

权利要求书

- 1、一种声音分离方法，包括：
获取输入声音频谱，所述输入声音频谱中包括对应于多种声源的声音频谱；
对所述输入声音频谱进行频谱分离处理，由所述输入声音频谱中分离出预测声音频谱；
5 从所述输入声音频谱中去除所述预测声音频谱，得到更新后的输入声音频谱；
通过所述更新后的输入声音频谱，继续获得下一个分离的预测声音频谱，直至更新后的输入声音频谱未包含声音频谱。
- 2、根据权利要求1所述的方法，其特征在于，对所述输入声音频谱进行频谱分离处理，由所述输入声音频谱中分离出预测声音频谱，包括：
10 获取所述输入声音频谱对应的输入视频帧，其中，所述输入视频帧中包含所述多个声源；
根据所述输入视频帧对所述输入声音频谱进行频谱分离处理，由所述输入声音频谱中分离出所述预测声音频谱。
- 3、根据权利要求2所述的方法，其特征在于，根据所述输入视频帧对所述输入声音频谱进行频谱分离处理，由所述输入声音频谱中分离出所述预测声音频谱，包括：
15 根据所述输入声音频谱，得到k个基本分量，所述k个基本分量分别表示所述输入声音频谱中的不同声音特征，所述k为自然数；
根据所述输入视频帧，得到视觉特征图，所述视觉特征图包括多个k维的视觉特征向量，每个视觉特征向量对应所述输入视频帧中的一个声源；
根据其中一个所述视觉特征向量和所述k个基本分量，获得所述预测声音频谱，所述预测声音
20 频谱的声源是所述视觉特征向量对应的声源。
- 4、根据权利要求3所述的方法，其特征在于，根据所述输入视频帧，得到所述视觉特征图，包括：
将所述输入视频帧输入特征提取网络，输出所述输入视频帧的视频特征；
对所述视频特征在时间维度进行最大池化，得到包含所述多个视觉特征向量的所述视觉特征图。
25 5、根据权利要求3所述的方法，其特征在于，根据其中一个所述视觉特征向量和所述k个基本分量，获得所述预测声音频谱，包括：
将所述k个基本分量和其中一个所述视觉特征向量中的k维元素，分别相乘后加和，得到所述预测声音频谱。
- 6、根据权利要求3所述的方法，其特征在于，根据其中一个所述视觉特征向量和所述k个基本
30 分量，获得所述预测声音频谱，包括：
将所述k个基本分量和其中一个所述视觉特征向量中的k维元素，分别相乘后加和；
对加和结果进行非线性激活处理，获得预测掩模；
将所述预测掩模和首次迭代时初始的输入声音频谱进行点乘，得到所述预测声音频谱。
- 7、根据权利要求3所述的方法，其特征在于，根据其中一个所述视觉特征向量和所述k个基本
35 分量，获得所述预测声音频谱，包括：

由所述多个视觉特征向量中，随机选取一个视觉特征向量；

根据选取出的视觉特征向量和所述k个基本分量，获得所述预测声音频谱。

8、根据权利要求3所述的方法，其特征在于，根据其中一个所述视觉特征向量和所述k个基本分量，获得所述预测声音频谱，包括：

5 由所述多个视觉特征向量中，选取音量最大的声源所对应的所述视觉特征向量；

根据选取出的视觉特征向量和所述k个基本分量，获得所述预测声音频谱。

9、根据权利要求8所述的方法，其特征在于，选取音量最大的声源所对应的所述视觉特征向量，包括：

将所述多个视觉特征向量中的每一个视觉特征向量，执行如下处理：

10 所述视觉特征向量与由所述k个基本分量组成的向量相乘，得到第一相乘结果；

对非线性激活后的第一相乘结果与首次迭代的初始输入声音频谱相乘，得到第二相乘结果；

求取所述第二相乘结果的平均能量；

选取平均能量最大值位置对应的视觉特征向量。

15 10、根据权利要求1~9任一所述的方法，其特征在于，由所述输入声音频谱中分离出预测声音频谱之后，所述方法还包括：

根据所述预测声音频谱和历史累计频谱，得到余量掩模，所述历史累计频谱是在所述声音分离的过程中，分离出的历史预测声音频谱的加和；

基于所述余量掩模和所述历史累计频谱，得到余量频谱；

将所述余量频谱和所述预测声音频谱相加，得到完整预测声音频谱。

20 11、根据权利要求10所述的方法，其特征在于，所述历史预测声音频谱的加和包括历史完整预测声音频谱的加和；

从所述输入声音频谱中去除所述预测声音频谱，得到所述更新后的输入声音频谱，包括：

从所述输入声音频谱中去除所述完整预测声音频谱，得到所述更新后的输入声音频谱。

12、根据权利要求10所述的方法，还包括：

25 根据所述完整预测声音频谱与频谱真实值之间的误差，调整第一网络、第二网络和第三网络中的至少任一网络的网络参数，其中，所述输入声音频谱通过所述第一网络，得到k个基本分量；与所述输入声音频谱对应的输入视频帧通过所述第二网络，得到视觉特征图，其中所述输入视频帧中包括所述多个声源；所述预测声音频谱和所述历史累计频谱通过第三网络，得到所述余量掩模。

