歌唱分音:训练数据的研究

Laure pracemet, Romain Hennequin, Jimena Royo-Letelier, Andrea Vaglio

歌唱声分离:训练数据的研究

Romain Hennequin? *JimenaRoyo-Letelier?* *安德里亚Vaglio吗?__*

吗?Deezer R&D，法国巴黎，research@deezer.com †LTCI,
T´el´ecom ParisTech，巴黎萨克雷大学，法国巴黎

摘要。
近年来，歌唱声音分离系统由于使用监督训练而表现出更高的性能。训练数据集的设计被认为是影响这类系统性能的关键因素。我们研究了训练数据集的特征如何影响最先进的歌唱语音分离算法的分离性能。我们表明，分离质量和多样性是一个好的训练数据集的两个重要和互补的资产。我们还提供了对执行此任务的数据增强的可能转换的见解。

索引术语——源分离、监督学习、训练数据、数据增强

1. 介绍

唱腔分离就是将一段音乐录音分解成两条音轨，一边是唱腔，另一边是器乐伴奏。典型的应用有自动卡拉ok创作、混音、音高跟踪[1]、歌手识别[2]、歌词誊写[3]。

这是音乐信息检索(MIR)文献中非常受欢迎的话题，SiSec MUS挑战等年度竞赛聚集了越来越多的团队(2016年评估了24个系统，2018年评估了30个系统)。2018年版的SiSec活动[4]表明，当前最好的系统依赖于基于监督的深度学习模型。特别是卷积神经网络(CNN)似乎特别适合这项任务。最近，一个U-Net[5]和几个基于densenet的系统[6]表现出了令人印象深刻的性能:最先进的模型第一次在仪器部分表现得与oracle系统相似[4]。

然而，尽管取得了这些成就，通常很难确定这些系统的主要成功因素是什么。结果通常是一个完整的过程，包括数据集构建、数据预处理和/或增强、架构设计、后处理，有时还需要长时间的工程工作来调整模型的超参数[7,8,5,9]。

在这项工作中，我们专注于训练数据集对最先进的基于深度学习的分离系统性能的影响。我们通过在改变训练数据集的同时训练相同的基线模型，研究了其中四个不同方面(大小、分离质量、使用数据增强技术和使用来自几种仪器的分离源来估计语音分离)的影响。在之前的文献[10,11,12,13,9]中，通常使用相同的训练/测试数据集来比较不同的架构，但据我们所知，之前没有专门研究这些数据集的影响的作品。与之前的作品相反，我们使用一个单一的最先进的架构，并在不同的数据集上训练它，以揭示训练数据的不同特征对分离性能的影响。我们特别检查了以下方面:数据多样性和分离质量、数据增强和分离源的数量。

多样性和分离质量。在文献中，数据稀缺性经常被引用为构建高效且可扩展的监督歌唱语音分离算法的主要限制之一[14,15,16]。的确，公共训练数据集已经定期发布(MIR-1K[17]、MedleyDB[18]、DSD100[19]、MUSDB[20])，并用于比较不同的方法，但它们规模较小，往往缺乏多样性。我们在这里建议使用几个不同大小和分离质量的数据集来评估具有更大数据量的训练系统的好处。其中包括一个相对较小的公共数据库(MUSDB)，一个大型私有数据集，以及一个大型数据集，其中包含根据[21]中提出的技术从Deezer的音乐目录中构建的估计分离的曲目。

数据增加。用于人为增加MIR任务数据集大小的常用方法是数据增强。例如，在歌唱声音检测中，一些数据增强(如音调移位或随机频率滤波器的应用)已被证明可以提高性能[22]。此外，在[8]中，作者研究了其他数据增强(通道交换、幅度缩放或随机分块)的使用，但没有改善结果。我们建议研究在小型数据集上使用几种数据增强技术的影响。

