

音频信号处理方法、设备、系统及存储介质

技术领域

本申请涉及音频处理技术领域，尤其涉及一种音频信号处理方法、设备、
5 系统及存储介质。

背景技术

麦克风阵列是由一定数目的麦克风组成，对声场的空间特性进行采样并滤波的系统。麦克风阵列对远场干扰噪声具有很强的抑制作用，可应用在一些具
10 有语音采集功能的产品中，例如拾音器、录音笔等，以便准确采集各种场景中的语音信号。

在一些应用场景中，既存在单人发言的情况，也可能发生多个发言人同时发言的情况，所采集到的语音信号既可能包括单一语音信号，也可能包括多个发言人的发言重叠在一起的重叠语音信号。为了准确识别会议中同时有几个发
15 言人，各自发言内容是什么，需要识别出重叠语音信号，进而对重叠语音信号做语音识别处理。

在现有技术中，可以通过大量音频数据训练得到的重叠语音检测模型，基于该检测模型对音频信号是否是重叠语音信号进行检测。但是，现有重叠语音检测大多是基于单通道音频的，直接将现有重叠语音检测技术应用到采用麦克
20 风阵列的多通道音频场景中，其准确性会下降，无法满足产品级别的检测需求。

发明内容

本申请的多个方面提供一种音频信号处理方法、设备、系统及存储介质，用以提高识别语音是否为重叠语音的准确性，以满足产品级别的检测需求。

25 本申请实施例提供一种音频信号处理方法，包括：获取麦克风阵列采集到

的当前音频信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

5 本申请实施例还提供一种音频信号处理方法，适用于会议设备，会议设备包括麦克风阵列，该方法包括：获取麦克风阵列在会议场景中采集到的当前会议信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史会议信号学习到的单一语音与重叠语音
10 之间的转换关系，识别当前会议信号是否为重叠语音。

 本申请实施例还提供一种音频信号处理方法，适用于教学设备，教学设备包括麦克风阵列，该方法包括：获取麦克风阵列在教学环境中采集到的当前课堂信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前课堂信号的相位差信息，生成当前课堂信号对应的当前声源空间分布信息；根据
15 当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

 本申请实施例还提供一种音频信号处理方法，适用于智能车载设备，智能车载设备包括麦克风采集阵列，该方法包括：获取麦克风阵列在车载环境中采集到的当前音频信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风
20 采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

 本申请实施例还提供一种终端设备，包括：存储器、处理器以及麦克风阵列；存储器，用于存储计算机程序；处理器，与存储器耦合，用于执行计算机
25 程序，以用于：获取麦克风阵列采集到的当前音频信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合

基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。本申请实施例还提供一种会议设备，包括：存储器、处理器以及麦克风阵列；存储器，用于存储计算机程序；处理器，与存储器耦合，用于执行计算机程序，以用于：获取麦克风阵列在会议场景中采集到的当前会议信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史会议信号学习到的单一语音与重叠语音之间的转换关系，识别当前会议信号是否为重叠语音。

本申请实施例还提供一种教学设备，包括：存储器、处理器以及麦克风阵列；存储器，用于存储计算机程序；处理器，与存储器耦合，用于执行计算机程序，以用于：获取麦克风阵列在教学环境中采集到的当前课堂信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前课堂信号的相位差信息，生成当前课堂信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

本申请实施例还提供一种智能车载设备，包括：存储器、处理器以及麦克风阵列；存储器，用于存储计算机程序；处理器，与存储器耦合，用于执行计算机程序，以用于：获取麦克风阵列在车载环境中采集到的当前音频信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

本申请实施例还提供一种音频信号处理系统，包括：终端设备和服务端设备；终端设备包括麦克风阵列，麦克风阵列包含至少两个麦克风，用于采集当前音频信号；终端设备用于将至少两个麦克风采集的当前音频信号上传至服务端设备；服务端设备，用于根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分

布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

本申请实施例还提供一种服务端设备，包括：存储器和处理器；存储器，用于存储计算机程序；处理器，与存储器耦合，用于执行计算机程序，以用于：
5 接收终端设备上传的麦克风阵列中至少两个麦克风采集的当前音频信号；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

10 本申请实施例还提供一种存储有计算机程序的计算机可读存储介质，当计算机程序被处理器执行时，致使处理器实现本申请实施例提供的音频信号处理方法中的步骤。

本申请实施例还提供一种计算机程序产品，包括计算机程序/指令，当计算机程序/指令被处理器执行时，致使处理器实现本申请实施例提供的音频信号处
15 理方法中的步骤。

在本申请实施例中，利用麦克风阵列采集音频信号，根据麦克风阵列中每个麦克风采集到音频信号的相位差信息，生成音频信号对应的声源空间分布信息，进而根据声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音，相比于单通道
20 音频，利用麦克风阵列采集的音频信号，包含了声源空间分布信息，从而可以准确地识别当前音频信号是否为重叠语音，满足产品级别的检测需求。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，
25 本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1a 为本申请示例性实施例提供的一种音频信号处理方法的流程示意图；

