(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2023 年 5 月 25 日 (25.05.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/088091 A1

(51) 国际专利分类号:
G10L 21/0272 (2013.01) *G10L 21/0308* (2013.01)

(21) 国际申请号: PCT/CN2022/129118

(22) 国际申请日: 2022 年 11 月 2 日 (02.11.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111386550.8 2021年11月22日 (22.11.2021) CN

(71) 申请人: 北京字跳网络技术有限公司 (BEIJING ZITIAO NETWORK TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市海淀区紫金数码园4号楼2层0207, Beijing 100190 (CN)。

(72) 发明人: 汪鑫 (WANG, Xin); 中国北京市海淀区知春路63号中国卫星通信大厦今日头条小邮局, Beijing 100086 (CN)。

(74) 代理人: 中国贸促会专利商标事务所有限公司 (CCPIT PATENT AND TRADEMARK LAW OFFICE); 中国北京市复兴门内大街158号远洋大厦F10层, Beijing 100031 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,

WO 2023/088091 A1

(54) **Title:** VOICE SEPARATION METHOD AND APPARATUS, ELECTRONIC DEVICE, AND READABLE STORAGE MEDIUM

(54) **发明名称:** 语音分离方法、装置、电子设备及可读存储介质

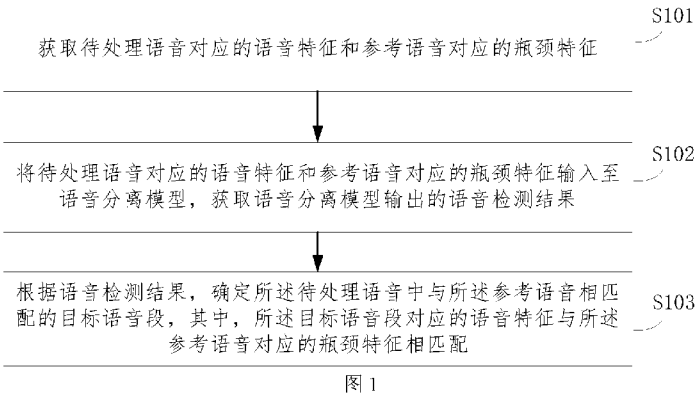

图 1

S101 Obtain a voice feature corresponding to a voice to be processed and a bottleneck feature corresponding to a reference voice

S102 Input, into a voice separation model, the voice feature corresponding to the voice to be processed and the bottleneck feature corresponding to the reference voice and obtain a voice detection result output by the voice separation model

S103 Determine, according to the voice detection result, a target voice segment matching the reference voice in the voice to be processed, wherein the voice feature corresponding to the target voice segment matches the bottleneck feature corresponding to the reference voice

(57) **Abstract:** A voice separation method and apparatus, an electronic device, and a readable storage medium. The method comprises: obtaining a voice feature corresponding to a voice to be processed and a bottleneck feature corresponding to a reference voice (S101); inputting, into a voice separation model, the voice feature corresponding to the voice to be processed and the bottleneck feature corresponding to the reference voice and obtaining a voice detection result output by the voice separation model (S102); and determining, on the basis of the voice detection result, a target voice segment matching the reference voice in the voice to be processed, wherein

[见续页]

PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

the voice feature of the target voice segment matches the bottleneck feature of the reference voice (S103). According to the method, the voice feature corresponding to the voice to be processed and the bottleneck feature corresponding to the reference voice are used as the joint input of a voice separation system, and a target voice is separated from the voice to be processed, so that the accuracy of the separated target voice can be improved.

(57) 摘要：一种语音分离方法、装置、电子设备及可读存储介质，该方法包括:获取待处理语音对应的语音特征和参考语音对应的瓶颈特征(S101)；将待处理语音对应的语音特征和参考语音对应的瓶颈特征输入至语音分离模型，获取语音分离模型输出的语音检测结果(S102);基于语音检测结果，确定待处理语音中与参考语音相匹配的目标语音段，其中，目标语音段的语音特征与参考语音的瓶颈特征相匹配(S103)。该方法将待处理语音对应的语音特征和参考语音对应的瓶颈特征作为语音分离系统的联合输入，从待处理语音中分离出目标语音，能够提高分离出的目标语音的准确度。

语音分离方法、装置、电子设备及可读存储介质

相关申请的交叉引用

本申请是以申请号为202111386550.8，申请日为2021年11月22日的中国申请为基础，并主张其优先权，该中国申请的公开内容在此作为整体引入本申请中。

技术领域

本公开涉及语音处理技术领域，尤其涉及一种语音分离方法、装置、电子设备及可读存储介质。

背景技术

语音活动检测（Voice Activity Detection，VAD）技术被广泛应用于语音识别的前端，用于检测语音与非语音。在一些场景下，不仅需要检测语音与非语音，还需要分离出目标语音。

相关技术中，进行语音分离是通过“VAD系统+语音识别分离系统”实现，具体地，首先，通过VAD系统对语音进行端点检测，再利用语音识别分离系统基于端点检测结果，分离出目标语音。

发明内容

本公开提供了一种语音分离方法、装置、电子设备及可读存储介质。

第一方面，本公开提供了一种语音分离方法，包括：

获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；

将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果；

根据所述语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段的语音特征与所述参考语音的瓶颈特征相匹配。

作为一种可能的实施方式，所述将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果，包括：

将所述待处理语音对应的语音特征输入至所述语音分离模型包括的第一神经

网络，获取所述第一神经网络输出的所述语音特征对应的向量表达；

将所述语音特征对应的向量表达和所述参考语音对应的瓶颈特征进行拼接，获得融合特征；

将所述融合特征输入至所述语音分离模型包括的第二神经网络，获取所述第二神经网络输出的矩阵，且基于所述矩阵获取所述语音检测结果。

作为一种可能的实施方式，所述基于所述矩阵获取所述语音检测结果，包括：

根据所述矩阵包括的各音频帧对应的元素，获取各所述音频帧分别属于第一类别和第二类别的概率值；所述第一类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征匹配，所述第二类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征不匹配；

