Computer Science and Application 计算机科学与应用, 2022, 12(12), 2788-2794
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212283

Hans汉斯

基于深度学习的音乐源分离技术研究

卞宇仁

天津工业大学，天津

收稿日期：2022年11月15日；录用日期：2022年12月15日；发布日期：2022年12月22日

摘　要

音乐源分离技术在音乐产业中起着重要的作用。随着深度学习的发展，音乐源分离技术也产生了巨大的变化，由传统的基于知识的源分离转变为数据驱动的源分离。本文将基于深度学习的音乐源分离分为基于频域的音乐源分离和基于时域的音乐源分离，探讨了这些深度学习模型的原理和优缺点，并介绍了音乐源分离数据集的发展历史，最后对音乐源分离技术的进一步发展做了展望。

关键词

深度学习，源分离，音乐

Research on Music Source Separation Technology Based on Deep Learning

Yuren Bian

Tiangong University, Tianjin

Received: Nov. 15th, 2022; accepted: Dec. 15th, 2022; published: Dec. 22nd, 2022

Abstract

Music source separation technology plays an important role in the music industry. With the development of deep learning, music source separation technology has also produced great changes, from the traditional knowledge-based source separation to data-driven source separation. In this paper, music source separation based on deep learning is divided into music source separation based on frequency domain and music source separation based on time domain. The principles, advantages and disadvantages of these deep learning models are discussed, and the development history of music source separation data sets is introduced. Finally, the further development of music source separation technology is prospected.

文章引用: 卞宇仁. 基于深度学习的音乐源分离技术研究[J]. 计算机科学与应用, 2022, 12(12): 2788-2794.
DOI: 10.12677/csa.2022.1212283

Keywords

Deep Learning, Source Separation, Music

1. 引言

音乐是一种重要的艺术表现形式，在娱乐产业中发挥着重要作用。数字化和互联网带来了一场音乐传播方式的巨大变革。在音乐录制中，已有各种应用程序实现了与单个音频对象交互的能力，如音乐上混和再混、对象均衡等。大多数公开的音乐唱片都是以单声道或立体声组合的形式发布——即多个声音对象共享一个音轨。因此，单个声音对象的操作需要将立体声混合音频分离成几个不同的音轨，每个不同的声源一个音轨，这一过程被称为音乐源分离[1]。

在过去的二十年里，人们对音乐源分离问题进行了大量的研究。已有的音乐源分离方法可大致分为传统音乐源分离方法和基于深度学习的音乐源分离方法两大类。早期的音乐源分离模型通过利用信号处理技术，如特定乐器的声谱图形状来进行源分离。采用的技术有主成分分析 Principal Component Analysis (PCA) [2]，非负矩阵分解 Non-Negative Matrix Factorization (NMF) [3]，多通道 NMF [4]等。这些传统方法已经实现了一定质量的音乐源分离效果并且具有可解释的优点。然而，他们中的大多数并不提供清晰的乐器标识，而是根据信号的特征执行无监督的分离。此外，他们的分离性能主要依赖于对音乐数据特征的了解。因此，基于信号处理技术的分离模型并不能达到最佳的分离性能[5]。

在过去的十年里，由于计算能力的提高和神经网络对 GPU 训练适应性的提高，机器学习成为一个可行的选择，大多数研究人员也开始将他们的努力集中在这条道路上。因此，音乐源分离方向从基于知识的源分离转变为数据驱动的源分离，在这种情况下，模型的训练数据越多，其性能就越好[6]。可以在最近的音乐信号分离评估运动(SiSEC) [7]中看出，现在所有性能较好的音乐源分离模型都使用了深度学习技术。

通常的音乐源分离深度学习模型由三个部分组成：a) 编码器，它接收混源信号并生成适当表示。b) 分离模块，也称为掩码模块，它计算一个要应用于混源信号表示的掩码，以便获得目标源的表示。c) 解码器，它根据源信号的编码重新构造源信号[8] [9]。

