(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115497495 B
(45)授权公告日 2025.01.21

(21)申请号 202211145279.3
(22)申请日 2022.09.20
(65)同一申请的已公布的文献号
申请公布号 CN 115497495 A
(43)申请公布日 2022.12.20
(30)优先权数据
21204004.2 2021.10.21 EP
(73)专利权人 汇顶科技(香港)有限公司
地址 中国香港上水龙琛路39号上水广场21楼2113室
(72)发明人 亚历山大·博伦德 奈尔什·马杜安·埃尔维尔·斯普利特 沃特·朱斯·蒂里
(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
专利代理师 田玉珺 毛威
(51)Int.Cl.
G10L 21/0216(2013.01)
G10L 25/06(2013.01)
G10L 25/30(2013.01)
G01S 5/22(2006.01)
(56)对比文件
CN 109313909 A,2019.02.05
CN 110517705 A,2019.11.29
审查员 何元

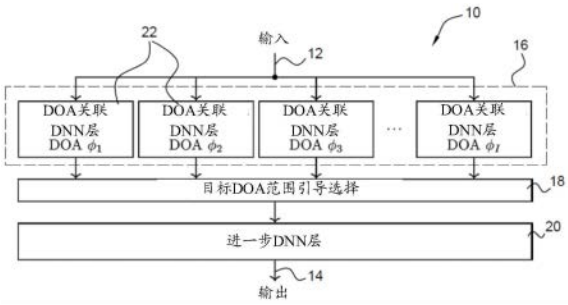
权利要求书2页 说明书11页 附图7页

(54)发明名称

用于检测或估计多个声源中的目标声源的方法和装置

(57)摘要

本申请公开了一种用于检测或估计多个声源中的目标声源的方法、装置和计算机程序,所述目标声源位于多个空间区域中的至少一个目标空间区域,所述方法包括:处理器接收多个信号,所述多个信号中的每一个与多个麦克风信号中的一个相关联,所述麦克风信号包括由所述多个声源产生的声音事件;处理器通过神经网络提取多个特征,所述多个特征是从所述多个信号中提取的,并通过针对所述多个空间区域中的不同空间区域训练所述神经网络,获得所述多个特征中的每一个;所述处理器基于所提取的多个特征生成对应于所述至少一个目标空间区域的另外多个特征;所述处理器基于所述另外多个特征,检测或估计所述目标空间区域中所述目标声源中的至少一个声源。



CN 115497495 B

1.一种用于检测或估计多个声源中的目标声源的方法,其中所述目标声源位于多个空间区域中的至少一个目标空间区域,所述方法包括:

处理器接收多个信号,其中所述多个信号中的每一个与多个麦克风信号中的一个相关联,其中所述麦克风信号包括由所述多个声源产生的声音事件;

所述处理器通过神经网络提取多个特征,其中所述多个特征是从所述多个信号中提取的,并且其中通过针对所述多个空间区域中的不同空间区域训练所述神经网络,获得所述多个特征中的每一个;

所述处理器基于所述神经网络以及所提取的多个特征生成另外多个特征,其中所述另外多个特征对应于所述至少一个目标空间区域;

所述处理器基于与所述至少一个目标空间区域对应的所述另外多个特征,检测或估计所述目标空间区域中所述目标声源中的至少一个声源。

2.根据权利要求1所述的方法,其中,所述多个信号中的每一个与麦克风时域信号的时频表示的多个相位相关联。

3.根据权利要求1或2所述的方法,其中,所述多个信号中的每一个包括麦克风时域信号的时频表示的多个归一化振幅中的至少一个,其中所述归一化基于麦克风时域信号的多个时频表示中的至少一个的范数。

4.根据权利要求1所述的方法,其中,所述多个信号中的每一个与麦克风时域信号的多个复时频表示相关联。

5.根据权利要求1或2所述的方法,其中,所述多个特征是通过多个卷积神经网络层中的至少一个从所述多个信号中提取的。

6.根据权利要求1或2所述的方法,其中,所述多个空间区域中的每一个包括多个到达方向。

7.根据权利要求1或2所述的方法,还包括:

所述处理器基于所述另外多个特征估计时频TF掩码,并且基于所述TF掩码在所述至少一个目标空间区域中检测或估计所述目标声源中的所述至少一个声源。

8.根据权利要求1或2所述的方法,还包括:

所述处理器基于所述另外多个特征估计所述至少一个声源的时域表示,并且基于所述时域表示检测或估计所述至少一个声源。

9.根据权利要求1或2所述的方法,还包括:

所述处理器基于所述另外多个特征估计所述至少一个声源的时频表示,并且基于所述时频表示检测或估计所述至少一个声源。

10.根据权利要求1或2所述的方法,还包括:

所述处理器基于所述另外多个特征估计波束成形权重,并且基于所述波束成形权重检测或估计所述至少一个声源。

11.根据权利要求1或2所述的方法,其中,所述另外多个特征是基于所述所提取的多个特征的加权组合生成的。

12.根据权利要求11所述的方法,其中,生成所述所提取的多个特征的所述加权组合包括:

为所提取的多个特征中的每一个生成权重值,其中,所提取的多个特征中的特征的所

述权重值表示所述特征是否对应于所述目标空间区域。

13.根据权利要求1或2所述的方法,其中,所述另外多个特征是通过对所提取的多个特征进行加权并取加权后的所提取的多个特征中的最大值生成的。

14.一种用于检测或估计多个声源中的目标声源的装置,其中,所述目标声源位于多个空间区域中的至少一个目标空间区域中,所述装置包括存储器和与所述存储器通信地连接的处理器,并且用于执行指令以执行根据权利要求1至13中任一项所述的方法。

15.一种计算机程序产品,其中,包括计算机程序,用于执行根据权利要求1至13中任一项所述的方法。

用于检测或估计多个声源中的目标声源的方法和装置

[0001] 本申请要求于2021年10月21日提交欧洲专利局、申请号为EP21204004.2的欧洲专利申请的优先权,其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及一种基于神经网络的音频处理系统,该系统用于通过使用空间相关特征提取技术,估计或检测位于定义的、连续的空间区域内的声源产生的信号,涉及一种通过使用空间相关特征提取技术,估计或检测位于定义的、连续的空间区域内的声源产生的信号的方法,以及涉及一种由处理器执行并且包括实现该方法的指令的程序。

背景技术

[0003] 音频提取的问题通常使用诸如单麦克风稳态噪声抑制或多麦克风设置等技术来解决,以从目标信号中去除不想要的信号。所提取的需要的信号用于语音增强、语音识别、音频转录和许多其他基于音频的应用。

[0004] 在短时傅里叶变换(short-time Fourier transform,STFT)域中表现出高度稀疏的信号,例如语音或音乐,可以借助时频(time-frequency,TF)掩码有效地分离,该掩码可以识别其中目标信号占主导地位的STFT点(bins)。为了获得这种掩码,通常采用深度学习方法。TF掩码是用于语音增强和定位的音频处理中的已知工具,其允许监督学习。TF掩码包括关于目标源在短时傅里叶变换(STFT)域中每个时间和频率的活动的信息。通常,TF掩码的每个元素都是介于0和1之间的一个数字。TF掩码也可以是复数。TF掩码可用于恢复目标源的相应的干净源信号,例如,通过掩码和参考麦克风信号的逐元素相乘,然后进行逆变换。正如所解释的,TF掩码识别由感兴趣信号主导的短时傅里叶变换(STFT)点(bins),并且可以通过利用光谱时间相关性和空间信息来获得。使用深度学习方法可以利用这一点。除了估计TF掩码之外,其他已知的方法是估计感兴趣信号的STFT点(bins),估计感兴趣的时域信号或估计应用于多通道麦克风录音的多通道波束成形器权重。

[0005] 为了将深度神经网络(deep neural network,DNN)应用于该项任务,选择可以识别目标分量的判别特征是一个关键方面。对于单个麦克风捕获,可以通过利用语音的频谱时间特性中的典型结构来分离例如语音和背景噪声。然而,当目标和干扰属于同一类型(例如,将两个不同的说话者分离)时,这种方法将会失败。当多个麦克风捕获可用时,可以额外利用空间信息来区分具有相似光谱时间属性的源。

[0006] 无论使用哪种信息,明确定义目标分类都至关重要,这样DNN才能学会正确抑制不想要的分量,同时保留目标。例如,对于语音与背景噪声的分离,可以基于频谱时间(语音/非语音)和/或空间(局部/扩散)信号特性进行区分。在神经网络的训练过程中,可以将这些判别特性结合起来。