根据各所述音频帧分别属于第一类别和第二类别的概率值中的最大值，确定各所述音频帧对应的语音检测结果，所述音频帧对应的语音检测结果用于指示所述音频帧对应的语音特征与所述参考语音对应的瓶颈特征是否匹配。

作为一种可能的实施方式，所述语音分离模型是基于样本语音对应的语音特征、所述样本语音对应的瓶颈特征以及标注的所述样本语音的语音检测结果进行训练获得的，所述样本语音包括所述参考语音。

作为一种可能的实施方式，所述语音特征包括 FBank 特征、梅尔频谱特征以及音色特征中的一项或多项。

第二方面，本公开提供了一种语音分离装置，包括：

获取模块，用于获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；

语音检测模块，用于将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果；

分离模块，用于根据所述语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段的语音特征与所述参考语音的瓶颈特征相匹配。

作为一种可能的实施方式，语音检测模块，具体用于将所述待处理语音对应的语音特征输入至所述语音分离模型包括的第一神经网络，获取所述第一神经网络输出的所述语音特征对应的向量表达；将所述语音特征对应的向量表达和所述

参考语音对应的瓶颈特征进行拼接，获得融合特征；将所述融合特征输入至所述语音分离模型包括的第二神经网络，获取所述第二神经网络输出的矩阵，且基于所述矩阵获取所述语音检测结果。

作为一种可能的实施方式，语音检测模块，具体用于根据所述矩阵包括的各所述音频帧对应的元素，获取各所述音频帧分别属于第一类别和第二类别的概率值；所述第一类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征匹配，所述第二类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征不匹配；根据各所述音频帧分别属于第一类别和第二类别的概率值中的最大值，确定各所述音频帧对应的语音检测结果；

其中，所述音频帧对应的语音检测结果用于指示所述音频帧对应的语音特征与所述参考语音对应的瓶颈特征是否匹配。

作为一种可能的实施方式，所述语音分离模型是基于样本语音对应的语音特征、所述样本语音对应的瓶颈特征以及标注的所述样本语音的语音检测结果进行训练获得的，所述样本语音包括所述参考语音。

作为一种可能的实施方式，所述语音特征包括 FBank 特征、梅尔频谱特征以及音色特征中的一项或多项。

第三方面，本公开提供了一种电子设备，包括：存储器和处理器；

所述存储器被配置为存储计算机程序指令；

所述处理器被配置为执行所述计算机程序指令，使得所述电子设备实现如第一方面任一项所述的语音分离方法。

第四方面，本公开提供了一种可读存储介质，包括：计算机程序指令；

电子设备的至少一个处理器执行所述计算机程序指令，已实现第一方面任一项所述的语音分离方法。

第五方面，本公开提供一种计算机程序产品，当所述计算机程序产品被计算机执行时，使得所述计算机实现如第一方面任一项所述的语音分离方法。

本公开实施例提供一种语音分离方法、装置、电子设备及可读存储介质，该方法包括：获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，

获取语音分离模型输出的语音检测结果；基于语音检测结果，确定待处理语音中与参考语音相匹配的目标语音段，其中，所述目标语音段的语音特征与所述参考语音的瓶颈特征相匹配。本公开实施例，用于实现语音分离的系统采用端到端的方式实现，且通过将待处理语音对应的语音特征和参考语音对应的瓶颈特征作为语音分离系统的联合输入，用于从待处理语音中分离出目标语音。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。

为了更清楚地说明本公开实施例或相关技术中的技术方案，下面将对实施例或相关技术描述中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本公开一实施例提供的语音分离方法的流程示意图；

图 2 为本公开一实施例提供的语音分离模型的结构示意图；

图 3 为本公开另一实施例提供的语音分离方法的流程示意图；

图 4 为本公开一实施例提供的语音分离装置的结构示意图；

图 5 为本公开一实施例提供的电子设备的结构示意图。

具体实施方式

为了能够更清楚地理解本公开的上述目的、特征和优点，下面将对本公开的方案进行进一步描述。需要说明的是，在不冲突的情况下，本公开的实施例及实施例中的特征可以相互组合。

在下面的描述中阐述了很多具体细节以便于充分理解本公开，但本公开还可以采用其他不同于在此描述的方式来实施；显然，说明书中的实施例只是本公开的一部分实施例，而不是全部的实施例。

VAD 系统的性能以及语音识别分离系统的性能均对语音分离结果的准确度影响极大，因此，采用上述语音分离系统时，VAD 系统和语音识别分离系统需要分离训练，较为复杂。

示例性地，本公开提供的语音分离方法可以由本公开提供的语音分离装置实现。语音分离装置可以通过任意的软件和/或硬件的方式实现。示例性地，语音分离装置可以为：平板电脑、手机（如折叠屏手机、大屏手机等）、可穿戴设备、车载设备、增强现实（augmented reality，AR）/虚拟现实（virtual reality，VR）设备、笔记本电脑、超级移动个人计算机（ultra-mobile personal computer，UMPC）、上网本、个人数字助理（personal digital assistant，PDA）、智能电视、智慧屏、高清电视、4K 电视、智能音箱、智能投影仪等物联网（the internet of things，IOT）设备，本公开对电子设备的具体类型不作任何限制。