几个来源。最后，我们研究了使用几种来源(*低音*，*鼓*和MUSDB中可用的*其他*部分)来估计*器乐*部分的影响。事实上，当只估计*人声*部分和*乐器*部分时，源分离系统倾向于在人声估计中包括来自其他乐器(特别是来自*鼓*)*的残余部分*。因此，使用多个源中包含的附加信息可以更好地建模*器乐*部分，从而实现更好的分离。

本文的其余部分组织如下。在第2节中，我们介绍了我们用于实验的三个数据集。在第3节中，我们详细介绍了我们用来比较不同数据集上的性能的方法。在第4节中，我们展示了我们的结果并讨论了可能的解释。最后，我们在第5节中得出结论。

2. 数据集

在本节中，我们介绍了我们在实验中使用的三个训练数据集，以及它们的主要特征。除了音频的总持续时间外，我们还定义了*质量*标准*和多样性*标准。数据集的质量反映了数据集音轨中源分离的质量:在两个数据集(MUSBD和Bean)中，分离的音轨来自不同的录音，而在最后一个数据集(Catalog)中，声乐部分不能作为单独的音轨使用，必须进行估计。在最后一种情况下，分离的音轨只是估计值，来自其他来源的残差可以存在于地面真实音轨中。这个标准没有考虑

针对的是制作质量，也不是音质。多样性标准-rion反映了构建数据集的歌曲的可变性。它可以通过数据集中一个或多个片段表示的不同歌曲的数量来量化。表1总结了这些信息。

2.1. MUSDB

MUSDB是用于源分离的最大和最新的公共数据集。MUSDB主要由来自DSD100和MedleyDB数据集的歌曲组成，在上次歌唱分音运动中作为训练和测试数据的参考[4]。该数据集由150首专业制作的歌曲组成。只有西方音乐流派存在，绝大多数是流行/摇滚歌曲，还有一些嘻哈、说唱和金属歌曲。100首歌属于训练集，50首属于测试集。

对于每首歌，有5个音频文件可用:混音，和4个独立的音轨(鼓、*贝斯*、*人声*等)。原始混音可以通过直接将四个音源的音轨相加来合成。为了创建*器乐*源，我们将*鼓*、*贝斯*等*对应的音轨加起来*。在我们的实验中，我们同时考虑了*器乐*/*人声*数据集和4茎数据集。

	MUSDB	Catalog	Bean
Diversity	150 songs	28,810 songs	24,097 songs
Quality	Separated recordings	Estimates	Separated recordings
Duration	10 hours	95 hours	79 hours
Train/val/test (%)	53/13/33	97/3/0	85/8/7

表1:三个数据集的主要特征。

2.2. 豆

除了MUSDB之外，我们还使用一个名为Bean的私有多轨道数据集。Bean数据集包含大多数流行/摇滚歌曲，并将人声和器乐曲目作为分离的录音包括在内。在该数据集中的24,097首可用歌曲中，21,597首用于训练，2,000首用于验证，剩下的500首用于测试。

总的来说，Bean数据集代表了5679位不同的艺术家。我们将训练/验证/测试分割为这样一种方式，即艺术家不能同时出现在分割的两个部分中，就像在MUSDB中一样。这是一个重要的预防措施，以确保分离系统不会对艺术家进行过拟合，这是MIR中经常提出的问题[23]。我们对Bean进行了类型统计，如图1(绿色直方图)所示。Bean的类型分布主要以流行和摇滚歌曲为主，这与MUSDB非常相似。

2.3. 目录

为了构建这个数据集，我们从[21]中获得了灵感，其中提出了一种基于音乐流媒体目录构建数据集的方法。我们改编了这种方法，通过利用一些艺术家与原始歌曲一起发布的*器乐版本*，从Deezer的目录中构建了一个数据集。