[0007] 对于多个局部声源的分离,可以基于源的位置进行区分,这可以例如就对应的到达方向(directions of arrival,DOA)进行描述。然而,由于任一源(例如,同时说话的多个谈话者中的一个)可被认为是目标,因此需要用户输入来解决目标观察方向上的这种模糊

性。通常的方法是训练网络以从各个方向提取信号分量。然后可以通过选择多个输出流(例如,目标跟踪波束中的所有流)中的一个或多个输出流的组合来获得想要的信号。在这种方法中,在没有每个流的相关性的先验知识的情况下,生成所有流的最优结果,然后只使用这些流的小部分重建目标信号,即DNN输出包含了大量冗余。因此,这种方法无法有效地使用可用的计算资源。另一种方法是针对固定的、先验已知的空间角度训练神经网络。然而,这种方法无法在运行时选择角度或空间,因此并不灵活。通常也考虑单个角度。

[0008] “使用全频带和子频带空间特征的神经网络进行基于掩码的源分离”由A·博伦德(A.Bohlender)、A·斯普里耶(A.Spriet)、W·提利(W.Tirry)和N·马杜(N.Madhu)在《欧洲信号处理国际会议(EUSIPCO)》,2021年中发表,其公开了使用深度神经网络架构(由卷积层(convolutional,CNN)和循环层组成)估计TF掩码,如图1所示。图1所示的CNN的输入由N个麦克风信号的相位谱图组成。由于同时估计了所有到达方向(DOA)的掩码,因此对于总共有I个DOA的离散网格来说,每个TF点(bin)的输出大小等于I。因此,每个返回的I个TF掩码对应一个特定的方向。为了提取单个目标,只需要与目标源DOA对应的掩码即可,可以丢弃所有其他估计的掩码。因此,该方法的计算能力没能得到有效利用,因为计算了不会使用的TF掩码。此外,性能可能不太理想,因为所有角度都需要同时到达最低性能。

[0009] “基于时频掩蔽的卷积递归神经网络在线多通道语音增强”由S·查克拉巴蒂(S.Chakrabarty)和E.A.P·哈贝茨(E.A.P.Habets)在《IEEE信号处理专题》第13卷,第4期,第787-799页,2019年8月,DOI:10.1109/JSTSP中发表,其公开了一种与图1所示类似的用于掩码估计的CNN架构。但是,仅单个TF掩码返回,因为此方法仅考虑将单个局部语音源与分散且不相关的背景噪声分离。在这种情况下,仅区分局部和非局部信号分量就足够了,目标源的特定位置不起作用。然而,这种方法不适合分离多个局部源。

[0010] “多通道语音增强的基于特征向量的语音掩码估计”由L·比芬博格(L.Pfeifenberger)、M·泽勒(M. Zöhrer)和F·彭科夫(F.Pernkopf)在《IEEE/ACM音频、声学、语言信号处理》,第27卷,第12期,第2162-2172页,2019年12月,DOI:10.1109/TASLP.2019.2941592中发表,其公开了使用神经网络架构进行TF掩码估计,其中神经网络的输出包括总共三个掩码,分别对应于目标源、干扰源和背景噪声。通过对明确定义的场景训练网络,可以解决关于哪个源被视为目标的模糊性。图2示出了一种可能的训练场景的示例,其中一个源的位置被限制在特定区域D1,而另一个源的位置被限制在不同的区域D2。但是,这种方法并不通用,因为目标DOA范围必须事先指定,并且在训练神经网络后无法由用户控制,因为它对源的位置和移动做出假设,以明确定义哪个源是目标,哪个是干扰源。

[0011] “利用从空间特征预测的时频掩模进行远距离语音分离”是由P·佩尔蒂莱(P. Pertilä)和J·尼库宁(J.Nikunen)在《语音通信68(2015)》,第97-106页中发表,其公开了一种方法,其中特别考虑了基于其DOA分离多个源的问题。采用一个简单的前馈神经网络。输入特征基于麦克风信号的相位谱图以及DOA。为了获得适合与麦克风信号相位一起使用的DOA表示,首先计算由沿指定的DOA的理想平面波传播产生的“理论”通道间相位差。然后,作为神经网络的输入,为观察方向和非观察方向确定分数(每个TF点(bin))。通过平均或聚合几个方向的输入分数,该方法可以扩展为允许指定DOA的目标范围。或者,聚合可以基于多个DOA的独立获得的输出。然而,后者由于其计算复杂性而效率不高,而前一种方法也并未达到最佳标准,因为神经网络必须达成妥协才能为所有可能的感兴趣的空间角度提

供合适的性能。

[0012] “具有位置引导语音提取网络的多通道重叠语音识别”由Z·陈(Z.Chen),X·肖(X.Xiao),T·吉冈(T.Yoshioka),H·埃尔多安(H.Erdogan),J·李(J.Li)和Y·龚(Y.Gong)在《2018IEEE口语技术研讨会(SLT)》,2018年,第558-565页,DOI:10.1109/SLT.2018.8639593中发表,“基于方向信息的端到端多通道目标语音分离”由R·顾(R.Gu)和Y·邹(Y.Zou)在《arXiv预印本arXiv:2001.00391》,2020年中发表,两者皆公开了将目标的定义合并到输入特征中,因而缺乏通用性并降低了效率。三种不同类型的输入特征得到了使用,即空间特征(通道间相位差)、频谱特征(波束成形的麦克风信号)和所谓的角度特征(麦克风信号频谱和转向矢量间的余弦距离),其中两个(部分)基于目标说话者的位置。将观察方向加入光谱和角度特征中,以定义目标。所谓的角度特征是一个分数,其量化了麦克风信号与指定的观察方向的对应程度。此外,为了获得特定于该方向的光谱特征,在目标方向上引导波束成形器。因此,这些方法在计算上也是复杂的。此外,由于其必须适用于用户定义的角度,因此需要对用户定义的角度的性能进行折衷,以在所有考虑的角度上实现良好的性能。

发明内容

[0013] 本申请提供一种用于检测或估计多个声源中的目标声源的方法和装置。

[0014] 在第一方面,本申请提供了一种用于检测或估计多个声源中的目标声源的方法,其中所述目标声源位于多个空间区域中的至少一个目标空间区域,所述方法包括:处理器接收多个信号,其中所述多个信号中的每一个与多个麦克风信号中的一个相关联,其中所述麦克风信号包括由所述多个声源产生的声音事件;所述处理器通过神经网络提取多个特征,其中所述多个特征是从所述多个信号中提取的,并且其中通过针对所述多个空间区域中的不同空间区域训练所述神经网络,获得所述多个特征中的每一个;所述处理器基于所提取的多个特征生成另外多个特征,其中所述另外多个特征对应于所述至少一个目标空间区域;所述处理器基于与至少一个目标空间区域对应的所述另外多个特征,检测或估计所述目标空间区域中的所述目标声源中的至少一个声源。

[0015] 可以基于对应于目标空间区域的特征和对应于其他非目标空间区域的特征生成另外多个特征。目标声源可以全部位于单个目标空间区域中,也可以分布在多个空间目标区域中。神经网络可能已经对于多个空间区域中的每一个进行了训练。神经网络可以包括用于多个空间区域中的每一个的空间区域相关层,并且可以对于每个空间区域训练神经网络的空间区域相关层。至少一个目标空间区域可以由用户在运行时通过用户界面或以任何其他合适的方式设置。或者,可以在运行时设置几个目标空间区域。训练空间相关层的多个空间区域可以比目标空间区域覆盖更宽的空间。实际上,多个空间区域可以覆盖用户可以设置的所有可能的空间区域。作为示例,多个空间区域可以对应于覆盖0度到360度的DOA区域的网格,使得例如多个空间区域包括,覆盖包括从0度到10度角度的区域的第一空间区域,覆盖包括从10到20度角度的区域的第二空间区域,覆盖包括从20度到30度角度的区域的第三空间区域,依此类推。

[0016] 多个空间区域可以包括多个到达方向。或者,空间区域可以包括三维体积。例如,三维体积可以定义为多个点,其中每个点由笛卡尔坐标系、椭圆坐标系或球面坐标系中的

一组三维坐标定义。

[0017] 在本申请的实施例中,所述神经网络的所述多个输入信号可以包括所述麦克风信号的所述时频表示的多个相位,和/或麦克风时域信号的时频表示的多个归一化振幅,其中,所述归一化基于麦克风时域信号的多个时频表示和/或麦克风信号的多个复时频表示中的至少一个的范数。