下面以电子设备执行语音分离方法为例，结合几个具体实施例，详细介绍本公开提供的语音分离方法。

图 1 为本公开一实施例提供的语音分离方法的流程示意图。参照图 1 所示，本实施例提供的方法包括：

S101、获取待处理语音对应的语音特征和参考语音对应的瓶颈特征。

本公开对于待处理语音的时长、存储格式、内容等等参数不做限定。电子设备能够获取待处理语音对应的语音特征，其中，语音特征可以包括FBank特征（filter bank 特征）、梅尔频谱特征、瓶颈特征（bottleneck 特征）以及音色特征（pitch 特征）中的一项或多项。

由于获取 FBank 特征的各滤波器之间相互重叠，因此，FBank 特征中每个维度的特征之间相关性较高，将 FBank 特征作为待处理语音的语音特征，语音分离系统能够利用 FBank 特征中每个维度的特征之间的相关性，输出较为准确的语音检测结果。

梅尔频谱特特征是在 mel 域进行的特征提取，mel 域更接近于人类听觉系统，从而更能够准确地表示声音。

音高特征（pitch 特征），音高是一种感知属性，允许在与频率相关的尺度上对声音进行排序。音高可以量化为频率，称为基频 (F0)。音高变化形成了声调语言的声调，是重要的说话人识别、语音识别特征。

此外，本公开对于电子设备获取待处理语音对应的语音特征的实现方式不做限定，例如，电子设备可以采用语音特征提取模型对待处理语音进行特征提取，

语音特征提取模型例如为 ASR 模型，或者，电子设备也可以利用数字处理技术对待处理语音的信号进行转换，获取待处理语音对应的语音特征。

瓶颈（bottleneck）是一种非线性的特征转换技术以及有效的降维技术。瓶颈特征可以包括韵律、内容等维度的信息。本公开中，参考语音对应的瓶颈特征主要用于实现区分待处理语音中哪些音频帧为要分离的目标音频帧，其中，电子设备可以基于一个单独的神经网络对参考语音进行特征提取，获取参考语音对应的瓶颈特征。

S102、将待处理语音对应的语音特征和参考语音对应的瓶颈特征输入至语音分离模型，获取语音分离模型输出的语音检测结果。

其中，语音检测结果

其中，若待处理语音包括的音频帧与参考语音匹配，则表示该音频帧是需要分离出来的目标音频帧。若待处理语音包括的音频帧与参考语音不匹配，则表示该音频帧不是需要分离出来的目标音频帧。

一种可能的实施方式，电子设备可以通过预先训练好的语音分离模型对待处理语音进行语音检测，并输出语音检测结果。具体地，可将待处理语音对应的语音特征和参考语音对应的瓶颈特征作为语音分离模型的输入，语音分离模型可以输出待处理语音的每一个音频帧是否与参考语音匹配的分类标签，该分类标签即为语音检测结果。

另一种可能的实施方式，对于待处理语音，采用与瓶颈特征相似的处理流程和方式，提取到待处理语音包括的每一音频帧的瓶颈特征，再通过计算每一音频帧的瓶颈特征与瓶颈特征之间的相似度，确定音频帧对应的分类结果。其中，相似度的计算方式可以但不限于 consine 距离，内积等，之后，再通过预预设距离阈值，将每一个音频帧对应的距离值与预设距离阈值进行对比得到该音频帧对应的分类结果。

即，在该实施方式中，待处理语音对应的语音特征即为待处理语音对应的瓶颈特征。

S103、根据语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段对应的语音特征与所述参考语音对应的瓶颈

特征相匹配。

电子设备可以根据每一音频帧对应的语音检测结果，从待处理语音中确定出与参考语音相匹配的目标语音段。其中，目标语音段对应的语音特征与参考语音对应的瓶颈特征相匹配，即表示目标语音段的音色与参考语音的音色匹配度满足要求。

本实施例提供的方法，获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；将待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入语音分离模型，获取语音分离模型输出的语音检测结果；基于语音检测结果，确定待处理语音中与参考语音匹配的目标语音段。本公开实施例，用于实现语音分离的系统采用端到端的方式实现，且通过将待处理语音对应的语音特征和参考语音对应的瓶颈特征作为语音分离系统的联合输入，用于从待处理语音中分离出目标语音，还能够提高分离出的目标语音的准确度。

图 2 为本公开一实施例提供的语音分离模型的结构示意图。参照图 2 所示，本实施例提供的语音分离模型 200 包括：第一神经网络 201、特征融合层 202、第二神经网络 203 以及分类层 204。

其中，第一神经网络 201 的输出端与特征融合层 202 的输入端连接，特征融合层 202 的输出端与第二神经网络 203 的输入端连接，第二神经网络 203 的输出端与分类层 204 连接。

本公开对于第一神经网络 201 以及第二神经网络 203 的网络结构、类型等参数不做限定。例如，第一神经网络 201 和第二神经网络 203 可以为前馈神经网络（FeedForward Neural Network）、卷积神经网络(Convlution Neural Network，CNN)、深层前馈记忆网络（Deep Feed Forward Sequential Memory Networks，DFSMN）、transformer、conformer 等等。

在实际应用中，可以通过试验选取性能最优的网络结构作为第一神经网络 201 以及第二神经网络 203。此外，还需要说明的是，第一神经网络 201 和第二神经网络 203 的类型可以相同，也可以不同，本公开对此不做限定。

第一神经网络 201 主要用于接收待处理语音的语音特征作为输入，并对语音特征进行降维等处理，获得语音特征对应的向量表达。其中，语音特征表示为 F1，

$F1 \in R^{N \times k1}$ 维；参考语音对应的瓶颈特征表示为 F2，$F2 \in R^{l \times k2}$ 维，其中，N 表示待处理语音的音频帧的总数。