第一步是在目录中找到所有可能的*器乐*/*混音*曲目对。这种匹配是使用元数据和音频指纹来完成的。然后，执行一些滤波和均质化操作:如果两个音轨的持续时间差异大于2秒，则删除一对。超过5分钟的歌曲会被过滤掉。然后，使用自相关技术暂时重新排列一对歌曲中的曲目。最后，对两条音轨的响度进行均衡化。

为了从这对*(混音，器乐)中产生一个三连音(混音，器乐,人声)*，我们对两个声谱图进行半波整流差分。最终，我们创造了28,810个三连音。我们将它们拆分为训练数据集和验证数据集，确保给定的艺术家不能同时出现在拆分的两个部分。我们把这个数据集称为*目录A*。

使用元数据，我们注意到与Bean(以及MUSDB)中的类型分布相比，该数据集中对儿童音乐和嘻哈音乐有重要的类型偏好，如图1所示。为了克服这个问题，我们建立了第二个数据集，通过重新平衡每个类型的表示，使最终的分布与Bean的分布相匹配。我们将这个数据集称为*目录B*。

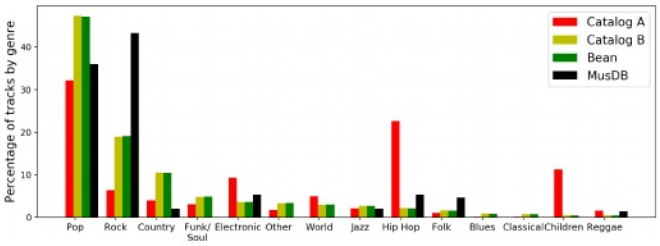

图1:Bean、Catalog和MUSDB数据集的类型分布。

尽管与MUSDB相比，Catalog受益于非常大的容量，但我们必须记住，它不是为分离目的而专业生产的，并且质量必然较低。我们在数据集中发现的两个主要问题是:

- 混音和器乐之间的半波校正差异并不完全对应于人声部分。这是因为该操作是在幅度谱图上执行的，而源的可加性并没有得到保证。此外，两个音轨之间最小的不对准也会在人声中产生器乐残余。对一小部分(40个音轨)进行的非正式听力测试显示，这种情况发生在几乎50%的音轨中。
- 如果元数据匹配不完美，混音中可能会有没有唱腔的歌曲。在这种情况下，*人声*部分只是残留的杂音。相反，有些*器乐*音轨包含唱诗班。这些情况很难被自动系统检测到。

因此，我们可以说Catalog数据库形成了大量的弱标记训练数据。器乐部分是专业制作的，而人声部分只是估计值。

3. 方法

3.1. 网络体系结构

在本文中，我们重点使用深度神经网络来执行分离。我们选择的基线模型是[5]中提出的U-Net。该架构在DSD100数据集[5]和上一次SiSeC活动[4]上显示了最先进的结果。在对其他架构(DenseNet和MMDenseNet[6])进行了一些试点实验后，我们选择了U-Net，即使在大型数据集上，它也可以在合理的时间内进行训练。它也是一种简单、通用的架构，可以应用于各种领域[24]。

U-Net与卷积自编码器共享相同的架构，具有额外的跳过连接，可以将编码阶段丢失的详细信息恢复到解码阶段。它在编码器中有5个跨步二维卷积层，在解码器中有5个跨步二维反卷积层。

与[5]相比，主要的修改是集成了立体处理:我们使用3D张量(通道、时间步长、频率箱)作为网络的输入和输出。其他层没有进行修改。

3.2. 数据准备

在原始数据集中，所有歌曲都是立体声的，采样频率为44100Hz。为了降低计算成本，我们将它们重新采样到22050Hz。我们

我们将所有歌曲分成11.88秒的片段。对于Catalog和Bean，我们从训练集和验证集的每首歌中随机选择一个片段，避免了引言(前20秒)和后场(最后20秒)这两个人声经常缺失的部分。我们还使用来自Bean的500首曲目构建了第二个测试数据集，从中我们能够提取1900个片段。我们确保在图1中最具代表性的10种类型上平衡其类型分布。最终的分割比例可以在表1中看到。