[0018] 在本申请的另一个实施例中,所述多个特征是通过多个卷积神经网络层中的至少一个从所述多个信号中提取的。

[0019] 在本申请的一个实施例中,所述方法可以包括:所述处理器基于所述另外多个特征估计时频TF掩码,并且基于所述TF掩码在所述至少一个目标空间区域检测或估计所述目标声源中的所述至少一个声源。该TF掩码可能具有复数或实数值。TF掩码可以应用于单个通道或多个通道。

[0020] 在本申请的另一个实施例中,所述方法可以包括:所述处理器基于所述另外多个特征估计所述至少一个声源的时域表示,并且基于所述时域表示检测或估计所述至少一个声源。

[0021] 在另一个实施例中,所述方法可以包括:所述处理器基于所述另外多个特征估计所述至少一个声源的时频表示,并且基于所述时频表示检测或估计所述至少一个声源。

[0022] 在另一个实施例中,所述方法可以包括:所述处理器基于所述另外多个特征估计波束成形权重(例如,多通道复掩码或时域滤波器系数),并且基于所述波束成形权重检测或估计所述至少一个声源。该波束成形权重可以应用于多通道信号。

[0023] 可以基于所提取的多个特征的加权组合生成另外多个特征。即,可以基于所提取的多个特征的任何合适的线性组合获得另外多个特征中的每一个。

[0024] 在另一个实施例中,所述另外多个特征是通过对所提取的多个特征进行加权并取加权后的所提取的多个特征中的最大值生成的。

[0025] 在另一个实施例中,加权组合可以基于与所述至少一个目标区域对应的所提取的多个特征的值中的平均值生成。即,可以通过计算与至少一个目标区域对应的所提取的多个特征的所有值中的平均值获得另外多个特征。

[0026] 本申请的另一个方面涉及一种用于检测或估计多个声源中的目标声源的装置,其中所述目标声源位于多个空间区域中的至少一个目标空间区域中,所述装置包括存储器和于所述存储器通信地连接的处理器,并且用于执行指令以执行上述方法。

[0027] 本申请的另一个方面涉及一种计算机程序,用于执行上述方法。

[0028] 通过提供可以集成到DNN或神经网络中的空间,例如DOA、相关特征提取和用户可控空间(例如DOA)选择,无需更改DNN的输入和输出以适应空间(例如DOA)依赖性。

[0029] 通过提供可以由一组目标DOA或空间区域定义的目标区域范围(在运行时)来定义目标源,在提供多功能性和可扩展性的同时,仍然允许分离多个局部源。

[0030] 此外,不需要预先定义目标源的位置,同时仍然避免为每个可能的场景训练单独的DNN,从而在节省资源的同时提供灵活性。

[0031] 在本申请中,可以只为所有目标分量估计一个掩码,而非为I个DOA中的每一个或空间相关特征估计一个掩码,从而有效地使用计算和内存资源。此外,由于最后的神经网络层只需要关注单个空间区域,因此有望提高性能。注意本申请不限于掩码估计。除了预测掩

码,网络可以估计目标信号的时域或时频域表示。或者,网络可以估计波束成形器系数。

[0032] 本申请允许DNN自行学习如何根据指定的空间区域或DOA处理输入数据的每个部分。通过这种方法,空间或DOA信息固有地与其他类型的信息相结合,而无需手工特征。

[0033] 要求保护的本申请在神经网络内执行降维(例如,通过逐元素最大运算)。因此,DNN可以更好地控制要丢弃的信息。

[0034] 要求保护的本申请还允许加入一系列空间区域或DOA的信息,从而允许更好地利用可用信息。这将改善源质量,并能更好地抑制噪声和干扰。

[0035] 通过要求保护的空间区域相关特征提取,神经网络可以自行学习如何将任何类型的输入特征与空间区域信息相结合。

[0036] 显然,其他方法对于本领域技术人员来说是可用的并且是已知的。

附图说明

[0037] 下面将参照附图更详细地讨论本申请,其中:

[0038] 图1示意性地示出了根据现有技术的基于神经网络的音频处理系统;

[0039] 图2示意性地示出了根据现有技术的基于神经网络的音频处理系统的训练场景;

[0040] 图3示意性地示出了根据现有技术的基于神经网络的另一个音频处理系统;

[0041] 图4A示意性地示出了根据本申请第一实施例的基于神经网络的音频处理系统;

[0042] 图4B示出了根据本申请实施例的空间区域和目标空间区域的示例;

[0043] 图5示意性地示出了根据本申请第二实施例的基于神经网络的音频处理系统;

[0044] 图6A和图6B示出了使用现有技术方法和根据本申请实施例的方法执行的信号估计的模拟结果;

[0045] 图7示出了根据本申请实施例的基于神经网络的音频处理系统中的处理音频方法的流程图。

[0046] 附图仅用于说明目的,并不作为权利要求所规定的范围或保护的限制。

具体实施方式

[0047] 本申请可以在任何多麦克风系统中用于估计和/或检测由位于定义的、连续的空间目标区域内的声源产生的信号。在增强的情况下,这意味着抑制指定目标区域之外的所有源,以及抑制非局部背景噪声。例如,一种应用可以是提取单个目标源,其位置可以使用源定位算法来获取。可以调整目标区域的大小,以控制位置的不确定性。本申请的另一种应用是允许空间音频捕获,其中目标区域的大小由应用(例如,声学缩放)适配/控制。或者,例如,在监测应用中,本申请可以用于检测位于感兴趣的空间区域中的特定声源。因此,本申请可以应用于对来自某个方向的特定声音进行音频增强、源分离以及源活动检测。本申请的其他应用也是可能的。

[0048] 在本申请中,在运行时指定目标空间范围。这可以作为一组附加输入参数加入神经网络,从而明确定义应该从中提取源的区域。

[0049] 这里特别感兴趣的是可以在运行时确定感兴趣的空间区域。为了实现这一点,建议执行空间范围相关的特征提取,其中DNN的第一层使用根据指定目标空间范围的权重。在不失一般性的情况下,下面我们将考虑由到达方向(DOA)表示的空间范围。然而,这并不将

本申请的范围限制为仅基于DOA的目标范围规范。例如,放弃使用由到达方向表示的空间区域,空间中的三维(three dimensional,3D)体积由多个点表示,其中每个点都可以根据笛卡尔坐标、球面坐标或椭圆坐标定义。因此,虽然通过描述可以参考包括DOA的空间区域,但是所有描述的实施例可以替代地通过使用任何其他类型的合适的空间区域来实现,例如3D体积。

[0050] 因此,指定了一个目标DOA范围,所有位于目标DOA范围内的声源都将被视为目标源,而位于目标DOA范围之外的所有声源都将被视为干扰源。对于由此产生的源分离问题,将使用可集成到DNN中的DOA相关特征提取和用户可控的DOA选择。这样,DNN的输入和输出都无需改变以适应DOA依赖性。这样,输入不需要包括包含目标DOA范围信息(例如理论上的通道间相位差)的特征,这些特征与其他特征(即,基于何种信息执行源分离,例如源自麦克风信号的输入特征)相结合(例如,通过沿一个维度连接或通过计算分数)。此外,关于输出,由于单个目标是基于指定的DOA范围定义的,因此相较于同时估计多个目标,可以更有效地利用DNN的计算能力,其中除了表示所需的方向的小子集外,其他所有子集随后都将被丢弃。这样可以节省资源并提高性能。

[0051] 图4A示意性地示出了根据本申请实施例的基于DNN的信号处理系统10。基于DNN的信号处理系统10包括输入12、输出14、第一DNN层16、第二DNN层18和进一步处理DNN层20。

[0052] 第一DNN层16包括多个DOA相关DNN层22,其中每个DOA相关DNN层22与I个离散的DOA中的不同DOA相关(即,I是不同的DOA的数量)。多个DOA相关DNN层22的可训练参数特定于I个离散的DOA中的每一个。由于定义目标取决于所选的DOA范围,因此DNN有望在训练期间自行学习每个DOA的合适的参数。图4B示出了I个离散的DOA区域和目标DOA范围的示例。在图4B的示例中,有I=12个DOA区域$\varphi_1,\varphi_2,\ldots,\varphi_{12}$,总共覆盖360度,其中,12个DOA区域中的每一个覆盖30度。在图4B的示例中,目标DOA范围包括从-30度到30度的范围。正如所解释的,图4B是示例,可以使用以任何其他合适方式布置的任何其他合适数量的DOA区域。例如,网格可以更精细,使得每个DOA区域覆盖例如5或10度。空间区域相关DNN层将使用来自空间区域中任何位置的源进行训练。