图 1b 为本申请示例性实施例提供的另一种音频信号处理方法的流程示意图；

图 2a 为本申请示例性实施例提供的麦克风阵列中麦克风布局的示意图；

图 2b 为本申请示例性实施例提供的声源空间分布信息的峰值信息的示意图；

图 3a 为会议设备在会议场景下的使用状态示意图；

5 图 3b 为拾音设备在商务合作商谈场景下的使用状态示意图；

图 3c 为教学设备在教学场景下的使用状态示意图；

图 3d 为智能车载设备在车载环境下的使用状态示意图；

图 3e 为本申请示例性实施例提供的又一种音频信号处理方法的流程示意图；

图 3f 为本申请示例性实施例提供的又一种音频信号处理方法的流程示意图；

10 图 3g 为本申请示例性实施例提供的又一种音频信号处理方法的流程示意图；

图 4 为本申请示例性实施例提供的一种音频信号处理系统的结构示意图；

图 5 为本申请示例性实施例提供的一种终端设备的结构示意图；

图 6 为本申请示例性实施例提供的一种服务端设备的结构示意图。

15 具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施
20 例，都属于本申请保护的范围。

以下结合附图，详细说明本申请各实施例提供的技术方案。

图 1a 为本申请示例性实施例提供的一种音频信号处理方法的流程示意图；
如图 1a 所示，该方法包括：

101a、获取麦克风阵列采集到的当前音频信号，麦克风阵列包含至少两个
25 麦克风；

102a、根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；

103a、根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

在本实施例中，声源是指可以通过振动产生声音的物体，例如，声源可以是乐器、振动的音叉、人类的发声器官（例如声带）或者动物的发声器官等，
5 声源可以产生语音，语音是指由人的发音器官发出的具有一定社会意义的声音。麦克风阵列可以采集声源发出的音频信号，音频信号中可能存在语音，也可能存在语音之外的其它声音，例如，混响（Reverberation）、回声（Echo）、环境噪声、动物叫声或者物体碰撞声等。

在本实施例中，麦克风阵列中包含有至少两个麦克风，其中，至少两个麦
10 克风的布局不做限定，如图 2a 所示，可以是线性阵列、平面阵列或者立体阵列。由于麦克风阵列中麦克风存在一定的布局，同一音频信号到达每个麦克风的时间会有所差异，即存在时延。该时延可体现为同一音频信号到达各麦克风时的相位差异，简称为相位差。

在本实施例中，麦克风阵列采集音频信号过程中，根据应用场景的不同，
15 随时可能发生“抢话”等现象，也即，其中一个人打断另一个人的发言，麦克风阵列采集到的当前音频信号可能是一个发言人产生的单一语音，也可能是多个发言人的发言重叠在一起的重叠语音。在本实施例中，假定音频信号存在两种状态，一种是单一语音，一种是重叠（overlap）语音。

在本实施例中，借助于麦克风阵列中各麦克风采集当前音频信号时存在的
20 相位差信息对音频信号的状态进行判断，即判断音频信号是单一语音还是重叠语音，或者判断音频信号是否为重叠语音。其中，该相位差信息在一定程度上可体现声源位置的空间分布情况，根据声源位置的空间分布情况可识别出有效声源的数量以及有效声源的位置，在识别到有效声源数量的情况下，可以判断音频信号是否为重叠语音。

25 具体地，可以获取麦克风阵列采集到的当前音频信号，音频信号的切分长度不做限定，可以以信号帧为单位，则当前音频信号可以是一个信号帧，每个信号帧通常是毫秒级的（例如，20ms），通常小于发言过程中单个字或音节的发

音持续时间，或者，也可以是连续几个信号帧作为当前音频信号，对此不做限定。接着，根据麦克风阵列中至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息。当前声源空间分布信息反应了当前声源在空间上的分布情况，根据当前声源在空间上的分布情况，可识别出有效声源的数量以及有效声源的位置，在识别到有效声源数量的情况下，
5 可以判断音频信号是否为重叠语音。

在实际应用中，鉴于音频信号的连续性，从一个状态转换为另一状态存在一定规律，例如，当前音频信号的状态可能与前一个音频信号对应的状态有关，也可能与前两个或者前 N ($N > 2$) 个音频信号对应的状态有关，基于此，在单一语音和重叠语音的初始化概率下，基于历史音频信号的状态不断学习单一语音与重叠语音之间的转换关系，该转换关系指的是音频信号对应的状态之间的转换概率，可以包括：单一语音与单一语音之间的转换概率、单一语音与重叠语音之间的转换该率、重叠语音与单一语音之间的转换概率以及重叠语音与重叠语音之间的转换概率。基于上述，在判断当前音频信号是否为重叠信号时，
10 可依赖于学习到的单一语音与重叠语音之间的转换关系，并结合当前声源空间分布信息，识别当前音频信号是否为重叠信号，相比于单通道音频，利用麦克风阵列采集的音频信号，包含了声源空间分布信息，从而可以准确地识别任一时刻的音频信号是否为重叠语音，满足产品级别的检测需求。

在本实施例中，相位差信息在一定程度上可体现声源位置的空间分布情况，
20 为了更好地体现声源位置的空间分布情况，在本申请一些可选实施例中，可根据至少两个麦克风采集到当前音频信号的相位差信息，计算当前音频信号对应的波达谱图，波达谱图可反应当前声源在空间上的分布情况。

进一步可选地，针对位置空间中的任一方位，对任意两个麦克风采集到当前音频信号的相位差信息进行累加，得到每个方位作为当前声源位置的概率；
25 根据位置空间中各方位作为当前声源位置的概率，生成当前音频信号对应的波达谱图。具体地，可以利用基于相位变换加权的可控响应功率的声源定位算法（Steered Response Power-PHase Transform, SRP-PHAT）得到每个方位作为当