由于本公开提供的方法是需要将待处理语音对应的语音特征与参考语音对应的瓶颈特征作为联合输入，如果直接将待处理语音对应的语音特征与参考语音对应的瓶颈特征进行拼接，由于两个特征维度不同，则无法拼接，也无法进行后续的语音检测流程，因此，需要对待处理语音对应的语音特征进行降维等处理。

之后，将第一神经网络 201 的输出与参考语音对应的瓶颈特征输入至特征融合层 202，以将上述语音特征的向量表达与参考语音对应的瓶颈特征进行拼接，获得特征融合层 202 输出的融合特征。本公开对于拼接的实现方式不做限定。

融合特征作为第二神经网络 203 的输入，第二神经网络 203 通过对融合特征进行空间映射，获得目标维度的矩阵，其中，第二神经网络 203 输出的矩阵为 N*2 维，其中，N 为待处理语音的音频帧的帧数。

矩阵中每个 1*2 的子矩阵对应 1 个音频帧，基于每个音频帧对应的子矩阵。将上述第二神经网络 203 输出的矩阵输入至分类层 204，以使分类层采用预设的分类函数，依据矩阵进行计算，输出每个音频帧的语音检测结果。

示例性地，分类层 204 采用 softmax 函数对每个音频帧对应的子矩阵进行计算，获得音频帧分别属于第一类别和第二类别的概率，再基于音频帧分别属于第一类别和第二类别的概率中的最大值确定音频帧是否与参考语音匹配。其中，第一类别包括的音频帧与参考语音匹配，第二类别包括的音频帧与参考语音不匹配。

其中，分类层 204 输出的每个音频帧对应的语音检测结果可以通过 0、1 向量表达，当语音检测结果为 0，则表示该音频帧对应的语音特征与参考语音对应的瓶颈特征不匹配，即，该音频帧与参考语音不匹配；当语音检测结果为 1，则表示该音频帧对应的语音特征与参考语音对应的瓶颈特征匹配，即，该音频帧与参考语音匹配。

之后，可以基于每个音频帧的语音检测结果，从待处理语音中确定出与参考语音匹配的目标音频段。例如，可将待处理语音中所有语音检测结果为 1 的音频帧确定为目标音频段包括的音频帧，继而可以分离出目标音频段。

还需要说明的是，由于待处理语音可能包括多个语音角色的语音段，每个语

音角色也可能对应多个语音段，且多个语音段不连续，因此，目标音频段可以包括一个或多个语音段。

在图 2 所示实施例的基础上，可知，与采用“VAD 系统+语音识别分离系统”相结合的方式进行语音分离相比，本公开提供的语音分离模型能够通过端到端的方式实现语音分离，本公开提供的语音分离模型的训练过程更加简单。且本公开提供的语音分离方法是将待处理语音的语音特征和参考语音的瓶颈特征作为语音分离模型的联合输入，用于实现语音分离，能够提高语音分离的准确度。

结合前述图 1 以及图 2 所示实施例可知，本公开提供的语音分离方法可以通过语音分离模型实现，而语音分离模型是需要预先经过训练的，下面通过图 3 所示实施例介绍如何训练获取语音分离模型

图 3 为本公开另一实施例提供的语音分离方法的流程示意图。参照图 3 所示，本实施例提供的方法包括：

S301、获取样本语音对应的语音特征、样本语音对应的瓶颈特征以及标注的样本语音的语音检测结果。

本方案中，样本语音包括上述参考语音，即在训练的过程中，要保证语音分离模型学习过参考语音的特征，从而保证训练好的语音分离模型能够正确地识别出与参考语音匹配的音频帧。

本公开对于样本语音和参考语音的数量、内容等参数不做限定。

一种可能的实现方式，获取样本语音的语音特征以及样本语音的瓶颈特征的实现方式可以参考前述图 1 所示实施例中 S101 的相关描述。

另一种可能的实现方式，若一些相关的数据库中预先存储有样本语音以及样本语音的语音特征、样本语音对应的瓶颈特征以及标注的样本语音的语音检测结果等等，也可以在对语音分离模型进行训练时，从数据库中读取。

S302、将样本语音对应的语音特征和样本语音对应的瓶颈特征，输入至语音分离模型中，获取语音分离模型的预测的样本语音的语音检测结果。

S303、基于标注的样本语音的语音检测结果以及预测的样本语音的语音检测结果，对语音分离模型进行优化，直至满足预设收敛条件，获取训练好的语音分离模型。

可以基于标注的样本语音的语音检测结果以及预测的样本语音的语音检测结果，采用预设的损失函数计算相应的损失值，根据损失值确定是否满足预设收敛条件，若满足预设收敛条件，则结束训练，若不满足预设收敛条件，则基于损失值对语音分离模型的相关参数进行优化。接着，再通过样本语音进行下一轮训练，直至满足预设收敛条件，则获得训练好的语音分离模型。

本公开对于预设收敛条件的实现方式不做限定。示例性地，预设收敛条件可以是训练迭代次数、损失阈值等评价指标。还需要说明的是，本公开对于预设损失函数不做限定。

在训练的过程中，可以先利用样本语音中非参考语音的相关数据对语音分离模型进行训练，使语音分离模型具备一定的语音分离的能力，在此基础上，再基于参考语音的相关数据，对语音分离模型进行训练，使语音分离模型具有能够分离出与参考语音相匹配的音频帧的能力。