[0053] 第二DNN层18包括目标DOA范围引导选择模块,该模块用于聚合对应于指定的目标DOA范围的DOA相关特征,以获得不再包括来自所有DOA的信息的特征表示。这可以通过下述方法实现,例如,通过丢弃用于I个DOA中目标DOA范围之外的DOA的特征(即,将它们与0相乘),然后在I个DOA中对目标DOA范围内的DOA中取最大值(每个特征)。

[0054] 这样,多个DOA相关DNN层22可以接收多个信号,每个信号与包括由声源产生的声音事件的麦克风信号相关联。DOA相关DNN层22可以从多个信号中提取多个特征。第二DNN层18可以从DOA相关DNN层22接收所提取的多个特征,并且基于所提取的多个特征生成另外多个特征,其中另外多个信号对应于至少一个目标空间区域。例如,可以将权重分配给所提取的特征中的每一个,并且另外多个特征可以是基于所提取的多个特征的加权组合生成的。每个权重可以是介于0和1之间的值。在另一个实施例中,每个权重可以是0或1。在一个实施例中,另外多个特征是通过对所提取的多个特征进行加权并且取加权后的所提取的特征中的最大值生成的。在另一个实施例中,加权组合可以通过计算所提取的多个特征的值中的平均值生成。进一步处理DNN层20不再需要执行DOA特定处理,因为第一DNN层16和第二DNN层18已经在DOA之间进行了区分。另一个实施例中可能不存在进一步处理DNN层20。基于DNN

的信号处理系统10可以估计基于另外多个特征计算的TF掩码,并且使用所估计的TF掩码检测或估计位于目标DOA范围内的声源。或者,基于DNN的信号处理系统10可以估计指定目标DOA范围内的声源的时域表示,并且所述时域表示可以用于估计或检测至少一个目标声源。在另一个实施例中,可以基于所述另外多个特征估计时频表示,并且时频表示可以用于估计或检测目标声源。或者,可以估计波束成形权重,使得基于所述波束成形权重检测或估计声源。

[0055] 现在将解释根据本申请的实施例的示例实施方式。$Y_{ref}(\mu,\lambda)$和$S(\mu,\lambda)$分别表示短时傅里叶变换(STFT)域中的参考麦克风信号和目标信号,其中,μ是频率指数,λ是时间框架指数。此外,Φ_t是目标DOA的集合,使得:

[0056] $$\Phi_t \subseteq \{\phi_1, \phi_2, \ldots, \phi_I\} \quad (1);$$

[0057] Φ_S是(实际)源DOA的集合,使得:

[0058] $$\Phi_S \subseteq \{\phi_1, \phi_2, \ldots, \phi_I\} \quad (2);$$

[0059] 其中,$\varphi_1, \varphi_2, \ldots, \varphi_I$是不同的离散DOA,其中I是离散DOA的数量,$\Phi_t$和$\Phi_s$是集合$\varphi_1, \varphi_2, \ldots, \varphi_I$内离散DOA的集合。$\Phi_t \cap \Phi_s$对应于包括同时属于集合$\Phi_t$和集合$\Phi_s$的DOA的离散DOA的集合,并表示目标源DOA的集合,而Φs\(Φt∩Φs)是干扰源DOA的集合,并对应于离散的DOA的集合,其中包括包含在集合$\varphi_1, \varphi_2, \ldots, \varphi_I$中但不在目标源DOA的集合$\Phi_t \cap \Phi_s$中的DOA。

[0060] 图1的CNN的目标输出包括对于所有i=1,...,I的TF掩码$\mathcal{M}_i(\mu,\lambda) \in [0,1]$。对于所有i来说,其中,$\phi_i \notin \Phi_s$,理想的TF掩码是$\mathcal{M}_i(\mu,\lambda) = 0$。使用目标源方向(或估计其$\hat{\mathcal{M}}_i(\mu,\lambda)$)的TF掩码$\mathcal{M}_i(\mu,\lambda)$,得到的目标信号的估计如下:

[0061] $$\hat{S}(\mu,\lambda) = \sum_{i:\phi_i \in \Phi_t \cap \Phi_s} \left(\mathcal{M}_i(\mu,\lambda) \cdot Y_{ref}(\mu,\lambda)\right) \quad (3);$$

[0062] 因此,尽管CNN总共返回I个TF掩码$\hat{\mathcal{M}}_i(\mu,\lambda), i = 1,\ldots,I$,只有与目标源DOA $\varphi_i \in \Phi_t \cap \Phi_s$之一对应的那些实际上需要获得目标信号的估计。

[0063] 为了更有效地利用CNN的计算能力,方程(3)可以改写为

[0064] $$\hat{S}(\mu,\lambda) = \mathcal{M}(\mu,\lambda) \cdot Y_{ref}(\mu,\lambda) \quad (4);$$

[0065] 和

[0066] $$\mathcal{M}(\mu,\lambda) = \sum_{i:\phi_i \in \Phi_t \cap \Phi_s} \mathcal{M}_i(\mu,\lambda) \quad (5)。$$

[0067] 因此,当预先指定Φ_t时,目标输出减少到单个TF掩码$\mathcal{M}(\mu,\lambda)$。

[0068] 对于本申请的一个特定的非限制性实施例,其中,DOA相关特征提取为此集成到图1的TF掩码估计器中,如下图5所示。

[0069] 除了在DNN的输入中加入Φ_t,第一层的参数是基于DOA索引i=1,...,I,从而使该层的输出特征是DOA相关的。DOA相关层的参数(即权重)是通过使用来自DOA索引i的源来训练这些层获得的。随后,针对所有目标DOA $\varphi_i \in \Phi_t$获得的特征的聚合用于进一步处理。这些

特征的聚合是通过对所提取的特征的加权组合得到的。例如,可以通过对所提取的特征进行加权并且取所有加权的所提取的特征中的最大值生成该加权组合。此外,加权组合可以通过计算所有加权的所提取的特征的平均值来获得。其他聚合特征的合适方式也可以使用。

[0070] 为了将图4描述的DOA相关特征提取集成到图1的TF掩码估计的CNN中,从而加入关于由Φ_t指定的目标DOA范围的信息,修改的DNN架构如下面图1所示。

[0071] 在第一层(在这种情况下,为N-1卷积层中的第一层,其中N是麦克风的数量)中引入了DOA相关性。由于DNN的这部分独立地处理与每个M'离散频率对应的子带,直至奈奎斯特频率,因此,这是针对每个$\mu=0,\ldots,M'-1$进行的。在DOA选择后,以下层(包括其他N-2卷积层)可以类似于图1中的那些。更一般地,DOA相关性可以合并到第一个L卷积层,所以只有剩下的N-L-1个卷积层与DOA无关。此外,由于只需要一个TF掩码(而非I个DOA中的每一个有一个TF掩码),输出大小(每个频率和每一帧)从I减至1。

[0072] 图5示意性地示出了根据本申请实施例的基于DNN的信号处理系统40。基于DNN的信号处理系统40包括输入42、输出44、具有DOA相关参数的DNN层46、目标DOA范围引导选择层48、进一步卷积层50和进一步处理DNN层52。

[0073] 具有DOA相关参数的DNN层46包括卷积层54的集合,每个卷积层54包括64个长度为2的过滤器,其仅在通道维度上运行(即,对于第一层,在麦克风信号相位向量的元素上$\angle Y(\mu,\lambda)$)。每个卷积层54可以包括任何合适数量的过滤器,并且每个过滤器可以具有任何合适的长度。每个频率的输入是单独处理的,但滤波器与频率无关(即,可以对所有频率联合执行训练)。在图5中,这通过一些块之间的水平连接指示(—)。通过引入DOA相关参数,每个DOA相关层的过滤器总数增加到(64·I)。

[0074] 目标DOA范围引导选择层48接收具有DOA相关参数46的DNN层的输出,并且从所述输出中丢弃那些DOA并非位于指定的目标DOA范围内的输出(即,那些φ_i不属于目标DOA的集合Φ_t)。为此,所述输出信号中对应于DOA的部分使得$\varphi_i \in \Phi_t$(即,DOA φ_i属于目标DOA的集合Φ_t)可以乘以1,而所述输出信号中对应于DOA的部分使得$\phi_i \notin \Phi_t$(即,这种φ_i不属于目标DOA的集合Φ_t)可以乘以0,从而提供加权的DOA相关特征。这样,可以之后生成DOA相关特征的加权组合。然而,这可以以任何其他合适的方式来执行。实际上,对于对应于$\phi_i \notin \Phi_t$的特征,可以完全跳过DNN层中具有DOA相关参数46的计算,因为无论如何都不需要这些。因为基于DNN的信号处理系统40中的所有进一步处理都与DOA无关,所以之后不再需要DOA维度,因此其大小应减小至1。例如,与所有(选择的)DOA对应的特征的最大值用于进一步处理。