前声源位置的概率。SRP-PHAT 算法的基本原理是：假设位置空间中的任一方位是声源所处的方位，麦克风阵列采集该方位处的声源发出的音频信号，利用广义互相关-相位变换加权（Generalized Cross Correlation-PHase Transformation, GCC-PHAT）算法，计算任意两个麦克风采集到的音频信号之间的互相关函数，并对互相关函数的互功率谱密度函数进行加权，接着，对计算出的所有任意两个麦克风之间的 GCC-PHAT 值进行累加，以得到任一方位对应的 SRP-PHAT 值。进一步，可以根据任一方位对应的 SRP-PHAT 值，得到各方位作为当前声源位置的概率，根据位置空间中各方位作为当前声源位置的概率，生成当前音频信号对应的波达谱图。例如，可以直接将各方位对应的 SRP-PHAT 值，作为各方位作为当前声源位置的概率，波达谱图中可以记录各方位及其对应的 SRP-PHAT 值，SRP-PHAT 值越大，表示 SRP-PHAT 值对应的方位是声源方位的概率越大。又例如，可以将各方位上的 SRP-PHAT 值与所有方位 SRP-PHAT 值的和的比值，作为每个方位作为当前声源位置的概率，波达谱图中可直接体现各方位作为当前声源位置的概率。

在一可选实施例中，可以利用隐马尔科夫模型（Hidden Markov Model, HMM）识别当前音频信号是否为重叠语音。具体地，可以将音频信号的状态，即单一语音和重叠语音，作为 HMM 的两个隐藏状态，计算音频信号的当前声源空间分布信息的峰值信息，作为 HMM 的观测状态。可选地，可以利用峰度（Kurtosis）算法或者过剩质量（Excessive Mass）算法计算当前声源空间分布信息的峰值信息。其中，峰值信息可以是波峰的个数，如图 2b 所示，为峰值信息的三种形态，也即观测状态的三种形态，三种形态分别为单峰（Unimodal）、双峰（Bimodal）以及多峰（Multimodal）。

在本实施例中，针对当前音频信号计算出观测状态之后，可以将当前观测状态输入 HMM，结合 HMM 学习到的两个隐藏状态之间的跳转关系，以历史观测状态为前提条件，计算当前观测状态对应隐藏状态的概率。具体地，可以设定隐藏状态的初始化概率，例如，可以是单一语音 0.6，重叠语音 0.4，在设定隐藏状态的初始化概率的情况下，基于历史音频信号的状态，不断学习隐藏状

态之间的转换关系以及隐藏状态到观测状态的发射关系，以得到 HMM 模型，将观测状态输入至 HMM 模型后，HMM 模型输出以历史观测状态为前提条件的情况下，当前观测状态是隐藏状态的概率。例如，历史观测状态为连续五个单峰，HMM 模型将连续五个历史观测状态对应的隐藏状态识别为单一语音，则在出现连续五个单峰观测状态的前提下，若当前观测状态为双峰观测状态，HMM 模型分别输出当前观测状态为重叠语音和单一语音的概率，且当前观测状态为重叠语音的概率大于当前观测状态为单一语音的概率。

在本实施例中，在 HMM 模型输出当前观测状态对应隐藏状态的概率之后，可根据当前观测状态对应隐藏状态的概率，识别当前音频信号是否为重叠语音。若当前观测状态对应为重叠语音的概率大于当前观测状态对应为单一语音的概率，则认为当前音频信号为重叠语音；若当前观测状态对应为重叠语音的概率小于或等于当前观测状态对应为单一语音的概率，则认为当前音频信号为单一语音。

在一可选实施例中，若识别到当前音频信号为重叠语音，则根据当前声源空间分布信息，确定至少两个有效声源方位。例如，在当前声源空间分布信息包括各方位作为当前声源位置的概率的情况下，将作为当前声源位置的最大的两个方位作为有效声源方位。又例如，若当前声源空间分布信息用波达谱图来体现，波达谱图包括各方位的 SRP-PHAT 值，则可以从波达谱图中选择 SRP-PHAT 值最大的两个方位，作为有效声源方位。接着，可以对至少两个有效声源方位上的音频信号进行语音增强，具体地，可以利用波束形成（Beam Forming, BF）技术在有效声源方位上的音频信号形成波束，利用该波束可以有效地对音频信号进行语音增强，而抑制有效声源方位之外的其它方位的音频信号，从而达到语音分离的效果，在此基础上对至少两个有效声源方位上增强后的音频信号分别进行语音识别，可以提高语音识别的准确性，提高用户的体验感。

在另一可选实施例中，若识别到当前音频信号为单一语音，则将作为当前声源位置的最大的一个方位作为有效声源方位；对有效声源方位上的音频

信号进行语音增强，并对有效声源方位上增强后的音频信号进行语音识别。关于对单一语音进行语音增强的实施方式，与上述实施例中对重叠语音进行语音增强的实施方式相同或相似，在此不再赘述。

在本申请一些应用场景中，例如，会议场景、教师授课场景或者商务合作会谈场景中，往往需要对语音信号进行识别，对于非语音信号，例如，环境噪声、动物叫声或者物体碰撞声，则不太关注。基于此，在识别当前音频信号是否为重叠语音之前，还可以判断当前音频信号是否为语音信号，若当前音频信号不是语音信号，则不对该当期音频信号进行识别，以提高音频处理的效率，若当前音频信号是语音信号，则识别当前音频信号是否为重叠语音。

10 基于上述，本申请实施例还提供一种音频信号处理方法，如图 1b 所示，该方法包括：

101b、获取麦克风阵列采集到的当前音频信号，麦克风阵列包含至少两个麦克风；

15 102b、根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；

103b、根据当前的声源空间分布信息，计算当前音频信号的波达方向；

104b、根据波达方向，从至少两个麦克风中选择一路麦克风作为目标麦克风；

20 105b、对目标麦克风采集到的当前音频信号进行语音端点检测 VAD，以得到当前音频信号是否为语音信号；

106b、若当前音频信号为语音信号，则执行步骤 107b，否则结束对当前音频信号的处理；

107b、根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

25 在本实施例中，关于步骤 101b、102b 以及 107b 的内容，可参见前述实施例中步骤 101a、102a 以及 103a 的详细内容，在此不再赘述。

在本实施例中，根据当前的声源空间分布信息，计算当前音频信号的波达

方向 (Direction of arrival, DOA), 波达方向是指当前音频信号到达麦克风阵列的方向角, 该波达方向与麦克风阵列中每个麦克风接收音频信号的方向角可能相同, 也可能不同, 具体跟麦克风的布局有关。其中, 在声源空间分布信息包括各方位作为当前声源位置的概率的情况下, 可以将作为当前声源位置的概率最大的方位直接作为波达方向, 也可以将与该作为当前声源位置的概率最大的方位成设定角度的方位作为波达方向, 对此不做限定。