13、根据权利要求1~12任一所述的方法，还包括：

30 当所述更新后的输入声音频谱的平均能量小于一个预设阈值时，确定所述输入声音频谱未包含声音频谱。

14、一种声音分离装置，包括：

输入获取模块，用于获取输入声音频谱，所述输入声音频谱中包括对应于多种声源的声音频谱；

35 频谱分离模块，用于对所述输入声音频谱进行频谱分离处理，由所述输入声音频谱中分离出预测声音频谱；通过更新后的输入声音频谱，继续获得下一个分离的预测声音频谱，直至更新后的输

入声音频谱未包含声音频谱;

频谱更新模块, 用于从所述输入声音频谱中去除所述预测声音频谱, 得到所述更新后的输入声音频谱。

15、根据权利要求 14 所述的装置, 其特征在于, 所述频谱分离模块, 包括:

5 视频处理子模块, 用于获取所述输入声音频谱对应的输入视频帧, 其中, 所述输入视频帧中包含所述多个声源;

声音分离子模块, 用于根据所述输入视频帧对所述输入声音频谱进行频谱分离处理, 由所述输入声音频谱中分离出所述预测声音频谱。

16、根据权利要求 15 所述的装置, 其特征在于,

10 所述视频处理子模块, 用于根据所述输入视频帧, 得到视觉特征图, 所述视觉特征图包括多个 k 维的视觉特征向量, 每个视觉特征向量对应所述输入视频帧中的一个声源;

所述声音分离子模块, 用于: 根据所述输入声音频谱得到 k 个基本分量, 所述 k 个基本分量分别表示所述输入声音频谱中的不同声音特征, 所述 k 为自然数; 以及根据其中一个所述视觉特征向量和所述 k 个基本分量, 获得所述预测声音频谱, 所述预测声音频谱的声源是所述视觉特征向量对应的声源。

15

17、根据权利要求 16 所述的装置, 其特征在于, 所述视频处理子模块, 用于:

将所述输入视频帧输入特征提取网络, 输出所述输入视频帧的视频特征;

对所述视频特征在时间维度进行最大池化, 得到包含所述多个视觉特征向量的所述视觉特征图。

18、根据权利要求 16 所述的装置, 其特征在于, 所述声音分离子模块, 用于:

20 将所述 k 个基本分量和其中一个所述视觉特征向量中的 k 维元素, 分别相乘后加和, 得到所述预测声音频谱。

19、根据权利要求 16 所述的装置, 其特征在于, 所述声音分离子模块, 用于:

将所述 k 个基本分量和其中一个所述视觉特征向量中的 k 维元素, 分别相乘后加和;

对加和结果进行非线性激活处理, 获得预测掩模;

25 将所述预测掩模和首次迭代时初始的输入声音频谱进行点乘, 得到所述预测声音频谱。

20、根据权利要求 16 所述的装置, 其特征在于, 所述声音分离子模块, 用于:

由所述多个视觉特征向量中, 随机选取一个视觉特征向量;

根据选取出的视觉特征向量和所述 k 个基本分量, 获得所述预测声音频谱。

21、根据权利要求 16 所述的装置, 其特征在于, 所述声音分离子模块, 用于:

30 由所述多个视觉特征向量中, 选取音量最大的声源所对应的所述视觉特征向量;

选取出的视觉特征向量和所述 k 个基本分量, 获得所述预测声音频谱。

22、根据权利要求 21 所述的装置, 其特征在于, 所述声音分离子模块, 用于:

将所述多个视觉特征向量中的每一个视觉特征向量, 执行如下处理:

所述视觉特征向量与由所述 k 个基本分量组成的向量相乘, 得到第一相乘结果;

35 对非线性激活后的第一相乘结果与首次迭代的初始输入声音频谱相乘, 得到第二相乘结果;

求取所述第二相乘结果的平均能量；

选取平均能量最大值位置对应的视觉特征向量。

23、根据权利要求 14~22 任一所述的装置，还包括：频谱调整模块，用于：

5 根据所述预测声音频谱和历史累计频谱，得到余量掩模，所述历史累计频谱是在所述声音分离的过程中，分离出的历史预测声音频谱的加和；

基于所述余量掩模和所述历史累计频谱，得到余量频谱；

将所述余量频谱和所述预测声音频谱相加，得到完整预测声音频谱。

24、根据权利要求 23 所述的装置，其特征在于，所述频谱更新模块，用于：

10 从所述输入声音频谱中去除所述完整预测声音频谱，得到所述更新后的输入声音频谱；其中，所述历史预测声音频谱的加和包括历史完整预测声音频谱的加和。

25、根据权利要求 14~24 任一所述的装置，其特征在于，所述频谱分离模块，用于：

当所述更新后的输入声音频谱的平均能量小于一个预设阈值时，确定所述输入声音频谱未包含声音频谱。

26、一种电子设备，其特征在于，所述设备包括存储器、处理器，所述存储器用于存储可在处
15 理器上运行的计算机指令，所述处理器用于在执行所述计算机指令时实现权利要求 1 至 13 任一所述的方法。

27、一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述程序被处理器执行时实现权利要求 1 至 13 任一所述的方法。

28、一种计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求 1 至 13 任一
20 所述的方法。