[0075] 另一个卷积层50可以类似于图1所示的第一层的对应部分。其中,在图1的(N-1)个卷积层中,DOA相关参数仅在这些(N-1)个卷积层中的第一个使用,使得以下(N-2)个卷积层以及DNN的其余部分,即进一步卷积层50和进一步处理DNN层52,独立于Φ_t。因此,这些部分的架构可能与图1中的类似(例如,包括FC、展平、堆叠和LSTM层)。然而,在图1中,由于同时估计了所有DOA的TF掩码,因此所有DOA信息仍包含在处理的数据中。因为这不是图4A中的情况,DNN的计算能力得到了更有效的利用。

[0076] DOA相关性由训练数据启用,训练数据由成对的目标DOA范围Φ_t和目标输出

$\mathcal{M}(\mu,\lambda)$组成。对于其他情况不变的场景(即麦克风处的相同信号混合),目标输出可以根据哪些源DOA落在目标DOA范围内而有所不同。

[0077] 本申请不限于图5示出的相位输入。除了(或代替)麦克风信号的时频表示的相位之外,复值时频表示和/或时频表示的归一化振幅可以用作图4和图5中的输入。振幅时频表示的归一化可以基于所有或麦克风信号的子集的时频表示的范数(例如,1-范数或2-范数),如等式(6)所示:

[0078] $$|Y(\mu,\lambda)|_{norm}=\frac{|Y(\mu,\lambda)|}{\left\|\left(Y(i,\lambda),\ldots,Y(j,\lambda)\right)\right\|_{i,\ldots j\in\Delta}} \quad (6)。$$

[0079] 本申请不限于图5示出的架构,本申请已如所解释的集成到该架构中。相反,根据图4描绘的通用框图,它可以很好地集成到任何其他DNN拓扑中,例如基于编码器/解码器的神经网络(即DOA特定编码器层)。此外,可能不存在进一步处理DNN层52。在其他实施例中,除了估计目标输出TF掩码之外,还可以进行其他估计,例如复TF掩码、目标信号的时域或时频表示,或波束成形器系数。

[0080] 以下将列出替代实施例的一些变形,并且也不特定于任何特定的DNN架构,而是可以用于任何合适的DNN。

[0081] 当在基于DNN的信号处理系统40的第一层中将DNN层与DOA相关参数46集成时,可训练的参数数量的增加最小。这是因为对于每个频率和帧,输入由长度为N的单个向量组成,然而,后续卷积层的输入和输出由与滤波器数量(图5中的64个)相同的向量组成。对于不同的架构,可能优选地将具有DOA相关参数46的DNN层集成到基于DNN的信号处理系统的不同层中。

[0082] 本申请的目的是在链的早期消除数据的DOA相关性,从而可以进行所有进一步的处理,而不考虑目标DOA。通过直接基于输入这样做,其余层可以使用与DOA无关的表示,在此基础上可以更有效地使用计算能力(甚至可以允许使用转移学习或其他有效的学习方法)。相比之下,在如图1所示的现有技术中,所有I个DOA的DOA信息仍然包含在数据中,直到最终估计所有I个DOA的掩码。由于实际上只需要这些掩码的小部分子集,因而处理的数据中存在相当大的冗余,因此计算资源没有得到有效利用。

[0083] 尽管如此,也可以将DOA相关处理以不同的形式集成到DNN中。例如,本申请可以集成在网络的更深层次(例如,图5中的全连接(fully-connected,FC)层),其中已经实现了更高程度的抽象。

[0084] 图5的目标DOA范围引导选择层48通过生成所提取的特征的加权组合执行DOA选择,例如,通过取加权的DOA相关特征$I(\mu,\lambda,\varphi_i)$在所有观察方向上的最大值,这可以表示为

[0085] $$O(\mu,\lambda)=\max_{1\le i\le I}\left(w(\phi_i)\cdot I(\mu,\lambda,\phi_i)\right) \quad (7);$$

[0086] 其中,$I(\mu,\lambda,\varphi_i)$是在目标DOA范围引导选择层48接收的矩阵输入,也是具有DOA相关参数46的DNN层的输出,max(•)是一个运算函数,其输出是其输入的最大值并按逐元素应用,权重$w(\varphi_i)$分配给每个所提取的特征,由下式给出

[0087] $$w(\phi_i)=\begin{cases}1,\phi_i\in\Phi_t\\0,\phi_i\notin\Phi_t\end{cases} \quad (8);$$

[0088] 无论Φ_t中元素的数量是多少,为所有DOA $\varphi_i\in\Phi_t$取最大值背后的直觉是Φ_t中包含的任何方向的源活动都应捕获。然而,其他解决方案也是可能的,并且下文解释了一些非限制性替代方案。

[0089] 上面等式(5)中示出的目标输出是所有单个TF掩码的总和。因此,为了生成所提取的特征的加权组合,可以考虑对所有DOA上的加权的所提取的特征求和或计算平均值,而非取最大值。方程(7)概括为

[0090] $$O(\mu,\lambda)=\underset{1\le i\le I}{\mathrm{op}}\left(w(\phi_i)\cdot I(\mu,\lambda,\phi_i)\right) \quad (9);$$

[0091] 其中,op(•)是一些运算,例如求和运算或平均运算。

[0092] 可以采用实值加权代替二进制值w(φ_i)。例如,当Φ_t覆盖一个连贯的DOA范围时,权重可以设置为在中心的最大值w(φ_i)=1,并朝着目标DOA范围的末端单调递减,以实现平稳过渡。或者,关于每个DOA的源活动的软信息,例如,从单独的系统(例如,神经网络)获得的信息,可以直接用于设置权重。

[0093] 最后,针对这种特殊情况下,即Φ_t受到约束,使得$|\Phi_t|=1$(其中,$|\mathcal{S}|$表示集合$\mathcal{S}$中的元素的数量),给定一个特定目标源DOA和从麦克风信号中提取的信息,将返回适当的TF掩码。因此,在这种形式下,可以进行直接比较。因为针对每个DOA分别计算TF掩码,所以与同时估计所有方向的掩码相比,当使用本申请时,能够更有效地使用神经网络的计算能力。

[0094] 图6A示出了模拟结果,其中,类似于图1的DNN用于同时估计所有方向的TF掩码。图6B示出了模拟结果,其中,根据本申请实施例,DNN用于估计目标DOA范围的TF掩码,其覆盖方位角宽度为15°。在存在附加噪音的情况下(SNR=5dB),图6A-图B中两种模拟的测试场景由两个同时说话的人组成(基本到达方位角相差60°)。其中一个语音信号服务于目标,而干扰性的说话者和噪音应被抑制。在图6A-图B中,横轴表示以秒为单位的时间,而纵轴表示以千赫兹为单位的频率。条60和62表示TF掩码颜色图例,其中TF掩码可以具有介于0和1之间的值,并且每个值与条60和62中所示的颜色相关联。显然,图6B所示的模拟中使用的DOA相关特征提取允许更准确地捕获语音的谐波结构。具体地,观察到以下关键改进,其中一些示例性部分也在图6A-B中标出:

[0095] 改善低频分离:在低频下,信号分量之间在空间信息上的差异(TF掩码估计仅基于此处)很小,这使得它们的分离变得更加困难。通过DOA相关特征提取,图6B中的低频段仍可看到明显的结构,如64所指示的,而图6A中的TF掩码已经模糊,如66所指示的。

[0096] 保留更大比例的目标语音:同样在更高的频率下,很明显,具有DOA相关特征提取的DNN可以更好地利用有限的信息。在这种情况下,信号在图6A中被完全抑制到更高的频率(TF掩码在任何地方接近0),如68所指示的,而图6B中使用的方法中的掩码仅仅抑制部分信号,如67所指示的。

[0097] 更准确捕捉精细结构:背景噪声是宽带的,而目标语音相对稀疏。因此,为了确保在具有显著语音活动的TF区域中也能抑制噪声,关键是在各个谐波之间也有相当大的抑制。在图6A中,谐波显得模糊,如69所指示的。由此产生的谐波之间抑制不足可能会引起可