与[5]类似，我们使用短时傅里叶变换(STFT)作为我们网络的输入和输出特征。窗口大小为2048，步长为512。我们选择这些设置是为了在去掉最高频带后，频谱图的维度是2的幂:(通道，时间步长，频率箱)=(2,512,1024)。这是必要的，因为我们使用的网络架构通过一个因子减少了频谱图的维度，这个因子是2的幂。

3.3. 培训

对于每个声源(人声和*器乐*)，我们训练了一个U-Net从混合的声级谱图中输出相应的声级谱图。我们使用带有Tensorflow后端的Keras对每个网络进行了500次epoch的训练。我们将一个epoch定义为800个梯度下降步骤。为了限制过拟合，我们使用每个数据集的验证分割进行提前停止。训练损失为目标谱图与被屏蔽输出谱图之差的L_1范数，如[5]所述。优化器为ADAM，学习率为0.0001。在进行短网格搜索后，将批大小设置为1。

3.4. 重建

一旦训练完成，我们在测试数据集上执行推理传递，同样被切成11.88秒的片段。通过从两个估计中计算一个比例掩模，并将其应用于原始混合频谱图，重建每个源的复杂频谱图。这样，输出相位就是混合物的相位。通过将源的谱图估计值(对应的U-Net输出)除以两个估计值的和，得到源的比率掩模。对于4音干分离的特殊情况，*乐器*谱图估计是通过对3个非声乐音干的谱图估计求和得到的。STFT是倒置的，通过简单地连接不同的片段来重建完整的歌曲。音频最终被上采样回44100Hz。

3.5. 评价

我们使用Museval[20]工具箱来计算标准的源分离度量:信失真比(SDR)、信干扰比(SIR)和信伪比(SAR)。我们使用所有1秒帧的中位数来聚合这些指标，以保持每首歌和每个源的单一指标，如[4]所示。我们在MUSDB和Bean测试数据集上运行评估过程。

为了比较不同方法的性能，我们还对每首歌指标进行了配对学生t检验。这一步的动机是观察到度量分布中的方差很高，这使得有时很难评估一种方法是否比另一种方法表现得更好。即使两种方法可能产生非常相似的度量分布，这些度量也可能以依赖的方式变化(例如，具有小但恒定的差异)。配对t检验有助于揭示这一现象。

4. 实验和结果

4.1. 数据增加

当在像MUSDB这样的小数据集上训练时，数据增强经常被引用为提高分离性能的一种方法[8]。在

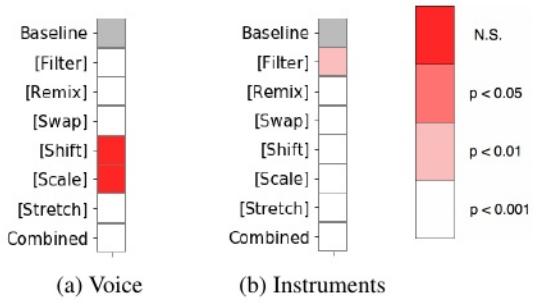

图2:数据增强实验:MUSDB Test数据集上SDR的Student配对t检验结果。

在这个实验中，我们试图弄清楚数据增强能在多大程度上提高分离性能。对于选择要执行的数据变换，我们从[22]中获得了灵感，其中作者在声谱图上使用了一组变换，并测试了对歌唱语音检测任务的效果。我们设置了一组类似的实验来评估各种形式的数据增强对分离结果的影响。我们采用了Sch¨ulter提出的用于源分离的变换(音高移动、时间拉伸、响度修改和滤波)，并添加了通道交换(参见[8])和源混音。在源分离的背景下，数据增强的特殊性在于，目标和输入都必须用完全相同的变换进行处理。下面是我们使用的各种变换的细节:

通道交换[Swap]:左右通道交换的概率为0.5。

时间拉伸[Stretch]:我们将频谱图的时间轴按一个因子$\beta_{stretch}$线性缩放，并保留中心部分。$\beta_{stretch}$是从每个样本的0.7到1.3(±30%)的均匀分布中随机抽取的。请注意，这是与音频速度的实际修改相比的近似值。

音高移位[Shift]:我们将频谱图的频率轴线性缩放一个因子β_{shift}，并保持底部部分，使最低频段保持与0 Hz对齐。β_{shift}是从每个样本的0.7到1.3(±30%)的均匀分布中随机抽取的。请注意，这是与音频的实际音高移动相比的近似值。

混音[Remix]:我们用随机的响度系数重新*混音器乐*和人声部分，在- 9dB和+9dB之间的对数尺度上均匀绘制。

反高斯滤波[Filter]:我们对每个样本应用频率响应为$f(s) = 1 - e^{-(s-\mu)2}/2\sigma^2$的滤波器，其中μ在0到4410Hz的线性尺度上随机选择，σ在500Hz到1000Hz的线性尺度上随机选择。

响度缩放[Scale]:我们将频谱图的所有系数乘以一个因子β_{scale}。β_{scale}在−10dB和+10dB之间的对数尺度上均匀绘制。

组合:我们同时执行通道交换，音高移动，时间拉伸和混合数据增强。

中位数源分离度量(SDR、SAR、SIR)见表2。为了了解度量差异的重要性，我们在数据增强训练和非数据增强基线之间执行了配对学生t检验:我们报告了图2中MUSDB测试集上应用于SDR的该检验的p值。

表2显示，在某些情况下，数据增强可能会对分离指标产生积极影响:特别是在Bean数据集上，通道交换、音调移动和时间拉伸似乎相当一致地改善了大多数指标。然而，必须注意到这一点

即使在我们进行的测试中，改进在统计上显着，但改进非常有限，在SDR中几乎不超过0.2dB，这是非常低的，甚至可能听不到。因此，我们测试的各种数据增强类型似乎对分离结果的影响相当低，同时在文献中被广泛使用。

		Voice			Instruments		
Test	Transform	SDR	SIR	SAR	SDR	SIR	SAR
MUSDB	*Baseline*	*4.32*	*12.62*	*4.1*	*10.65*	*13.46*	*11.51*
	[Filter]	3.9	**13.35**	3.33	10.27	12.57	11.66
	[Remix]	3.75	12.89	3.6	10.45	11.81	**12.05**
	[Swap]	4.37	**13.01**	4.08	**10.69**	13.08	**11.74**
	[Shift]	4.0	**15.3**	3.5	10.58	12.46	**12.11**
	[Scale]	4.05	12.6	3.64	10.68	12.38	**11.85**
	[Stretch]	4.19	**13.44**	3.57	10.96	12.76	**12.09**
	Combined	3.76	**13.86**	3.3	10.48	12.35	**11.72**
Bean	*Baseline*	*5.91*	*9.23*	*5.73*	*9.33*	*12.43*	10.9
	[Filter]	5.58	**10.8**	5.2	9.18	11.53	10.75
	[Remix]	5.7	**10.18**	5.44	9.43	11.1	**11.4**
	[Swap]	**5.98**	**9.94**	**5.83**	**9.5**	12.25	**11.24**
	[Shift]	**6.06**	**11.53**	5.82	**9.57**	11.67	**11.63**
	[Scale]	5.87	**9.55**	5.66	9.42	11.71	**11.32**
	[Stretch]	**6.12**	**10.68**	**5.94**	**9.64**	12.18	**11.35**
	Combined	5.98	**11.45**	**5.99**	9.4	11.1	**11.07**