在计算出波达方向之后, 可以根据波达方向, 从至少两个麦克风中选择一路麦克风作为目标麦克风, 例如, 可以计算每个麦克风接收当前音频信号的方向角, 从多个方向角中选择与波达方向一致的方向角, 将该方向角对应的麦克风作为目标麦克风。在确定目标麦克风之后, 可以对目标麦克风采集到的当前音频信号进行语音端点检测 (Voice Activity Detection, VAD), 以得到当前音频信号是否为语音信号。其中, VAD 的基本原理是从带有噪声的音频信号中准确的定位出语音信号的开始和结束的端点, 从而确定当前音频信号是否为语音信号, 即, 若可以从音频信号中检测出语音信号开始和结束的端点, 则认为该音频信号为语音信号, 若无法从音频信号中检测出语音信号开始和结束的端点, 则认为该音频信号不是语音信号。

需要说明的是, 在本实施例中, 并不限定对目标麦克风采集到的当前音频信号进行 VAD 的实施方式。在一可选实施例中, 可以利用软件 VAD 功能对目标麦克风采集到的当前音频信号进行 VAD, 软件 VAD 功能是指通过软件实现 VAD 功能, 并不对实现 VAD 功能的软件进行限定, 例如可以是通过人声模型训练而成的神经网络 VAD (Nature Network - VAD, NN-VAD) 模型等。在另一可选实施例中, 可以利用硬件 VAD 功能对目标麦克风采集到的当前音频信号进行 VAD, 硬件 VAD 功能是指通过在语音芯片或者设备上内置 VAD 模块实现 VAD 功能, VAD 模块可以固化在语音芯片上, 通过配置参数来修改 VAD 功能。

本实施例提供的音频信号处理方法, 可应用于到各种多人发言场景中, 例如多人会议场景、庭审场景或者教学场景等。在这些应用场景中, 本实施例的终端设备会被部署在这些场景中, 用于采集应用场景中的音频信号, 并实现本

申请上述各方法实施例以及下述系统实施例中所描述的其它功能。其中，终端设备可以实现为录音笔、录音棒、录音机或拾音器等拾音设备，也可以实现为带有录音功能的终端设备等，例如，会议设备、教学设备、机器人、智能机顶盒、智能电视、智能音箱以及智能车载设备等。为了有更好的采集效果，便于识别音频信号是否为重叠语音，进一步，根据音频信号是否为重叠语音，对音频信号进行语音增强以及语音识别，可以根据多人发言场景的具体部署情况合理确定终端设备的放置位置。如图 3a 所示，在多人会议场景中，以终端设备是会议设备为例进行图示，该会议设备包含麦克风阵列，具有拾音功能，考虑到多个发言人分布在会议设备的不同方位，优选地，将会议设备部署在会议桌的中央；如图 3b 所示，在商务合作会谈场景下，以终端设备是拾音设备为例进行图示，第一商务方和第二商务方相对落座，会议组织方位于第一商务方和第二商务方之间，负责组织两方商谈，拾音设备部署在会议组织方、第一商务方、第二商务方的中心位置，第一商务方、第二商务方和会议组织方拾音设备不同方位上，方便拾音设备拾音；如图 3c 所示，在教学场景中，以终端设备是教学设备为例进行图示，教学设备部署在讲课桌上，教师与学生位于教学设备不同方位，方便同时拾取教师与学生的语音。如图 3d 所示，在车载场景中，终端设备实现为车机设备上的智能车载设备，智能车载设备位于汽车的中央，座位 A、座位 B、座位 C 以及座位 D 上的乘客位于智能车载设备不同方位上，方便拾取不同乘客的语音。

下面针对不同应用场景下的音视频信号处理过程进行详细说明。

针对图 3a 所示的会议场景，本申请实施例提供一种音频信号处理方法，适用于会议设备，如图 3e 所示，该方法包括：

301e、获取麦克风阵列在会议场景中采集到的当前会议信号，麦克风阵列包含至少两个麦克风；

302e、根据至少两个麦克风采集到当前会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；

303e、根据当前声源空间分布信息，结合基于历史会议信号学习到的单一

语音与重叠语音之间的转换关系，识别当前会议信号是否为重叠语音。

关于步骤 301e 至步骤 303e 的内容，可参见前述图 1a 和图 1b 所示的实施例，在此不再赘述。

针对图 3c 所示的教学场景，本申请实施例提供一种音频信号处理方法，适用于教学设备，如图 3f 所示，该方法包括：

301f、获取麦克风阵列在教学环境中采集到的当前课堂信号，麦克风阵列包含至少两个麦克风；

302f、根据至少两个麦克风采集到当前课堂信号的相位差信息，生成当前课堂信号对应的当前声源空间分布信息；

10 303f、根据当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

关于步骤 301f 至步骤 303f 的内容，可参见前述图 1a 和图 1b 所示的实施例，在此不再赘述。

针对图 3d 所示的车载场景，本申请实施例提供一种音频信号处理方法，适用于智能车载设备，如图 3g 所示，该方法包括：

301g、获取麦克风阵列在车载环境中采集到的当前音频信号，麦克风阵列包含至少两个麦克风；

302g、根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；

20 303g、根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

关于步骤 301g 至步骤 303g 的内容，可参见前述图 1a 和图 1b 所示的实施例，在此不再赘述。

需要说明的是，本申请实施例提供的方法可以全部由终端设备完成，也可以将一部分功能在服务端设备上实现，对此不做限定。基于此，本实施例提供一种音频信号处理系统，对音频信号处理方法基于终端设备和服务端设备共同实现的过程进行说明。如图 4 所示，该音频信号处理系统 400 包括：终端设备