感知的目标语音失真。相比之下,通过DOA相关特征提取,可以在掩码中看到更明显的谐波,从而显着降低可感知的目标语音失真(参见图6B中的63)。

[0098] 图7示出了一种用于检测或估计多个声源中的目标声源的方法的流程图,其中目标声源位于多个空间区域中的至少一个目标空间区域。

[0099] 在步骤70中,该方法包括:处理器接收多个信号,其中多个信号中的每一个与多个麦克风信号中的一个相关联,其中麦克风信号包括由多个声源产生的声音事件。

[0100] 在步骤72中,该方法包括:处理器通过神经网络提取多个特征,其中多个特征是从多个信号中提取的,并且其中通过针对多个空间区域中的不同空间区域训练神经网络,获得多个特征中的每一个。

[0101] 在步骤74中,该方法包括:处理器基于所提取的多个特征生成另外多个特征,其中另外多个特征对应于至少一个目标空间区域。

[0102] 最终,在步骤76中,该方法包括:处理器基于与至少一个目标空间区域对应的另外多个特征,检测或估计目标空间区域中的目标声源中的至少一个声源。

[0103] 尽管已经参考示例性实施例描述了本申请,但是本领域技术人员将理解,在不脱离本申请的范围的情况下,可以进行各种改变,并且可以用等价物代替其元件。此外,可以进行许多修改以使特定情况或材料适应本申请的教导,而不脱离本申请的基本范围。因此,旨在本申请不限于所公开的特定实施例,而是本申请将包括落入所附权利要求范围内的所有实施例。

[0104] 特别地,可以对本申请的各个方面的特定特征进行组合。通过添加关于本申请的另一方面描述的特征可以进一步有利地增强本申请的一个方面。应当理解,本申请仅受所附权利要求及其技术等价物的限制。在本文件及其权利要求中,动词“包括”及其变位以其非限制性意义使用,表示包括该词之后的项目,但不排除未具体提及的项目。此外,不定冠词“a”或“an”对元素的引用不排除存在多个元素的可能性,除非上下文明确要求存在一个且只有一个元素。因此,不定冠词“a”或“an”通常表示“至少一个”。

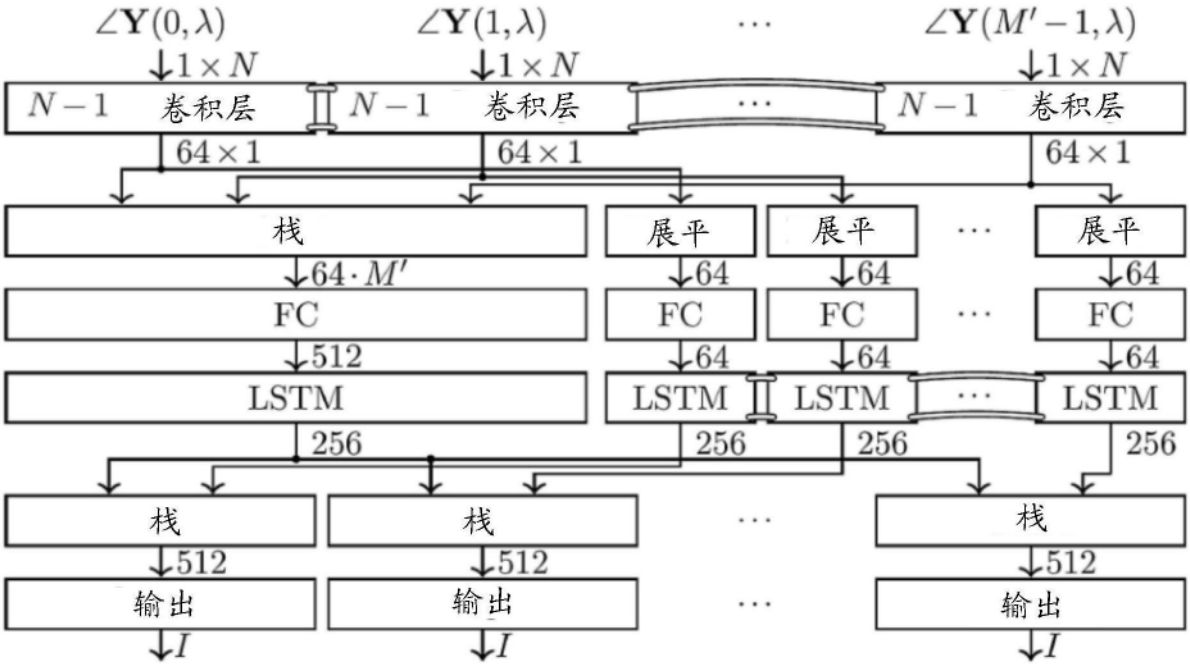

$\angle\mathbf{Y}(0,\lambda)$
$\angle\mathbf{Y}(1,\lambda)$
$\angle\mathbf{Y}(M'-1,\lambda)$
$1\times N$
$N-1$ 卷积层
64×1
栈
展平
$64\cdot M'$
64
FC
512
LSTM
256
输出
I

