有层输出的连接。

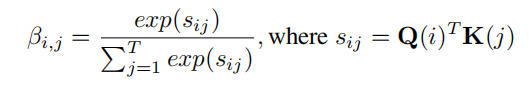


针对一个DenseBlock来说，Hi代表一个卷积层操作加上一个ELU激活函数，Xi-1…x0为前i-1层的输出。Xi为当前第i层的输出。

**·自注意力机制网络结构**



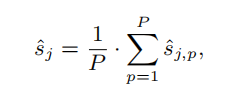
输入特征图有三个维度：通道、频率和时间，用C×F×T表示，其中T覆盖长时间范围（本研究约20秒），F覆盖谱图的整个频率范围，C对应一个DenseBlock输出的通道的数量。

注意力map计算方式如下：  


**·多上下文平均集成技术（MCA）**

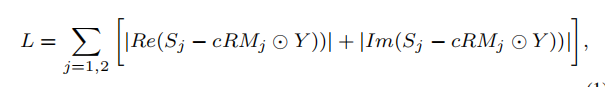
为了让注意力子网络在不同的层次上模拟音乐重复，我们通过使用不同的窗口来利用输入的不同上下文。假设P(P>1)DNNs形成一个DNN集合，每个DNN都有不同的窗口长度。给定Y和相应的真值歌声S1和伴奏S2，训练第p个DNN，将窗口长度设置为wp，估计S1和S2的STFT的实分量和虚分量。

在测试阶段，对于第p个DNN，也用窗长wp计算相应的complex STFT。**所有P DNNs的输出**通过逆STFT转换到时域。然后计算



输出波形的平均值，其中*sˆj* j=1对应的最终估计的歌声，j=2时对应最终估计的伴奏，*s*ˆ*j,p*对应第p个DNN估计的声波形。

### ③Loss损失

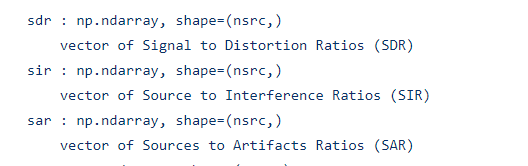
Loss：

*j=1*时代表分离出的歌声人声部分，*j=2*时代表分离出的音乐部分。

*cRMj*——网络输出的预测mask，*Y*——表示输入的complex STFT ，

*Sj*——真实的complex STFT，S1代表歌声，S2代表音乐伴奏。

### ④评价指标



SDR——信号失真率向量

SIR——源干扰率向量

SAR——源与伪影比率向量

以上三种评价指标使用mir\_eval计算。mir\_eval用于计算各种音乐/音频信息检索/信号处理任务的常见启发式准确度分数的 Python 库。三种指标均为越高越好。

## 3、数据集

①DSD100：由Dev和Test组成，每个包含50首歌曲。

②MedleyDB：包括122首歌曲，其中70首歌曲包含声乐。

③CCMixter：包括50首声乐歌曲。

训练集包含450首歌曲，其中50首来自DSD100的开发集，另外400首经过随机缩放，混音等将这50首歌曲进行数据增强。

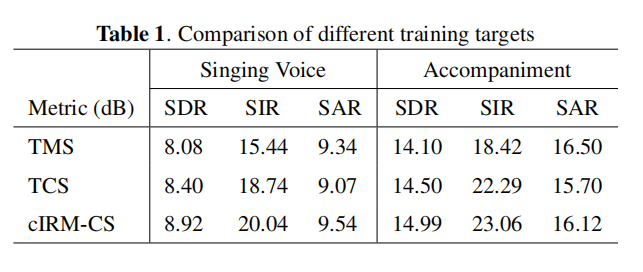
验证集包含了来自MedleyDB和CCMixter的三分之一，以及来自DSD100的测试集的一半。