根据编码和解码的执行方式，可以区分为两种类型的模型：a) 基于频域的模型，使用短时傅里叶变换、梅尔变换或类似的固定时频变换进行编码。b) 基于时域的模型，它具有可学习的编码、解码模块和分离模块。

2. 评估指标

对音乐源分离效果的评估是一个开放的问题。在评估音乐源分离模型时，通常使用客观和主观方法相结合的方式。客观评价方法的目的是利用客观指标对来音乐源分离方法进行系统评估，为音乐源分离研究提供一个标准化的评价程序，从而使不同的分离方法可以进行一个相对公平的比较。然而，客观的评价指标往往不能反映音频信号的感知质量[10]。

为了解决这一问题，研究人员开发了感知驱动的音乐源分离评估客观度量[11]和常规音乐源分离方法

的主观评价[12]。主观评价方法旨在定性地评估分离后的音乐质量。定性评估的一个缺点是它依赖于听力测试，通常需要多名受过训练的参与者。这使得定性评估很难在大规模评价研究中进行运用。

2.1. 客观指标

音乐源分离的客观评估指标已经有了一套成熟的评估流程，根据此评估程序，目标源信号可以被分解为如下几个部分[13]：

$$\hat{s} = s + e_{spat} + e_{interf} + e_{artif} \tag{1}$$

其中 s 表示干净的源信号，e_{spat} 表示空间失真引起的错误，e_{interf} 表示由其他干扰源引起的错误。e_{artif} 表示由伪影引起的误差。e_{spat}，e_{interf}，e_{artif} 可以通过下面的投影矩阵来进行计算：

$$P_{s_j} = \Pi\left\{s_j\right\} \tag{2}$$

$$P_a = \Pi\left\{\left(s_{j'}\right)_{1\le j'\le n}\right\} \tag{3}$$

$$e_{spat} = P_{s_j}\hat{s}_j - s_j \tag{4}$$

$$e_{interf} = P_s\hat{s}_j - s_j - e_{spat} \tag{5}$$

$$e_{artif} = \hat{s}_j - s_j - e_{spat} - e_{interf} \tag{6}$$

从上面的分解中可以得到下列指标：

源失真比：

$$\mathrm{SDR} = 10\log_{10}\frac{\|s\|^2}{\left\|e_{spat} + e_{interf} + e_{artif}\right\|^2} \tag{7}$$

源干扰比：

$$\mathrm{SIR} = 10\log_{10}\frac{\left\|s + e_{spat}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{8}$$

源伪影比：

$$\mathrm{SAR} = 10\log_{10}\frac{\left\|s + e_{spat} + e_{interf}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{9}$$

还有一些文献使用尺度不变源失真比(SISDR) [14]作为评估指标，定义如下：

$$\text{SI-SDR}\left(s,\hat{s}\right) = 10\log_{10}\frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2}, \alpha = \frac{\hat{s}^{\mathrm{T}}s}{\|s\|^2} \tag{10}$$

由于 SISDR 的可微分性，其也可以直接用作模型的损失函数。

2.2. 主观指标

音乐源分离的主观评价主要是指对分离音源质量的感性评价。在音乐源分离的研究中，音乐源分离的主观评价主要通过听力测试来实现，由于听力测试的实现较为困难，目前大多数文献仍很少使用听力测试，尽管研究表明应该高度重视听力测试[12]。

听力测试的进行必须遵循一个标准的流程。对于音乐源分离，国际电信联盟制定了一个标准(ITU-R

BS.1534-1) [15]，称为多激励隐藏参考基准测试方法。这是一种双盲的多激励音频信号听音比较测试方法，要求选用有经验的听音人，首先进行培训以熟悉测听过程并了解测试样本的损伤程度，再开展正式测试[16]。