图1

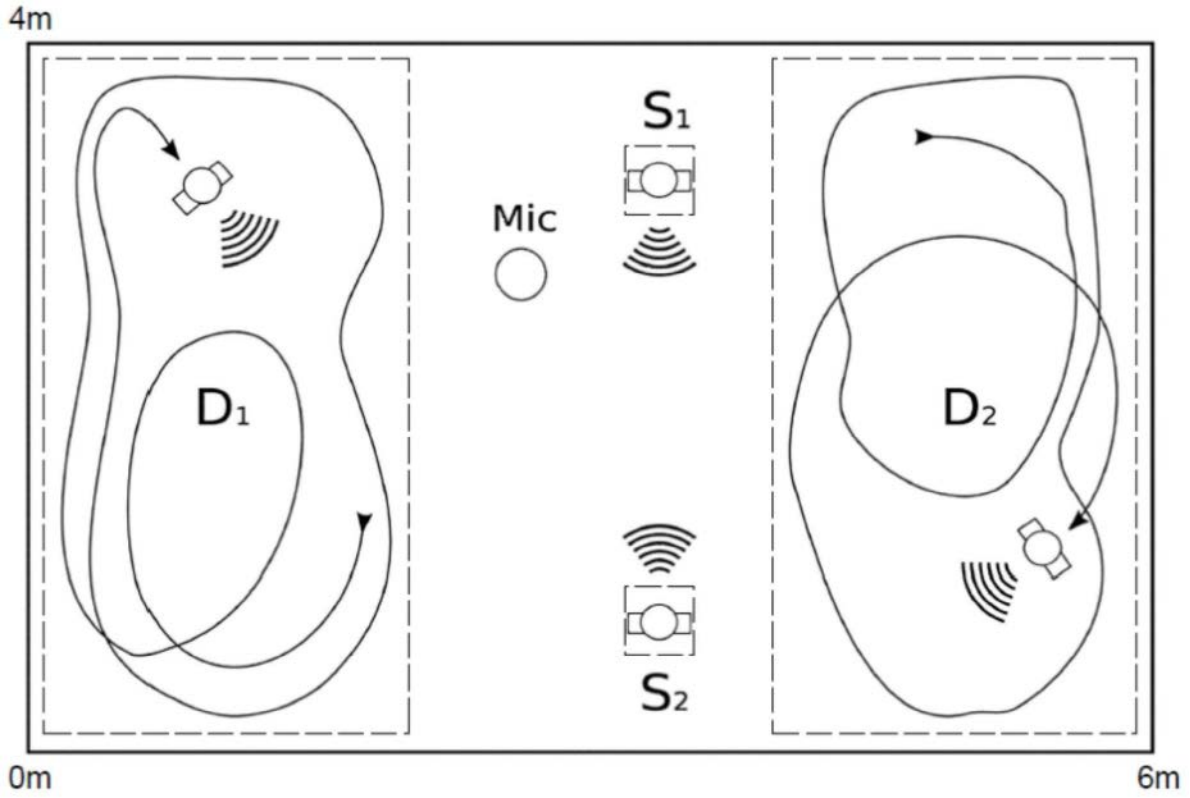

4m
0m
6m
Mic
S1
S2
D1
D2

图2

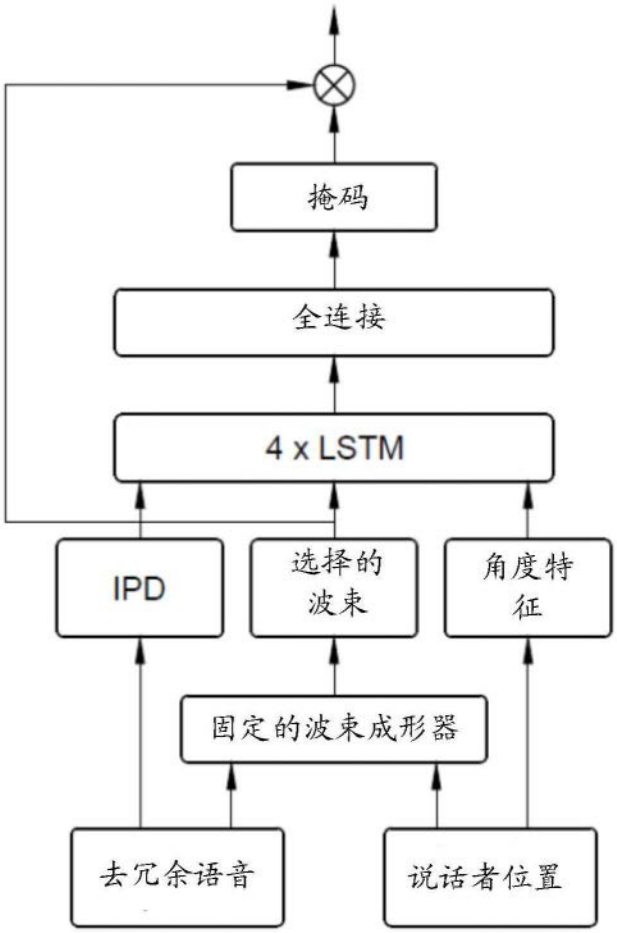

掩码
全连接
4 x LSTM
IPD
选择的波束
角度特征
固定的波束成形器
去冗余语音
说话者位置

图3

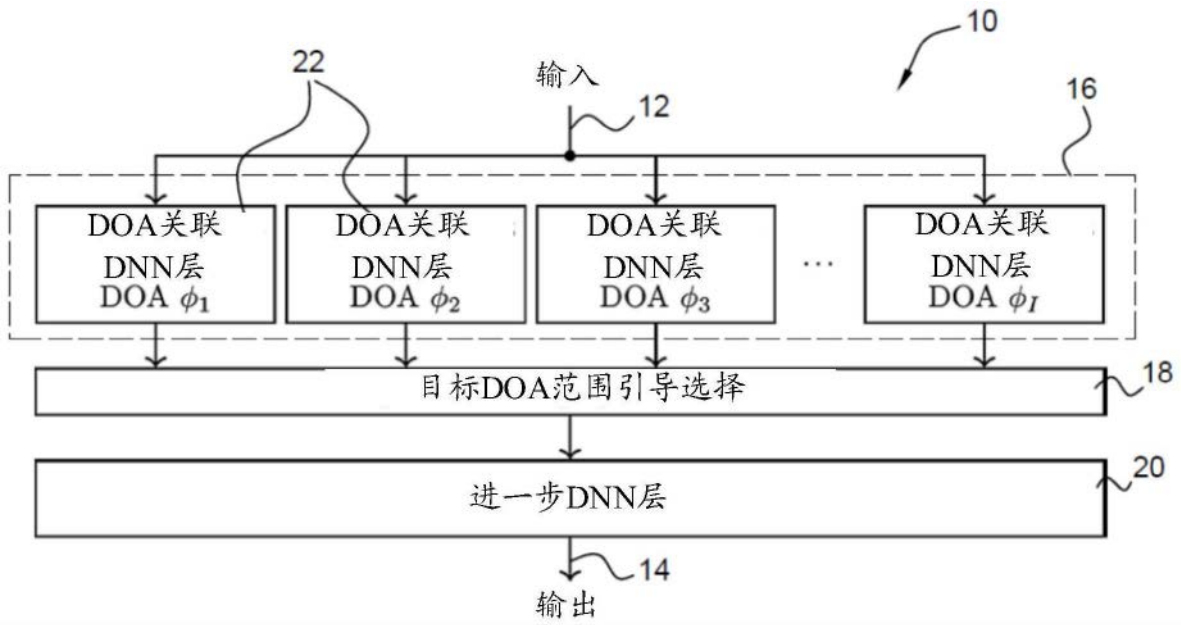

10
22
输入
12
16
DOA关联 DNN层 DOA ϕ_1
DOA关联 DNN层 DOA ϕ_2
DOA关联 DNN层 DOA ϕ_3
...
DOA关联 DNN层 DOA ϕ_I
目标DOA范围引导选择
18
进一步DNN层
20
14
输出