401 和服务端设备 402。该音频信号处理系统 400 可以应用到多人发言场景中，例如图 3a 所示的多人会议场景，图 3b 所示的商务合作会谈场景、图 3c 所示的教学场景以及图 3d 所示的车载环境等。在这些场景中，终端设备 401 可与服务端设备 402 配合实现本申请上述各方法实施例，在图 3a 至图 3d 所示多人发言场景中未示出服务端设备 402。

本实施例的终端设备 401 具有开机按键、调节按键、麦克风阵列以及扬声器等功能模块，其中，麦克风阵列包含有至少两个麦克风，进一步可选地，还可以包括显示屏。终端设备 401 可以实现自动录音、MP3 播放、FM 调频、数码相机功能、电话录音、定时录音、外部转录、复读机或编辑等功能。如图 4 所示，终端设备 401 可以利用麦克风阵列中的至少两个麦克风采集当前音频信号，将至少两个麦克风采集的当前音频信号上传至服务端设备 402；服务端设备 402 接收至少两个麦克风采集的当前音频信号，根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

在本申请一些可选实施例中，服务端设备 402 若识别到当前音频信号为重叠语音，则根据当前声源空间分布信息，确定至少两个有效声源方位；对至少两个有效声源方位上的音频信号进行语音增强，并对至少两个有效声源方位上增强后的音频信号分别进行语音识别。进一步可选地，在当前声源空间分布信息包括各方位作为当前声源位置的概率的情况下，将作为当前声源位置的概率最大的两个方位作为有效声源方位。

在本申请一些可选实施例中，服务端设备 402 若识别到当前音频信号为单一语音，则将作为当前声源位置的概率最大的一个方位作为有效声源方位；对有效声源方位上的音频信号进行语音增强，并对有效声源方位上增强后的音频信号进行语音识别。

需要说明的是，在音频信号处理系统应用到不同场景时，终端设备的实现形态有所不同。例如，在会议场景中，终端设备实现为会议设备；在商务合作

会谈场景中，终端设备实现为拾音设备；在教学场景中，终端设备实现为教学设备；在车载环境中，终端设备实现为智能车载设备。

需要说明的是，上述实施例所提供方法的各步骤的执行主体均可以是同一设备，或者，该方法也由不同设备作为执行主体。比如，步骤 101a 至步骤 103a 5 的执行主体可以为设备 A；又比如，步骤 101a 和 102a 的执行主体可以为设备 A，步骤 103a 的执行主体可以为设备 B；等等。

另外，在上述实施例及附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，操作的序号如 101a、102a 等，仅仅是用于区分开各个 10 不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。

图 5 为本申请示例性实施例提供的一种终端设备的结构示意图。如图 5 所示，该终端设备包括：麦克风阵列 53、存储器 54 以及处理器 55。

存储器 54，用于存储计算机程序，并可被配置为存储其它各种数据以支持 15 在终端设备上的操作。这些数据的示例包括用于在终端设备上操作的任何应用程序或方法的指令。

存储器 54 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器 (SRAM)，电可擦除可编程只读存储器 (EEPROM)，可擦除可编程只读存储器 (EPROM)，可编程只读存储器 (PROM)，只读存 20 储器 (ROM)，磁存储器，快闪存储器，磁盘或光盘。

处理器 55，与存储器 54 耦合，用于执行存储器 54 中的计算机程序，以用于：获取麦克风阵列 53 采集到的当前音频信号，麦克风阵列 53 包含至少两个麦克风；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于 25 历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

在一可选实施例中，处理器 55 在根据至少两个麦克风采集到当前音频信号

的相位差信息,生成当前音频信号对应的当前声源空间分布信息时,具体用于:
根据至少两个麦克风采集到当前音频信号的相位差信息,计算当前音频信号对
应的波达谱图,波达谱图反应当前声源在空间上的分布情况。

在一可选实施例中,处理器 55 在根据至少两个麦克风采集到当前音频信号
5 的相位差信息,计算当前音频信号对应的波达谱图时,具体用于:针对位置空
间中的任一方位,对任意两个麦克风采集到当前音频信号的相位差信息进行累
加,得到方位作为当前声源位置的概率;根据位置空间中各方位作为当前声源
位置的概率,生成当前音频信号对应的波达谱图。

在一可选实施例中,处理器 55 在根据当前声源空间分布信息,结合基于历
10 史音频信号学习到的单一语音与重叠语音之间的转换关系,识别当前音频信号
是否为重叠语音时,具体用于:计算当前声源空间分布信息的峰值信息,作为
隐马尔科夫模型 HMM 的当前观测状态,并将单一语音与重叠语音作为 HMM
的两个隐藏状态;将当前观测状态输入 HMM,结合 HMM 学习到的两个隐藏状
态之间的跳转关系,以历史观测状态为前提条件,计算当前观测状态对应隐藏
15 状态的概率;根据当前观测状态对应隐藏状态的概率,识别当前音频信号是否
为重叠语音。

在一可选实施例中,处理器 55 还用于:若识别到当前音频信号为重叠语音,
则根据当前声源空间分布信息,确定至少两个有效声源方位;对至少两个有效
声源方位上的音频信号进行语音增强,并对至少两个有效声源方位上增强后的
20 音频信号分别进行语音识别。