在对语音模型进行训练的过程中，还可以将参考语音的语音特征和其他样本语音的瓶颈特征进行组合，作为语音分离模型的输入，以使语音分离模型进行学习。

这样训练不仅能够使语音分离模型学习到正确的样本，还能够学习到错误的样本，从而提高语音分离模型的性能。

本实施例提供的方法，通过获取样本语音对应的语音特征、样本语音对应的瓶颈特征以及标注的样本语音的语音检测结果；将样本语音对应的语音特征和样本语音对应的瓶颈特征，输入至语音分离模型中，获取语音分离模型的预测的样本语音的语音检测结果；基于标注的样本语音的语音检测结果以及预测的样本语音的语音检测结果，对语音分离模型进行优化，直至满足预设收敛条件，获取训练好的语音分离模型。

示例性地，本公开还提供一种语音分离装置。

图 4 为本公开一实施例提供的语音分离装置的结构示意图。参照图 4 所示，本实施例提供的语音分离装置 400 包括：

获取模块 401，用于获取待处理语音对应的语音特征和参考语音对应的瓶颈特征.

语音检测模块 402，用于将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取语音分离模型输出的语音检测结果。

分离模块 403，用于根据所述语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段的语音特征与所述参考语音的瓶颈特征相匹配。

作为一种可能的实施方式，所述语音检测结果用于指示所述待处理语音包括的各音频帧的语音特征是否与所述参考语音的瓶颈特征相匹配。

作为一种可能的实施方式，语音检测模块 402，具体用于将所述待处理语音对应的语音特征输入至所述语音分离模型包括的第一神经网络，获取所述第一神经网络输出的所述语音特征对应的向量表达；将所述语音特征对应的向量表达和所述参考语音对应的瓶颈特征进行拼接，获得融合特征；将所述融合特征输入至所述语音分离模型包括的第二神经网络，获取所述第二神经网络输出的矩阵，且基于所述矩阵获取所述语音检测结果。

作为一种可能的实施方式，语音检测模块 402，具体用于根据所述矩阵包括的各所述音频帧对应的元素，获取各所述音频帧分别属于第一类别和第二类别的概率值；所述第一类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征匹配，所述第二类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征不匹配；根据各所述音频帧分别属于第一类别和第二类别的概率值中的最大值，确定各所述音频帧对应的语音检测结果。

其中，所述音频帧对应的语音检测结果用于指示该音频帧对应的瓶颈特征是否匹配。

作为一种可能的实施方式，所述语音分离模型是基于样本语音对应的语音特征、所述样本语音对应的瓶颈特征以及标注的所述样本语音的语音检测结果进行训练获得的，所述样本语音包括所述参考语音。

作为一种可能的实施方式，所述语音特征包括 FBank 特征、梅尔频谱特征、瓶颈特征以及音色特征中的一项或多项。

本实施例提供的语音分离装置可以用于实现如上任一方法实施例的技术方案，

其实现原理以及技术效果类似，可以参照前述方法实施例的详细描述，简明起见，此处不再赘述。

示例性地，本公开还提供一种电子设备。

图 5 为本公开一实施例提供的电子设备的结构示意图。参照图 5 所示，本实施例提供的电子设备 500 包括：存储器 501 和处理器 502。

其中，存储器 501 可以是独立的物理单元，与处理器 502 可以通过总线 503 连接。存储器 501、处理器 502 也可以集成在一起，通过硬件实现等。

存储器 501 用于存储程序指令，处理器 502 调用该程序指令，执行以上任一方法实施例的技术方案。

可选地，当上述实施例的方法中的部分或全部通过软件实现时，上述电子设备 500 也可以只包括处理器 502。用于存储程序的存储器 501 位于电子设备 500 之外，处理器 502 通过电路/电线与存储器连接，用于读取并执行存储器中存储的程序。

处理器 502 可以是中央处理器（central processing unit，CPU），网络处理器（network processor，NP）或者 CPU 和 NP 的组合。

处理器 502 还可以进一步包括硬件芯片。上述硬件芯片可以是专用集成电路（application-specific integrated circuit，ASIC），可编程逻辑器件（programmable logic device，PLD）或其组合。上述 PLD 可以是复杂可编程逻辑器件（complex programmable logic device，CPLD），现场可编程逻辑门阵列（field-programmable gate array，FPGA），通用阵列逻辑（generic array logic，GAL）或其任意组合。

存储器 501 可以包括易失性存储器（volatile memory），例如随机存取存储器（random-access memory，RAM）；存储器也可以包括非易失性存储器（non-volatile memory），例如快闪存储器（flash memory），硬盘（hard disk drive，HDD）或固态硬盘（solid-state drive，SSD）；存储器还可以包括上述种类的存储器的组合。

本公开还提供一种可读存储介质，包括：计算机程序指令；计算机程序指令被电子设备的至少一个处理器执行时，实现上述任一方法实施例所示的语音分离方法。

本公开还提供一种计算机程序产品，所述计算机程序产品被计算机执行时，

使得所述计算机实现上述任一方法实施例所述的语音分离方法。

需要说明的是，在本文中，诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、物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、物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、物品或者设备中还存在另外的相同要素。

以上所述仅是本公开的具体实施方式，使本领域技术人员能够理解或实现本公开。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本公开的精神或范围的情况下，在其它实施例中实现。因此，本公开将不会被限制于本文所述的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权 利 要 求

1、一种语音分离方法，包括：

获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；

将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果；

根据所述语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段对应的语音特征与所述参考语音对应的瓶颈特征相匹配。

2、根据权利要求 1 所述的方法，其中，所述将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果，包括：