图4A

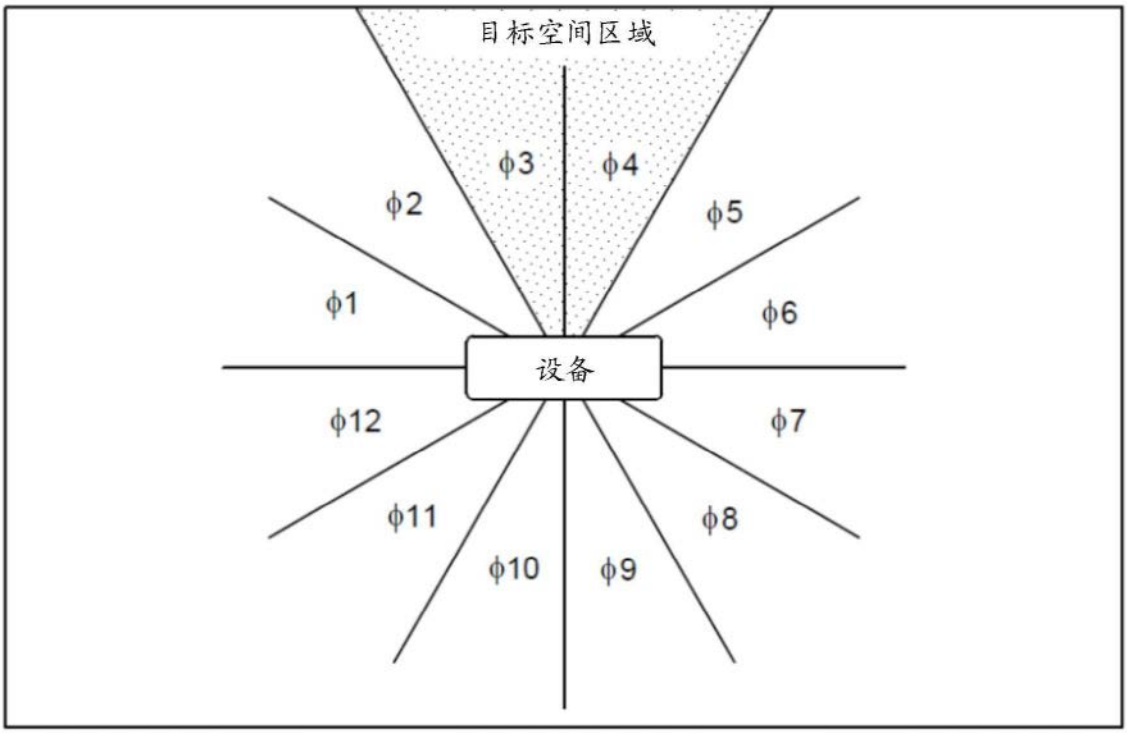


图4B

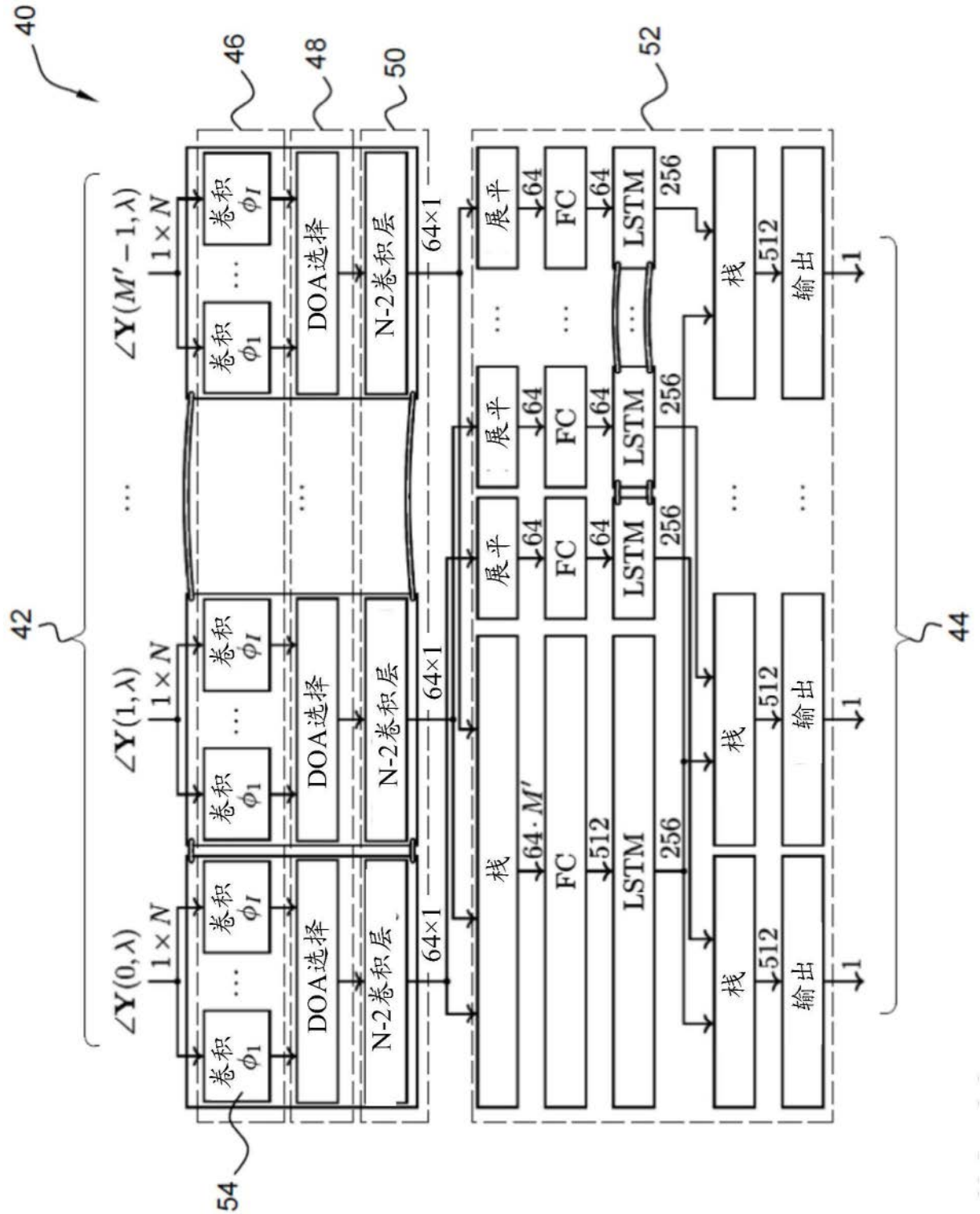

40
42
44
46
48
50
52
54
$\angle \mathbf{Y}(0,\lambda)$
$\angle \mathbf{Y}(1,\lambda)$
$\angle \mathbf{Y}(M'-1,\lambda)$
1×N
卷积 ϕ_1
卷积 ϕ_I
DOA选择
N-2卷积层
64×1
展平
64
FC
LSTM
256
栈
$64 \cdot M'$
512
输出
1

图5

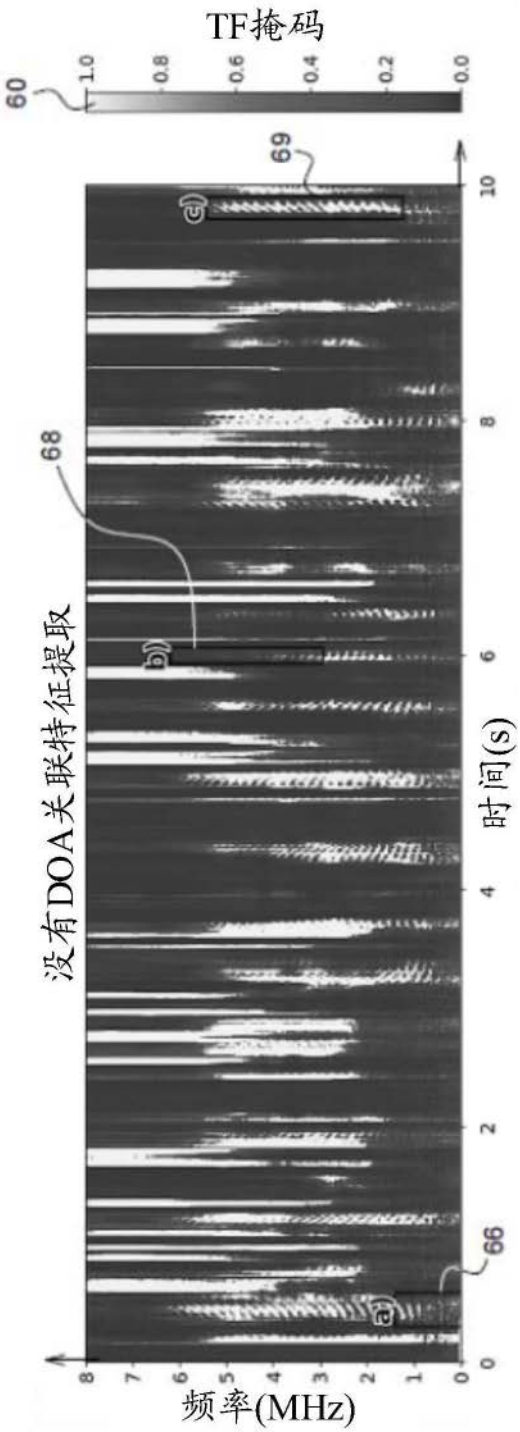

TF掩码
60
1.0
0.8
0.6
0.4
0.2
0.0
69
68
66
没有DOA关联特征提取
时间(s)
频率(MHz)
0
2
4
6
8
10
8
7
6
5
4
3
2
1
0

图6A

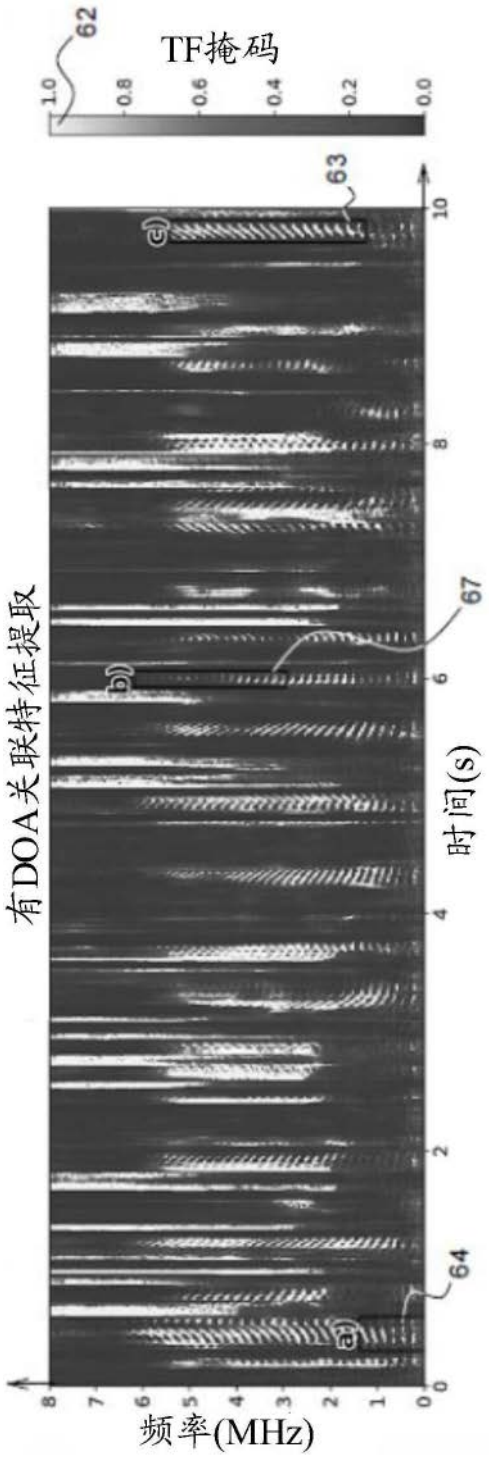

62
TF掩码
1.0
0.8
0.6
0.4
0.2
0.0
63
a)
10
8
67
b)
6
有DOA关联特征提取
时间(s)
4
2
64
a)
0
8 7 6 5 4 3 2 1 0
频率(MHz)

图6B

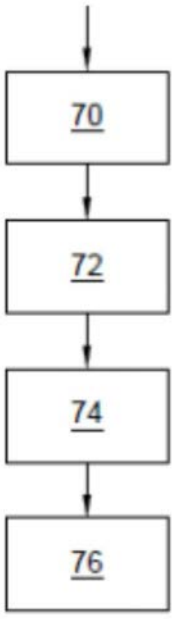


图7