在一可选实施例中,处理器 55 在根据当前声源空间分布信息,确定至少两
个有效声源方位时,具体用于:在当前声源空间分布信息包括各方位作为当前
声源位置的概率的情况下,将作为当前声源位置的最大的两个方位作为有
效声源方位。

25 在一可选实施例中,处理器 55 还用于:若识别到当前音频信号为单一语音,
则将作为当前声源位置的最大的一个方位作为有效声源方位;对有效声源
方位上的音频信号进行语音增强,并对有效声源方位上增强后的音频信号进行

语音识别。

在一可选实施例中，处理器 55 在识别当前音频信号是否为重叠语音之前，还用于：根据当前的声源空间分布信息，计算当前音频信号的波达方向；根据波达方向，从至少两个麦克风中选择一路麦克风作为目标麦克风；对目标麦克风采集到的当前音频信号进行语音端点检测 VAD，以得到当前音频信号是否为语音信号。

在一可选实施例中，终端设备为会议设备、拾音设备、机器人、智能机顶盒、智能电视、智能音箱以及智能车载设备。

本申请实施例提供的终端设备可以利用麦克风阵列采集音频信号，根据麦克风阵列中每个麦克风采集到音频信号的相位差信息，生成音频信号对应的声源空间分布信息，进而根据声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音，相比于单通道音频，利用麦克风阵列采集的音频信号，包含了声源空间分布信息，从而可以准确地识别当前音频信号是否为重叠语音，满足产品级别的检测需求。

进一步，如图 5 所示，该终端设备还包括：通信组件 56、显示器 57、电源组件 58、扬声器 59 等其它组件。图 5 中仅示意性给出部分组件，并不意味着终端设备只包括图 5 所示组件。需要说明的是，图 5 中虚线框内的组件为可选组件，而非必选组件，具体可视终端设备的产品形态而定。

在一可选实施例中，上述终端设备可以应用到不同的应用场景，在应用到不同的应用场景时，具体实现为不同设备形态。

例如，终端设备可以实现为会议设备，该会议设备的实现结构与图 5 所示的终端设备的实现结构相同或类似，可参照图 5 所示终端设备的结构实现。该会议设备与图 5 所示实施例中终端设备的区别主要在于：处理器执行存储器中存储的计算机程序所实现的功能不同。对于会议设备来说，其处理器执行存储器中存储的计算机程序，可用于：获取麦克风阵列在会议场景中采集到的当前会议信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前

会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史会议信号学习到的单一语音与重叠语音之间的转换关系，识别当前会议信号是否为重叠语音。

再例如，终端设备可以实现为教学设备，该教学设备的实现结构与图 5 所示的终端设备的实现结构相同或类似，可参照图 5 所示终端设备的结构实现。该教学设备与图 5 所示实施例中终端设备的区别主要在于：处理器执行存储器中存储的计算机程序所实现的功能不同。对于教学设备来说，其处理器执行存储器中存储的计算机程序，可用于：获取麦克风阵列在教学环境中采集到的当前课堂信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前课堂信号的相位差信息，生成当前课堂信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

又例如，终端设备可以实现为智能车载设备，该智能车载设备的实现结构与图 5 所示的终端设备的实现结构相同或类似，可参照图 5 所示终端设备的结构实现。该智能车载设备与图 5 所示实施例中终端设备的区别主要在于：处理器执行存储器中存储的计算机程序所实现的功能不同。对于智能车载设备来说，其处理器执行存储器中存储的计算机程序，可用于：获取麦克风阵列在车载环境中采集到的当前音频信号，麦克风阵列包含至少两个麦克风；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

相应地，本申请实施例还提供一种存储有计算机程序的计算机可读存储介质，计算机程序被处理器执行时，致使处理器能够实现本申请实施例提供的各方法实施例中的各步骤。

相应地，本申请实施例还提供一种计算机程序产品，包括计算机程序/指令，当计算机程序/指令被处理器执行时，致使处理器实现本申请实施例提供的各方法中的步骤。

图 6 为本申请示例性实施例提供的一种服务端设备的结构示意图。如图 6 所示，该服务端设备包括：存储器 64 以及处理器 65。

存储器 64，用于存储计算机程序，并可被配置为存储其它各种数据以支持在服务端设备上的操作。这些数据的示例包括用于在服务端设备上操作的任何应用程序或方法的指令。

存储器 64 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器 (SRAM)，电可擦除可编程只读存储器 (EEPROM)，可擦除可编程只读存储器 (EPROM)，可编程只读存储器 (PROM)，只读存储器 (ROM)，磁存储器，快闪存储器，磁盘或光盘。

处理器 65，与存储器 64 耦合，用于执行存储器 64 中的计算机程序，以用于：接收终端设备上传的麦克风阵列中至少两个麦克风采集的当前音频信号；根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

在一可选实施例中，处理器 65 在根据至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息时，具体用于：根据至少两个麦克风采集到当前音频信号的相位差信息，计算当前音频信号对应的波达谱图，波达谱图反应当前声源在空间上的分布情况。

在一可选实施例中，处理器 65 在根据至少两个麦克风采集到当前音频信号的相位差信息，计算当前音频信号对应的波达谱图时，具体用于：针对位置空间中的任一方位，对任意两个麦克风采集到当前音频信号的相位差信息进行累加，得到方位作为当前声源位置的概率；根据位置空间中各方位作为当前声源位置的概率，生成当前音频信号对应的波达谱图。

在一可选实施例中，处理器 65 在根据当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音时，具体用于：计算当前声源空间分布信息的峰值信息，作为