将所述待处理语音对应的语音特征输入至所述语音分离模型包括的第一神经网络，获取所述第一神经网络输出的所述语音特征对应的向量表达；

将所述语音特征对应的向量表达和所述参考语音对应的瓶颈特征进行拼接，获得融合特征；

将所述融合特征输入至所述语音分离模型包括的第二神经网络，获取所述第二神经网络输出的矩阵，且基于所述矩阵获取所述语音检测结果。

3、根据权利要求 2 所述的方法，其中，所述基于所述矩阵获取所述语音检测结果，包括：

根据所述矩阵包括的各音频帧对应的元素，获取各所述音频帧分别属于第一类别和第二类别的概率值；所述第一类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征匹配，所述第二类别包括的音频帧对应的语音特征与所述参考语音对应的瓶颈特征不匹配；

根据各所述音频帧分别属于第一类别和第二类别的概率值中的最大值，确定各所述音频帧对应的语音检测结果。

4、根据权利要求 1 至 3 任一项所述的方法，其中，所述语音分离模型是基于样本语音对应的语音特征、所述样本语音对应的瓶颈特征以及标注的所述样本语

音的语音检测结果进行训练获得的，所述样本语音包括所述参考语音。

5、根据权利要求 1 至 4 任一项所述的方法，其中，所述语音特征包括 FBank 特征、梅尔频谱特征以及音色特征中的一项或多项。

6、一种语音分离装置，包括：

获取模块，被配置为获取待处理语音对应的语音特征和参考语音对应的瓶颈特征；

语音检测模块，被配置为将所述待处理语音对应的语音特征和所述参考语音对应的瓶颈特征输入至语音分离模型，获取所述语音分离模型输出的语音检测结果；

分离模块，被配置为根据所述语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段的语音特征与所述参考语音的瓶颈特征相匹配。

7、根据权利要求 6 所述的装置，其中，所述语音检测模块，进一步被配置为将所述待处理语音对应的语音特征输入至所述语音分离模型包括的第一神经网络，获取所述第一神经网络输出的所述语音特征对应的向量表达；将所述语音特征对应的向量表达和所述参考语音对应的瓶颈特征进行拼接，获得融合特征；将所述融合特征输入至所述语音分离模型包括的第二神经网络，获取所述第二神经网络输出的矩阵，且基于所述矩阵获取所述语音检测结果。

8、一种电子设备，包括：存储器和处理器；

所述存储器被配置为存储计算机程序指令；

所述处理器被配置为执行所述计算机程序指令，使得所述电子设备实现如权利要求 1 至 5 任一项所述的语音分离方法。

9、一种可读存储介质，包括：计算机程序指令；

所述计算机程序指令被电子设备的至少一个处理器执行时，使得所述电子设备实现如权利要求 1 至 5 任一项所述的语音分离方法。

10、一种计算机程序产品，当所述计算机程序产品被计算机运行时，使得所述计算机实现如权利要求 1 至 5 任一项所述的语音分离方法。

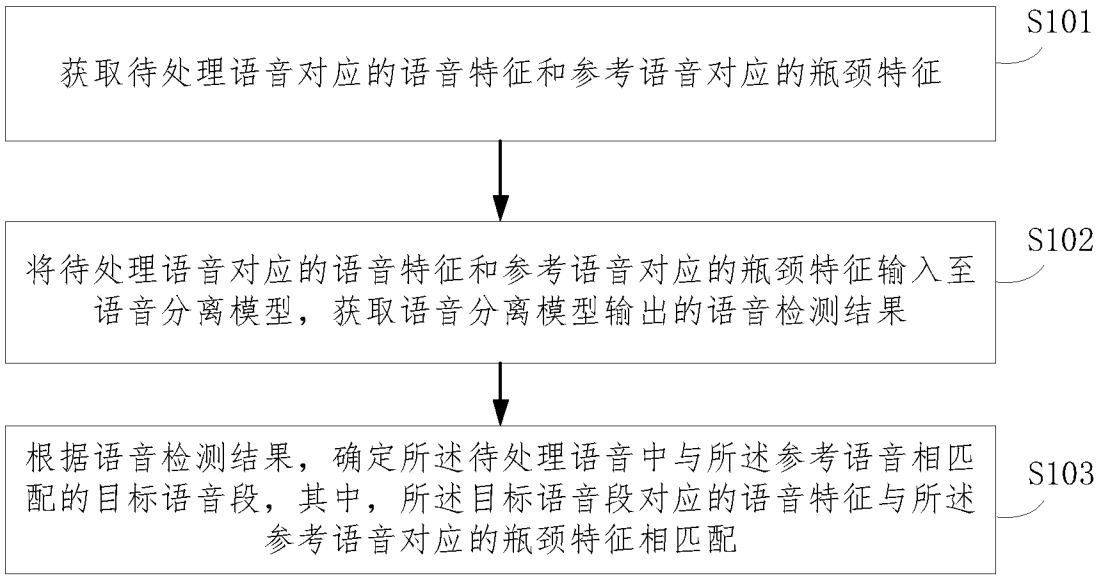
获取待处理语音对应的语音特征和参考语音对应的瓶颈特征
S101
将待处理语音对应的语音特征和参考语音对应的瓶颈特征输入至语音分离模型，获取语音分离模型输出的语音检测结果
S102
根据语音检测结果，确定所述待处理语音中与所述参考语音相匹配的目标语音段，其中，所述目标语音段对应的语音特征与所述参考语音对应的瓶颈特征相匹配
S103