3. 基于频域的音乐源分离

随着深度学习技术的发展，在音乐源分离领域性能靠前的深度学习模型大多采用频域与时域相结合的方式。但是基于频域的模型因为有着较为深厚的研究基础，也表现出了不错的性能。

以在 MUSDB18 数据集[17]为基础的音源分离模型为背景。目前，基于频域最好的模型是字节跳动公司研发的 CWS-PResUNet [18]，该模型基于信道子带相位感知 ResUNet [19]，将信号分解为子带宽，并为每个源估计一个无约束复理想比掩码。CWS-PResUNet 利用信道子带特性限制声谱图上不必要的全局权值共享，减少了计算资源的消耗。节省的计算成本和内存反过来可以支持更大的体系结构，在 MUSDB18HQ 测试集上，作者提出了深度达 276 层的 CWS-PResUNet 的架构，并在人声分离上达到了 8.92 的 SDR 分数，实现了目前最好的人声分离性能。

另外基于频域的音源分离模型还有索尼公司提出的 MMDenseLSTM [20]，它结合了 DenseNet [21]和 LSTM，使其能够对语音环境下的长距离依赖信息进行建模。该方法与使用理想二值掩模进行歌唱声音分离的方法相比得到了更好的结果。在音乐源分离开源模型中使用最广泛的模型是 Open-Unmix [22]，它的核心部分是一个三层双向的 LSTM。Open-Unmix 分离模块的输入是一个幅值谱图，因此在重构信号时不使用相位信息，而是将混源的相位应用到最后。该模型得到的输出是应用于输入谱图的掩模，将掩膜与输入谱图相乘得到分离源的谱图。最终使用多通道维纳滤波器对模型输出的源进行初始化分离。Open-unmix 使用的损失函数是单个乐器信号的模型输出和实际幅度谱图之间的 L2 损失。另外 open-unmix 模型每次只分离一个源，因此想要对多乐器进行分离时，需要为每种乐器训练一个单独的模型。

4. 基于时域的音乐源分离

通常的音乐源分离过程包括首先计算音乐信号的时频表示，并将该时频表示进一步传入分离模块中，最后再根据时频表示的编码方式进行重构源信号。这一过程虽然有助于处理较长的语音信号，但是也丢失了重要的相位信息。为了避免相位信息的丢失，研究人员试图构建端到端系统[23]。

基于时域的端到端音乐源分离系统允许对相位信息进行建模并且避免了固定的时频变换。由于音频的高采样率，在样本级别上使用长距离输入上下文是困难的，但高质量的分离结果又要求模型可以捕捉到语音的长距离依赖信息。在此背景下，Wave-U-Net [24]被提出，它是 U-Net 架构在一维时域上的一种适应，通过计算和组合不同时间尺度上的特征来进行重复重采样特征映射。歌唱声音分离的实验表明，在相同的数据条件下，Wave-U-Net 的性能可与基于声谱图的 U-Net 体系结构相媲美。

音乐源分离可以理解为声源分离技术的一个子领域，音乐源分离的发展与声源分离的发展息息相关。近几年的声源分离研究已经证明，直接在时域操作的端到端系统可以成功地优于基于频域的系统。在声源分离领域中的一个里程碑式的成果就是罗毅博士 2018 年发布的 Conv-Tasnet [9]架构，由于声源分离与音乐源分离的相似性，这个架构被广泛借鉴用于音乐源分离。

Conv-Tasnet 继承了 Tasnet [25]编码器、分离模块、解码器的组成部分。其中编码器由一个一维卷积块组成，分离模块是一个时间卷积网络，由堆叠的卷积块组成，同时伴随着跳跃连接与残差连接，最后使用 sigmoid 激活函数为每个分离源生成一个掩膜。随着卷积层数的不断增加，时间卷积网络的感受野也不断扩大，最终得以捕获足够多的语音信号上下文信息。同时 Conv-Tasnet 直接使用评估指标 SISDR 作为损失函数。