隐马尔科夫模型 HMM 的当前观测状态，并将单一语音与重叠语音作为 HMM 的两个隐藏状态；将当前观测状态输入 HMM，结合 HMM 学习到的两个隐藏状态之间的跳转关系，以历史观测状态为前提条件，计算当前观测状态对应隐藏状态的概率；根据当前观测状态对应隐藏状态的概率，识别当前音频信号是否

5 为重叠语音。

在一可选实施例中，处理器 65 还用于：若识别到当前音频信号为重叠语音，则根据当前声源空间分布信息，确定至少两个有效声源方位；对至少两个有效声源方位上的音频信号进行语音增强，并对至少两个有效声源方位上增强后的音频信号分别进行语音识别。

10 在一可选实施例中，处理器 65 在根据当前声源空间分布信息，确定至少两个有效声源方位时，具体用于：在当前声源空间分布信息包括各方位作为当前声源位置的概率的情况下，将作为当前声源位置的概率最大的两个方位作为有效声源方位。

在一可选实施例中，处理器 65 还用于：若识别到当前音频信号为单一语音，

15 则将作为当前声源位置的概率最大的一个方位作为有效声源方位；对有效声源方位上的音频信号进行语音增强，并对有效声源方位上增强后的音频信号进行语音识别。

在一可选实施例中，处理器 65 在识别当前音频信号是否为重叠语音之前，还用于：根据当前的声源空间分布信息，计算当前音频信号的波达方向；根据

20 波达方向，从至少两个麦克风中选择一路麦克风作为目标麦克风；对目标麦克风采集到的当前音频信号进行语音端点检测 VAD，以得到当前音频信号是否为语音信号。

本申请实施例提供的服务端设备可以利用麦克风阵列采集音频信号，根据麦克风阵列中每个麦克风采集到音频信号的相位差信息，生成音频信号对应的

25 声源空间分布信息，进而根据声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音，相比于单通道音频，利用麦克风阵列采集的音频信号，包含了声源空间分布信

息，从而可以准确地识别当前音频信号是否为重叠语音，满足产品级别的检测需求。

进一步，如图 6 所示，该服务端设备还包括：通信组件 66、电源组件 68 等其它组件。图 6 中仅示意性给出部分组件，并不意味着服务端设备只包括图 6 所示组件。需要说明的是，图 6 中虚线框内的组件为可选组件，而非必选组件，具体可视服务端设备的产品形态而定。

相应地，本申请实施例还提供一种存储有计算机程序的计算机可读存储介质，计算机程序被处理器执行时，致使处理器能够实现本申请实施例提供的各方法实施例中的各步骤。

相应地，本申请实施例还提供一种计算机程序产品，包括计算机程序/指令，当计算机程序/指令被处理器执行时，致使处理器实现本申请实施例提供的各方法中的步骤。

上述图 5 和图 6 中的通信组件被配置为便于通信组件所在设备和其他设备之间有线或无线方式的通信。通信组件所在设备可以接入基于通信标准的无线网络，如 WiFi，2G、3G、4G/LTE、5G 等移动通信网络，或它们的组合。在一个示例性实施例中，通信组件经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件还包括近场通信（NFC）模块，以促进短程通信。例如，在 NFC 模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

上述图 5 中的显示器包括屏幕，其屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。

上述图 5 和图 6 中的电源组件，为电源组件所在设备的各种组件提供电力。电源组件可以包括电源管理系统，一个或多个电源，及其他与为电源组件所在

设备生成、管理和分配电力相关联的组件。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任

何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带, 磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读介质不包括暂存电脑可读媒体 (transitory media), 如调制的数据信号和载波。

10 还需要说明的是, 术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含, 从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素, 而且还包括没有明确列出的其他要素, 或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下, 由语句“包括一个……”限定的要素, 并不排除在包括所述要素的过程、方法、商品或者设备中
15 还存在另外的相同要素。

以上所述仅为本申请的实施例而已, 并不用于限制本申请。对于本领域技术人员来说, 本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等, 均应包含在本申请的权利要求范围之内。

权利要求书

1、一种音频信号处理方法，其特征在于，包括：

获取麦克风阵列采集到的当前音频信号，所述麦克风阵列包含至少两个麦克风；

根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；

根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

2、根据权利要求1所述的方法，其特征在于，根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息，包括：

根据所述至少两个麦克风采集到当前音频信号的相位差信息，计算当前音频信号对应的波达谱图，所述波达谱图反应当前声源在空间上的分布情况。

3、根据权利要求2所述的方法，其特征在于，根据所述至少两个麦克风采集到当前音频信号的相位差信息，计算当前音频信号对应的波达谱图，包括：

针对位置空间中的任一方位，对任意两个麦克风采集到当前音频信号的相位差信息进行累加，得到所述方位作为当前声源位置的概率；

根据所述位置空间中各方位作为当前声源位置的概率，生成当前音频信号对应的波达谱图。

4、根据权利要求1-3任一项所述的方法，其特征在于，根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音，包括：