图 1

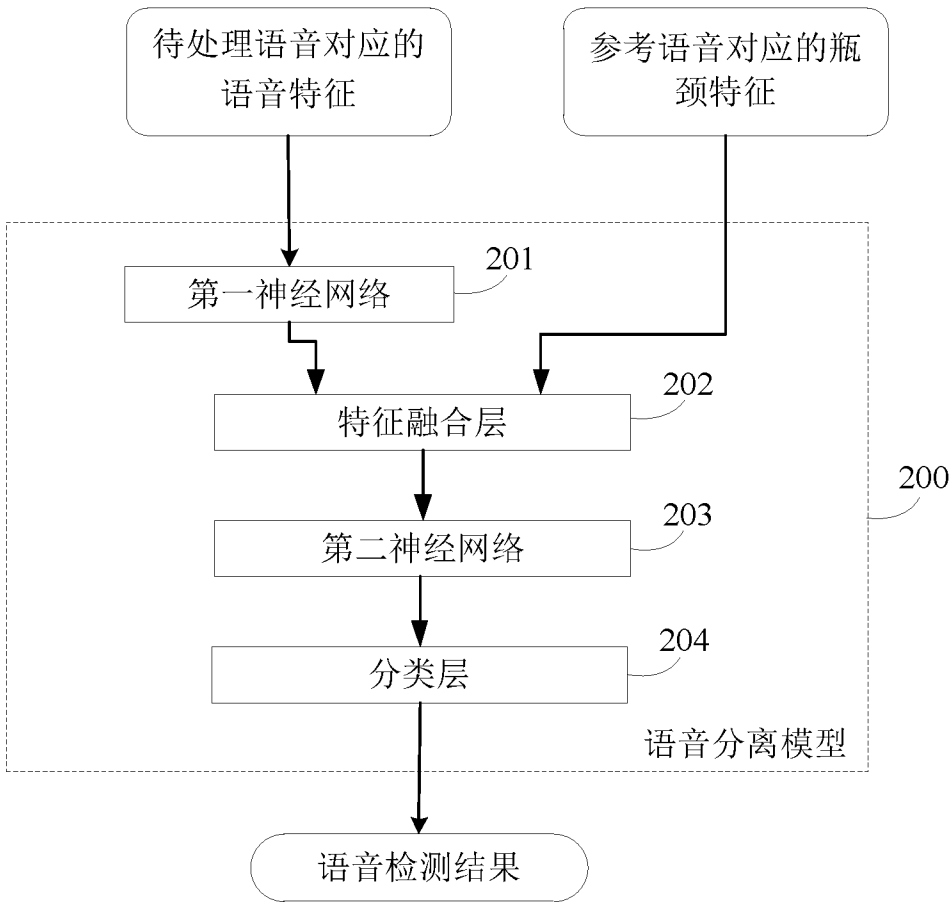
待处理语音对应的语音特征
参考语音对应的瓶颈特征
第一神经网络
201
特征融合层
202
第二神经网络
203
分类层
204
200
语音分离模型
语音检测结果

图 2

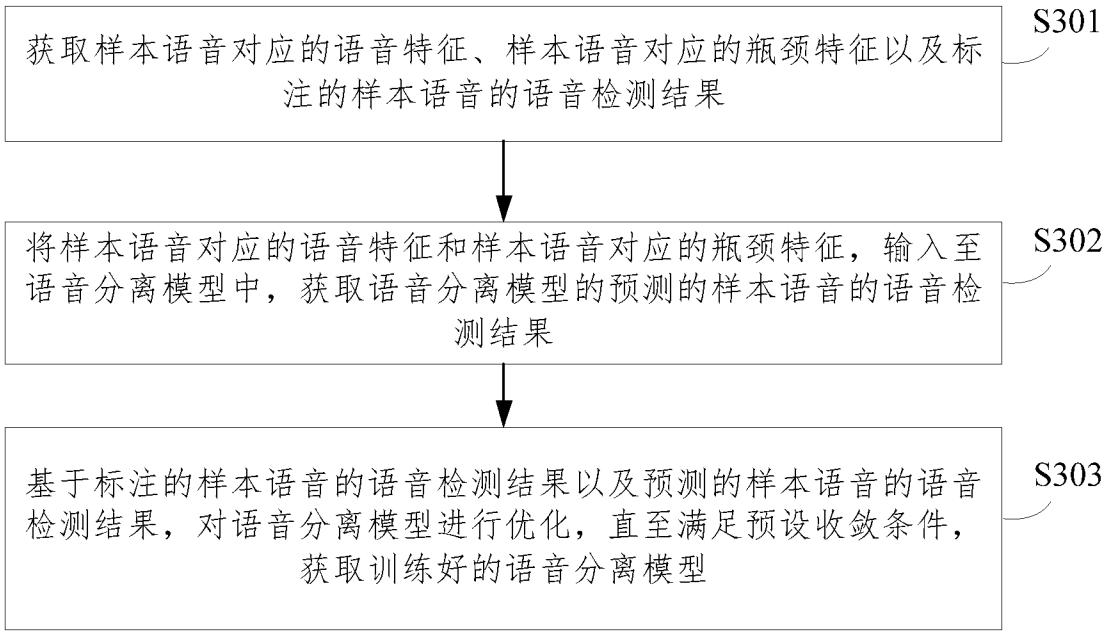
获取样本语音对应的语音特征、样本语音对应的瓶颈特征以及标注的样本语音的语音检测结果
S301
将样本语音对应的语音特征和样本语音对应的瓶颈特征，输入至语音分离模型中，获取语音分离模型的预测的样本语音的语音检测结果
S302
基于标注的样本语音的语音检测结果以及预测的样本语音的语音检测结果，对语音分离模型进行优化，直至满足预设收敛条件，获取训练好的语音分离模型
S303