Facebook 人工智能研究院在 2019 年提出了一种端到端弱监督训练模型，使用 U-Net 结构和双向 LSTM，名为 Demucs [26]，并在 MUSDB18 数据集上进行了实验验证，实验结果表明，Demucs 要明显优

于 Conv-TasNet 的分离效果。

音源分离的在时域的第二个里程碑式突破是 DPRNN [27]的发布，与 Conv-TasNet 对长语音信号的处理方式不同，DPRNN 通过组织两个 RNN 层来建模极长序列。DPRNN 将语音长序列分割成较小的块儿，并迭代得运用块内和块间操作，其中每次操作的输入长度与原始序列长度的平方根成正比。与 TasNet 等基于 CNN 的架构(由于感受野固定，只能进行局部建模)相比，DPRNN 能够通过块间的 RNN 充分利用全局信息，并在更小的模型尺寸下获得更好的性能。

Svoice [28]由 Facebook 团队提出的 DPRNN 改进版本，引入了多尺度损失，把单个双向 LSTM 换成了平行的两个 LSTM(类似于多头注意力机制)，同时放弃了 DPRNN 中基于掩膜的分离方法。Svoice 采用门控神经网络，经过训练后可以处理多个说话人同时说话的混源，同时保持每个说话人在固定的输出通道中。Svoice 虽然主要用来做人声分离，但在音乐源分离中也取得了不错的效果。

5. 数据集

一个好的训练数据集对于以数据驱动的音源分离模型是至关重要的。音源分离模型的训练数据不只是由一组歌曲组成，还需要包括构成它们的伴奏和其他子源(如，贝斯、吉他、鼓等等)。而在专业制作的音乐中，这些子源往往没办法获得或者受到版权限制，而且也很难在不侵犯版权的情况下找到与之相似的子源。另一个难点是音乐的生成有一定的规律性，并不只是简单地将这些子源随机叠加。因此，迄今为止，还没有模拟数据集能很好地模拟真实音乐场景[29]。

在这种情况下，歌唱声和伴奏分离数据集的开发是一个漫长的过程。在早期，研究人员用一些私人数据来进行测试。第一次尝试发布一个公共数据集来评估人声和伴奏分离是音乐音频信号分离数据集(MASS)。MASS 的发布有力地推动了音乐源分离领域的研究，即使它只有 2.5 分钟的数据[29]。

MASS 数据集是早期信号分离评估运动(SiSEC)的核心内容，用来评估各种音乐源分离方法的性能。虽然 MASS 数据集的数据量较小，但是因为技术的限制，人们在很长一段时间里都没办法做出较大的突破。近些年来，随着技术的进步，新的数据集也被不断地提出，并在许多方面改进了 MASS 数据集。下面将简要介绍几个重要数据集。

5.1. MUSDB18

MUSDB18 包含 150 首完整长度的不同类型(爵士乐、电子音乐、金属音乐等等)的音乐曲目(约 10 小时时长)，以及被分解后的独立声音，包括鼓、贝斯、歌唱声和其他声音。数据集被分成训练集和测试集，分别包含 100 首和 50 首歌曲。所有歌曲都是采样率为 44.1 kHz 的立体声。

5.2. MIR-1K [30]

MIR-1K 包含有 1000 首歌曲片段，音乐伴奏和歌唱声分别录制为左右声道，声音的各种属性信息都经过人工标注。MIR-1K 从 5000 首中国流行歌曲中随机选择出 110 首歌曲，由 8 名女性和 11 名男性业余歌手进行演唱。然后从这 110 首歌曲中提取出总长度为 133 分钟的片段，包含一个混合音轨和一个音乐伴奏音轨，每个片段的时长从 4 秒到 13 秒不等。