测试集由MedleyDB和CCMixter的另外三分之一以及DSD100测试集的另一半构建。

所有歌曲降采样到16kHz。

## 4、实验结果

**①训练目标对分离结果的影响：**



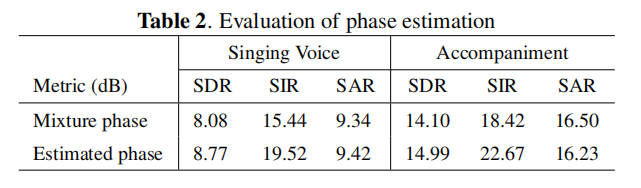
TMS——训练目标为幅度谱

TCS——训练目标为目标复谱（实部谱和虚部谱）

cIRM-CS——训练目标是在复谱中应用了损失函数的cIRM。

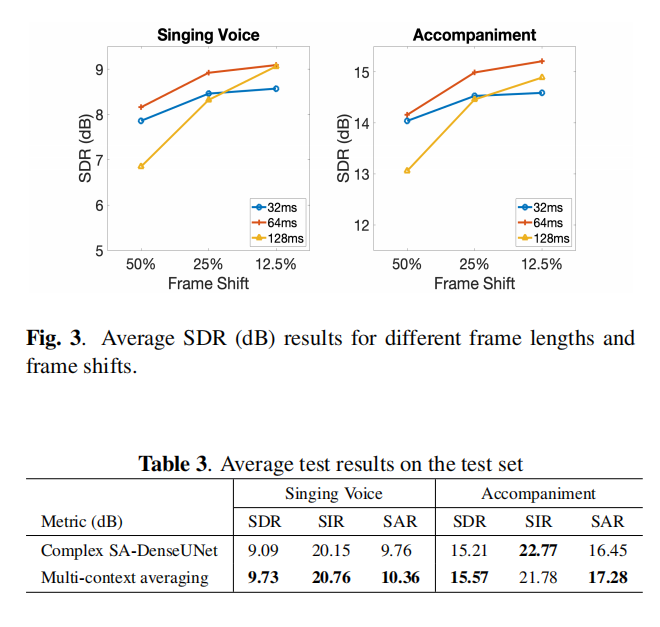
由结果可以看出两个在复数域上做歌声分离的目标都优于单一在幅度域上分离效果。且本文所提出的是当前最优的。

**②**



应用实分量和虚分量可以更好地估计出相位，取SA-DenseUNet估计的STFT值，用混合相和cRM估计的相位重新合成源波形。结果显示歌声和音乐伴奏的SDR都有提升，证明相位估计的重要性。

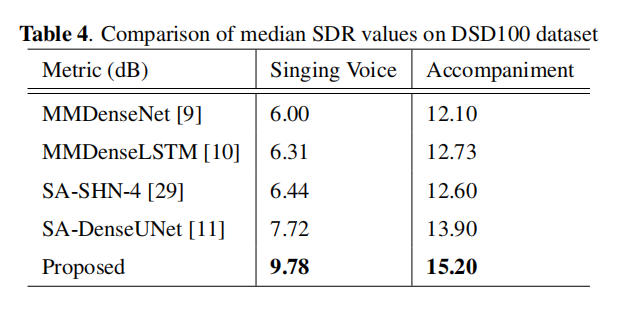
**③实验了使用不同窗口和帧移对实验结果的影响（MCA部分）**



如图3所示，如果只应用一种窗长，窗口长为64ms，且为12.5%的帧移，效果最好。

本文选择了3个complex SA-DenseUNets，其窗口长度为32ms，64ms，128ms来创建多上下文平均(MCA)集成，所有的帧移都为12.5%。与单个最佳窗长相比，MCA将唱歌声音和伴奏的平均sdr分别提高了0.64dB和0.36dB。

④与其他先进模型比较



本例中的训练集使用了DSD100，采样率保留了原始的44.1kHz。验证集包括MedleyDB和CCMixter，并且不包括来自DSD100测试集的任何歌曲。由结果可见效果得到了有效的提升。