表2:数据增强实验:在MUSDB上进行数据增强训练的U-Net结果。加粗为较基线显著改善的结果($p < 0.001$)。

4.2. 训练数据集的影响

在这个实验中，我们评估了训练数据集对所选分离系统性能的影响。该系统使用第2节中提供的5个数据集进行训练:*目录A*, *Cata-log B*, Bean，具有两个词干(伴奏和*人声*)的MUSDB和具有四个词干(人声，*鼓*，*贝司*等)*的MUSDB*。在每个数据集上训练系统后，我们评估了它在两个测试数据集上的性能:MUSDB和Bean。表3报告了源分离指标所有轨迹的中位数，图3报告了在MUSDB测试数据集上获得的SDR之间的配对学生t检验的p值。

正如预期的那样，在Bean数据集上的训练在人声和伴奏部分以及两个测试数据集上的大多数指标上都获得了最高分。值得注意的是，在Bean上训练的系统的*人声*部分的SDR值比所有其他系统的SDR值在MUSDB测试集上高出1dB以上，在Bean测试集上高出1.5dB以上，这是非常重要的(并且在感知上非常明显)。这证实了拥有具有干净分离音轨的大型数据集是提高源分离系统性能的好方法。更令人惊讶的是，所有其他训练数据集彼此之间都提供了相当相似的性能。特别地，用4个琴干而不是2个琴干进行训练并没有显著提高MUSDB上的指标:然后在这个特定的设置中，添加额外的信息来帮助建模伴奏谱图实际上并没有提高性能。

我们还注意到，使用两个*Catalog*数据集训练系统对分离性能的影响非常有限。与单独的MUSDB相比，它产生更高的SAR，但更低的SIR，从而产生类似的SDR。这种效果在人声中尤为明显。从Catalog训练数据集的构建方式来看，这是有道理的:录音是专业制作的，因此混合质量很好，但在声乐目标中仍然存在明显的泄漏。此外，使用*目录A*或*目录B*进行训练似乎提供了非常相似的结果，这意味着*目录A*和*Bean*之间类型分布的差异并不是造成性能高差异的原因，而性能低的实际原因可能是分离的质量较低

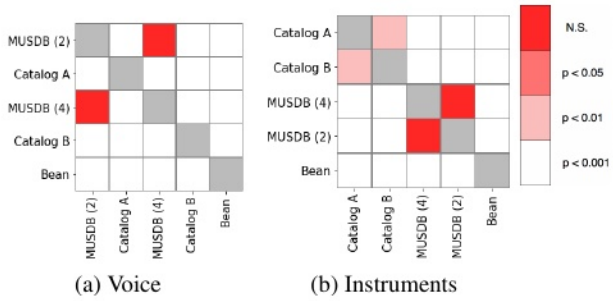

图3:训练数据集对比实验:MUSDB测试数据集上SDR的student - student配对t检验结果。SDR由左上至右下递增。

数据集的轨迹。

因此，与使用Bean等大型干净数据集相比，在具有低质量半自动获得的源的大型多样化数据集上训练系统似乎对性能指标的影响非常有限。这与[5]中的建议相矛盾，在[5]中，数据集大小的影响被认为是重要的(尽管这方面没有在所有其他因素固定的情况下进行测试)。

		Voice			Instruments		
Test	Train	SDR	SIR	SAR	SDR	SIR	SAR
MUSDB	MUSDB (2 stems)	4.32	12.62	4.1	10.65	13.46	11.51
	MUSDB (4 stems)	4.44	12.26	4.2	10.61	13.7	11.48
	Catalog A	4.2	7.6	**7.44**	10.47	12.84	12.03
	Catalog B	4.34	8.04	7.05	10.6	12.8	12.12
	Bean	**5.71**	**14.82**	5.19	**11.99**	**16.04**	**12.21**
Bean	MUSDB (2 stems)	5.91	9.23	5.73	9.33	12.43	10.9
	MUSDB (4 stems)	5.88	8.56	5.71	9.3	12.87	10.92
	Catalog A	5.85	7.26	7.16	9.56	11.68	12.3
	Catalog B	6.05	7.62	6.79	9.74	11.85	**12.42**
	Bean	**7.67**	**12.33**	**7.51**	**11.09**	**15.35**	12.17