5.3. MedleyDB [31]

MedleyDB 是一个多声部音乐数据集。MedleyDB 的开发主要是为了支持旋律提取研究。对于每首歌曲的旋律和乐器都进行了标注，以评估自动乐器识别任务。原始数据集包含 122 首多音轨歌曲，其中 108 首包含旋律标注。MedleyDB 包含各种音乐类型的歌曲：古典，摇滚，民谣，爵士，流行，音乐剧，说唱。所有类型的音频文件都是采样率为 44.1 kHz 的 wav 文件。

6. 结论

本文对音乐源分离在深度学习上的近期发展做了系统性总结，分析上述深度学习模型可以发现，基于频域和时域的方法的共同点是它们在重建目标信号之前隐式计算应用于混合信号的源依赖掩膜。因为基于时域的方法是使用时域信号进行优化的，时域信号包含携带重要信号信息的相位信息，而许多基于频域的方法忽略了这些信号，理论上基于时域的方法将显著优于频域的方法。然而，实验证据表明，基于频域的方法与基于时域的方法相比，具有相当或略好的分离性能。很明显，两种不同方法的性能差异可以归因于所使用的信号表示。因此，学习音乐信号的广义信号表示也是音乐源分离研究的一个有趣的方向。

参考文献

[1] Bahmaninezhad, F., Wu, J., Gu, R., Zhang, S.-X., Xu, Y., Yu, M. and Yu, D. (2019) A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation. *Interspeech* 2019, 20*th Annual Conference of the International Speech Communication Association*, Graz, 15-19 September 2019, 4574-4578. https://doi.org/10.21437/Interspeech.2019-3181

[2] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *Wiley Interdisciplinary Reviews*: *Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[3] Abdali, S. and NaserSharif, B. (2017) Non-Negative Matrix Factorization for Speech/Music Separation Using Source Dependent Decomposition Rank, Temporal Continuity Term and Filtering. *Biomedical Signal Processing and Control*, **36**, 168-175. https://doi.org/10.1016/j.bspc.2017.03.010

[4] Ozerov, A., Vincent, E. and Bimbot, F. (2012) A General Flexible Framework for the Handling of Prior Information in Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**, 1118-1133. https://doi.org/10.1109/TASL.2011.2172425

[5] Pons, J., Janer, J. and Rode, T. (2016) Remixing Music Using Source Separation Algorithms to Improve the Musical Experience of Cochlear Implant Users. *Journal of the Acoustical Society of America*, **140**, 4338-4349. https://doi.org/10.1121/1.4971424

[6] Heo, W.H., Kim, H. and Kwon, O.W. (2020) Source Separation Using Dilated Time-Frequency DenseNet for Music Identification in Broadcast Contents. *Applied Sciences-Basel*, **10**, Article No. 1727. https://doi.org/10.3390/app10051727

[7] Stöter, F.-R., Liutkus, A. and Ito, N. (2018) The 2018 Signal Separation Evaluation Campaign. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-93764-9_28

[8] Jao, P.-K., Su, L., Yang, Y.-H. and Wohlberg, B. (2016) Monaural Music Source Separation Using Convolutional Sparse Coding. *IEEE-ACM Transactions on Audio Speech and Language Processing*, **24**, 2158-2170. https://doi.org/10.1109/TASLP.2016.2598323

[9] Luo, Y. and Mesgarani, N. (2019) Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**, 1256-1266. https://doi.org/10.1109/TASLP.2019.2915167

[10] Brandenburg, K. and Sporer, T. (1992) NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria. *Audio Engineering Society Conference*: 11*th International Conference*: *Test & Measurement*, Portland, Oregon, 29-31 May 1992, Paper No. 11-020. https://www.aes.org/e-lib/online/browse.cfm?elib=6276

[11] Emiya, V., Vincent, E., Harlander, N. and Hohmann, V. (2011) Subjective and Objective Quality Assessment of Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**, 2046-2057. https://doi.org/10.1109/TASL.2011.2109381

[12] Cano, E., FitzGerald, D. and Brandenburg, K. (2016) Evaluation of Quality of Sound Source Separation Algorithms: Human Perception vs Quantitative Metrics. 2016 24*th European Signal Processing Conference* (*EUSIPCO*), Budapest, 29 August-2 September 2016, 1758-1762. https://doi.org/10.1109/EUSIPCO.2016.7760550

[13] Févotte, C., Gribonval, R. and Vincent, E. (2005) BSS_EVAL Toolbox User Guide—Revision 2.0.