计算所述当前声源空间分布信息的峰值信息，作为隐马尔科夫模型 HMM 的当前观测状态，并将单一语音与重叠语音作为 HMM 的两个隐藏状态；

将当前观测状态输入 HMM，结合 HMM 学习到的两个隐藏状态之间的跳转

关系，以历史观测状态为前提条件，计算当前观测状态对应隐藏状态的概率；

根据所述当前观测状态对应隐藏状态的概率，识别当前音频信号是否为重叠语音。

5、根据权利要求 1-3 任一项所述的方法，其特征在于，还包括：

- 5 若识别到当前音频信号为重叠语音，则根据所述当前声源空间分布信息，确定至少两个有效声源方位；

对所述至少两个有效声源方位上的音频信号进行语音增强，并对所述至少两个有效声源方位上增强后的音频信号分别进行语音识别。

- 6、根据权利要求 5 所述的方法，其特征在于，根据所述当前声源空间分布
10 信息，确定至少两个有效声源方位，包括：

在所述当前声源空间分布信息包括各方位作为当前声源位置的概率的情况下，将作为当前声源位置的概率最大的两个方位作为有效声源方位。

7、根据权利要求 6 所述的方法，其特征在于，还包括：

- 15 若识别到当前音频信号为单一语音，则将作为当前声源位置的概率最大的一个方位作为有效声源方位；

对所述有效声源方位上的音频信号进行语音增强，并对所述有效声源方位上增强后的音频信号进行语音识别。

8、根据权利要求 1-3 任一项所述的方法，其特征在于，在识别当前音频信号是否为重叠语音之前，还包括：

- 20 根据所述当前的声源空间分布信息，计算当前音频信号的波达方向；

根据所述波达方向，从所述至少两个麦克风中选择一路麦克风作为目标麦克风；

对所述目标麦克风采集到的当前音频信号进行语音端点检测 VAD，以得到当前音频信号是否为语音信号。

- 25 9、一种音频信号处理方法，其特征在于，适用于会议设备，所述会议设备包括麦克风阵列，所述方法包括：

获取所述麦克风阵列在会议场景中采集到的当前会议信号，所述麦克风阵

列包含至少两个麦克风；

根据所述至少两个麦克风采集到当前会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；

根据所述当前声源空间分布信息，结合基于历史会议信号学习到的单一语音与重叠语音之间的转换关系，识别当前会议信号是否为重叠语音。

10、一种音频信号处理方法，其特征在于，适用于教学设备，所述教学设备包括麦克风阵列，所述方法包括：

获取所述麦克风阵列在教学环境中采集到的当前课堂信号，所述麦克风阵列包含至少两个麦克风；

10 根据所述至少两个麦克风采集到当前课堂信号的相位差信息，生成当前课堂信号对应的当前声源空间分布信息；

根据所述当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

11、一种音频信号处理方法，其特征在于，适用于智能车载设备，所述智能车载设备包括麦克风采集阵列，所述方法包括：

获取所述麦克风阵列在车载环境中采集到的当前音频信号，所述麦克风阵列包含至少两个麦克风；

根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；

20 根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

12、一种终端设备，其特征在于，包括：存储器、处理器以及麦克风阵列；

所述存储器，用于存储计算机程序；

25 所述处理器，与所述存储器耦合，用于执行所述计算机程序，以用于：获取麦克风阵列采集到的当前音频信号，所述麦克风阵列包含至少两个麦克风；根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信

号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

13、根据权利要求 12 所述的终端设备，其特征在于，所述终端设备为会议
5 设备、拾音设备、机器人、智能机顶盒、智能电视、智能音箱以及智能车载设备。

14、一种会议设备，其特征在于，包括：存储器、处理器以及麦克风阵列；
所述存储器，用于存储计算机程序；

所述处理器，与所述存储器耦合，用于执行所述计算机程序，以用于：获
10 取所述麦克风阵列在会议场景中采集到的当前会议信号，所述麦克风阵列包含至少两个麦克风；根据所述至少两个麦克风采集到当前会议信号的相位差信息，生成当前会议信号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史会议信号学习到的单一语音与重叠语音之间的转换关系，识别当前会议信号是否为重叠语音。

15 15、一种教学设备，其特征在于，包括：存储器、处理器以及麦克风阵列；
所述存储器，用于存储计算机程序；

所述处理器，与所述存储器耦合，用于执行所述计算机程序，以用于：获
取所述麦克风阵列在教学环境中采集到的当前课堂信号，所述麦克风阵列包含至少两个麦克风；根据所述至少两个麦克风采集到当前课堂信号的相位差信息，
20 生成当前课堂信号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史课堂信号学习到的单一语音与重叠语音之间的转换关系，识别当前课堂信号是否为重叠语音。

16、一种智能车载设备，其特征在于，包括：存储器、处理器以及麦克风阵列；

25 所述存储器，用于存储计算机程序；

所述处理器，与所述存储器耦合，用于执行所述计算机程序，以用于：获取所述麦克风阵列在车载环境中采集到的当前音频信号，所述麦克风阵列包含

至少两个麦克风；根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

- 5 17、一种音频信号处理系统，其特征在于，包括：终端设备和服务端设备；所述终端设备包括麦克风阵列，所述麦克风阵列包含至少两个麦克风，用于采集当前音频信号；所述终端设备用于将所述至少两个麦克风采集的当前音频信号上传至所述服务端设备；

所述服务端设备，用于根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号是否为重叠语音。

- 18、一种服务端设备，其特征在于，包括：存储器和处理器；

所述存储器，用于存储计算机程序；

- 15 所述处理器，与所述存储器耦合，用于执行所述计算机程序，以用于：

接收终端设备上传的麦克风阵列中至少两个麦克风采集的当前音频信号；根据所述至少两个麦克风采集到当前音频信号的相位差信息，生成当前音频信号对应的当前声源空间分布信息；根据所述当前声源空间分布信息，结合基于历史音频信号学习到的单一语音与重叠语音之间的转换关系，识别当前音频信号
20 号是否为重叠语音。

19、一种存储有计算机程序的计算机可读存储介质，其特征在于，当所述计算机程序被处理器执行时，致使所述处理器实现权利要求 1-11 任一项所述方法中的步骤。

- 20、一种计算机程序产品，包括计算机程序/指令，其特征在于，当所述计算机程序/指令被处理器执行时，致使所述处理器实现权利要求 1-11 任一项所述方法中的步骤。