图 3

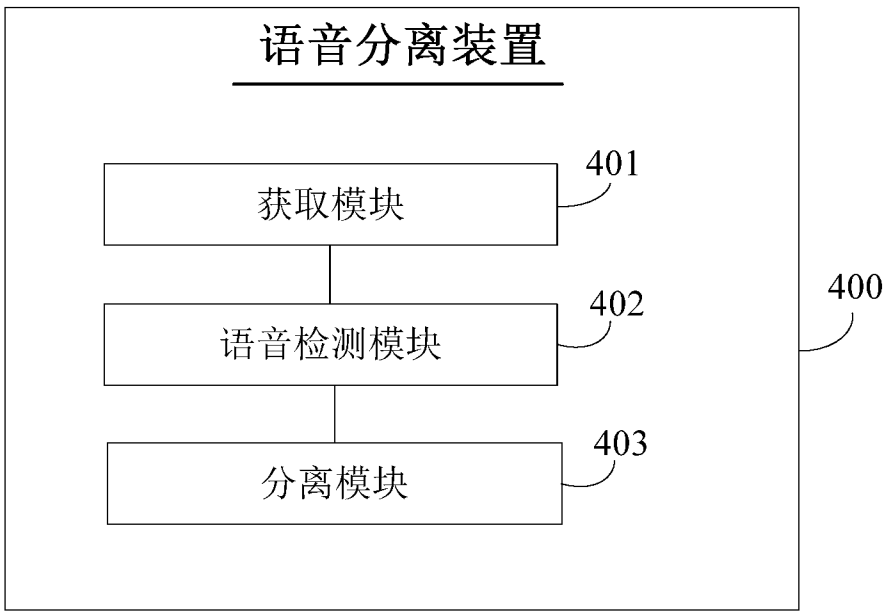
语音分离装置
401
获取模块
402
语音检测模块
403
分离模块
400

图 4

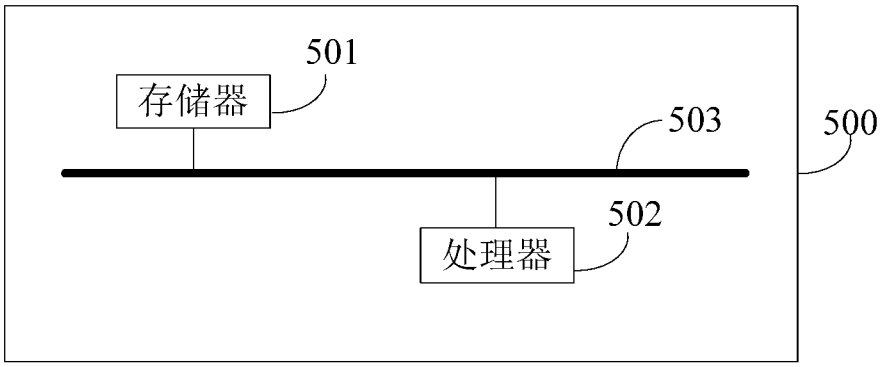

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/129118

A. CLASSIFICATION OF SUBJECT MATTER

G10L21/0272(2013.01)i;G10L21/0308(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L21, G10L17, G10L15

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, ENTXT, ENTXTC, DWPI, IEEE, CNKI: 端到端, 语音, 音频, 话音, 分离, 瓶颈, 特征, 层, 降维, 目标语音, 识别, Speech, voice, audio, separation, Bottleneck, +BN+, feature.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113113044 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 12 July 2021 (2021-07-12) description, paragraphs [0063]-[0084], [0095], and [0110]	1-10
A	CN 108461085 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 27 August 2018 (2018-08-27) entire document	1-10
A	CN 112466336 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 08 March 2021 (2021-03-08) entire document	1-10
A	CN 112241467 A (BEIJING MAGIC DATA CO., LTD.) 18 January 2021 (2021-01-18) entire document	1-10
A	US 2019/0325880 A (ID R&D, INC.) 23 October 2019 (2019-10-23) entire document	1-10

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2023.01.12	**10 February 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/129118

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113113044 A	12 July 2021	None	
CN 108461085 A	27 August 2018	None	
CN 112466336 A	08 March 2021	None	
CN 112241467 A	18 January 2021	None	
US 2019/0325880 A	23 October 2019	None	

国际检索报告

国际申请号
PCT/CN2022/129118

A. 主题的分类

G10L21/0272(2013.01)i;G10L21/0308(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L21, G10L17, G10L15

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT, ENTXT, ENTXTC, DWPI, IEEE, CNKI:端到端，语音，音频，话音，分离，瓶颈，特征，层，降维，目标语音，识别，Speech, voice, audio, separation, Bottleneck, +BN+, feature.

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 113113044 A (北京小米移动软件有限公司) 2021年7月12日 (2021-07-12) 说明书第[0063]-[0084]段，第[0095]段，第[0110]段）	1-10
A	CN 108461085 A (南京邮电大学) 2018年8月27日 (2018-08-27) 全文	1-10
A	CN 112466336 A (平安科技（深圳）有限公司) 2021年3月8日 (2021-03-08) 全文	1-10
A	CN 112241467 A (北京爱数智慧科技有限公司) 2021年1月18日 (2021-01-18) 全文	1-10
A	US 2019/0325880 A (ID R&D, Inc.) 2019年10月23日 (2019-10-23) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023.01.12	2023年2月10日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	谭雪艳
传真号 (86-10)62019451	电话号码 (+86) 010-62089709

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/129118

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 113113044 A	2021年7月12日	无	
CN 108461085 A	2018年8月27日	无	
CN 112466336 A	2021年3月8日	无	
CN 112241467 A	2021年1月18日	无	
US 2019/0325880 A	2019年10月23日	无	