[14] Roux, J.L., Wisdom, S., Erdogan, H. and Hershey, J.R. (2019) SDR—Half-Baked or Well Done? *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 626-630. https://doi.org/10.1109/ICASSP.2019.8683855

[15] ITU (2014) Recommendation ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level

of Audio Systems.

[16] 赵毅. 空间音频编码及多声道音频恢复技术研究[D]: [硕士学位论文]. 北京: 北京理工大学, 2015.

[17] Rafii, Z., *et al*. (2017) MUSDB18—A Corpus for Music Separation.

[18] Liu, H., Kong, Q. and Liu, J. (2021) CWS-PResUNet: Music Source Separation with Channel-Wise Subband Phase-Aware Resunet.

[19] Diakogiannis, F.I., Waldner, F., Caccetta, P. and Wu, C. (2020) ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

[20] Takahashi, N., Goswami, N. and Mitsufuji, Y. (2018) Mmdenselstm: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation. 2018 16*th International Workshop on Acoustic Signal Enhancement* (*IWAENC*), Tokyo, 17-20 September 2018, 106-110. https://doi.org/10.1109/IWAENC.2018.8521383

[21] Iandola, F., *et al*. (2014) Densenet: Implementing Efficient Convnet Descriptor Pyramids.

[22] Stöter, F.-R., Uhlich, S., Liutkus, A. and Mitsufuji, Y. (2019) Open-Unmix—A Reference Implementation for Music Source Separation. *Journal of Open Source Software*, **4**, Article No. 1667. https://doi.org/10.21105/joss.01667

[23] Lluís, F., Pons, J. and Serra, X. (2018) End-to-End Music Source Separation: Is It Possible in the Waveform Domain? *Interspeech* 2019, 20*th Annual Conference of the International Speech Communication Association*, Graz, 15-19 September 2019, 4619-4623. https://doi.org/10.21437/Interspeech.2019-1177

[24] Stoller, D., Ewert, S. and Dixon, S. (2018) Wave-u-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation.

[25] Luo, Y. and Mesgarani, N. (2018) Tasnet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, 15-20 April 2018, 696-700. https://doi.org/10.1109/ICASSP.2018.8462116

[26] Défossez, A., *et al*. (2019) Demucs: Deep Extractor for Music Sources with Extra Unlabeled Data Remixed.

[27] Luo, Y., Chen, Z. and Yoshioka, T. (2020) Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 46-50. https://doi.org/10.1109/ICASSP40776.2020.9054266

[28] Nachmani, E., Adi, Y. and Wolf, L. (2020) Voice Separation with an Unknown Number of Multiple Speakers. *Proceedings of the* 37*th International Conference on Machine Learning*, Vienna, 12-18 July 2020, 7164-7175.

[29] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S.I., FitzGerald, D. and Pardo, B. (2018) An Overview of Lead and Accompaniment Separation in Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**, 1307-1335. https://doi.org/10.1109/TASLP.2018.2825440

[30] Hsu, C.-L. and Jang, J.-S.R. (2010) On the Improvement of Singing Voice Separation for Monaural Recordings Using the Mir-1k Dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**, 310-319. https://doi.org/10.1109/TASL.2009.2026503

[31] Bittner, R.M., *et al*. (2014) MedleyDB: A Multitrack Dataset for Annotation-Intensive Mir Research. 15*th International Society for Music Information Retrieval Conference* (*ISMIR* 2014), Taipei, 27-31 October 2014, 155-160.