表3:训练数据集对比实验:U-Net系统在5个不同数据集上的训练结果。每个测试数据集上的最佳结果以粗体显示。

5. 结论

在本研究中，我们考虑了训练数据集的哪些方面对特定的最先进的源分离系统(U-Net)的分离性能有影响。在这个设置中，我们表明，虽然在文献中经常使用数据增强，但当在小型训练数据集上执行时，对分离结果的影响非常有限。我们还表明，通过访问比执行分离任务所需的更多的来源(4个茎而不是只有*人声*和*伴奏*)带来的额外信息并不能提高系统性能。此外，我们表明，与文献中假设的相反，与具有单独记录源的较小数据集相比，具有半自动获得的人声源的大型数据集对所研究的系统没有太大帮助。最后，我们证实了一个共同的信念，即拥有一个具有干净分离源的大数据集比一个小数据集能显著提高分离结果。

在未来的工作中，我们可能会尝试将这些结果推广到其他最先进的源分离系统中。此外，我们关注的是客观的源分离指标，这些指标被认为不能很好地解释系统之间的感知差异。然后，评估更关注感知影响的数据的影响将是这项工作的相关延续。

6. 引用

[1] Emanuele Pollastri, "A pitch tracking system dedicated to process singing voice for music retrieval," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. 341–344.

[2] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods.," in *ISMIR*, 2007, pp. 375–378.

[3] Annamaria Mesaros, "Singing voice recognition for music information retrieval," *Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication; 1064*, 2012.

[4] Fabian-Robert St¨oter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[5] Andreas Jansson, Eric J Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 323–332.

[6] Naoya Takahashi and Yuki Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 21–25.

[7] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[8] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 261–265.

[9] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," *arXiv preprint arXiv:1711.00048*, 2017.

[10] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," *arXiv preprint arXiv:1805.02410*, 2018.

[11] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent, "Multichannel music separation with deep neural networks," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1748–1752.

[12] Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 61–65.

[13] Zhe-Cheng Fan, Yen-Lin Lai, and Jyh-Shing Roger Jang, "Svs-gan: Singing voice separation via generative adversarial net-work," *arXiv preprint arXiv:1710.11428*, 2017.

[14] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia G´omez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 258–266.

[15] Stylianos Ioannis Mimilakis, Konstantinos Drossos, Tuomas Virtanen, and Gerald Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," *arXiv*, vol. 1709, 2017.

[16] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 429–436.

[17] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[18] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research.," in *ISMIR*, 2014, vol. 14, pp. 155–160.

[19] Antoine Liutkus, Fabian-Robert St¨oter, Zafar Rafii, Daichi Ki-tamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave, "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation - 12th Inter-national Conference, LVA/ICA 2015, Liberec, Czech Repub-lic, August 25-28, 2015, Proceedings*, Petr Tichavsk´y, Mas-soud Babaie-Zadeh, Olivier J.J. Michel, and Nad`ege Thirion-Moreau, Eds., Cham, 2017, pp. 323–332, Springer Interna-tional Publishing.

[20] Zafar Rafii, Antoine Liutkus, Fabian-Robert St¨oter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[21] Eric Humphrey, Nicola Montecchio, Rachel Bittner, Andreas Jansson, and Tristan Jehan, "Mining labeled data from web-scale collections for vocal activity detection in music," in *Pro-ceedings of the 18th ISMIR Conference*, 2017.

[22] Jan Schl¨uter, *Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals*, Ph.D. thesis, Johannes Kepler University Linz, Austria, July 2017, Chapter 9.

[23] Arthur Flexer, "A closer look on artist filters for musical genre classification," *World*, vol. 19, no. 122, pp. 16–17, 2007.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.