

(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
2021 年 1 月 28 日 (28.01.2021)



(10) 国际公布号
WO 2021/013255 A1

(51) 国际专利分类号:
G10L 17/20 (2013.01) **G10L 17/04** (2013.01)

(21) 国际申请号: PCT/CN2020/104545

(22) 国际申请日: 2020 年 7 月 24 日 (24.07.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910673696.7 2019 年 7 月 24 日 (24.07.2019) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 昆山杜克大学 (DUKE KUNSHAN

UNIVERSITY) [CN/CN]; 中国江苏省昆山市杜克大道 8 号, Jiangsu 215316 (CN)。

(72) 发明人: 曾夕娟 (ZENG, Xijuan); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 周小鹏 (ZHOU, Xiaopeng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 芦宇 (LU, Yu); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 胡伟湘 (HU, Weixiang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 蔡丹蔚 (CAI, Danwei); 中国江苏省昆山市杜克大道 8 号, Jiangsu 215316 (CN)。 李明 (LI, Ming); 中国江苏省昆山市杜克大道 8 号, Jiangsu 215316 (CN)。

(54) Title: VOICEPRINT RECOGNITION METHOD AND APPARATUS

(54) 发明名称: 一种声纹识别方法及装置

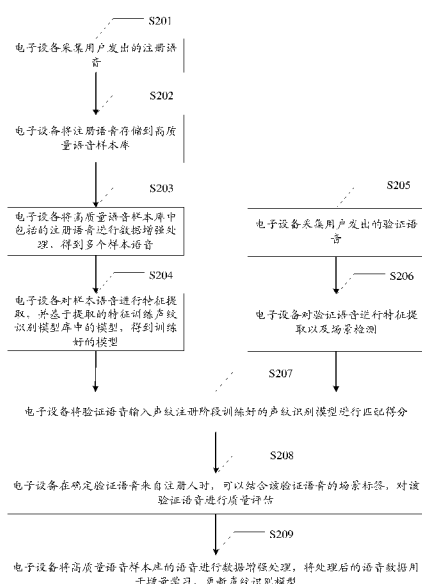


图 2

S201 An electronic device acquires a registration voice issued by a user
S202 The electronic device stores the registration voice in a high-quality voice sample library
S203 The electronic device performs data augmentation processing on the registration voice comprised in the high-quality voice sample library to obtain a plurality of sample voices
S204 The electronic device performs feature extraction on the sample voices, and trains, on the basis of an extracted feature, a model in a voiceprint recognition model library to obtain a trained model
S205 The electronic device acquires a verification voice issued by the user
S206 The electronic device performs feature extraction and scene detection on the verification voice
S207 The electronic device inputs the verification voice into the voiceprint recognition model, which is trained at a voiceprint registration stage, to perform matching and scoring
S208 When it is determined that the verification voice comes from a registered user, the electronic device can perform quality assessment on the verification voice in conjunction with a scene label of the verification voice
S209 The electronic device performs data augmentation processing on voices in the high-quality voice sample library, uses processed voice data for incremental learning, and updates the voiceprint recognition model

(57) Abstract: Provided are a voiceprint recognition method and apparatus. The method comprises: acquiring a registration voice input by a user (S201); generating, on the basis of the registration voice, a sample voice under a far-field condition (S203); training a voiceprint recognition model on the basis of the sample voice (S204); acquiring a verification voice input by the user (S205); inputting the verification voice into the voiceprint recognition model for matching to obtain a matching result (S207); and determining, on the basis of the matching result, whether the user is registered to the voiceprint recognition model (S208). The method relates to relevant

(74) 代理人: 北京同达信恒知识产权代理有限公司
(TDIP & PARTNERS); 中国北京市西城区裕民路
18号北环中心A座2002, Beijing 100029 (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,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, 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, 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, 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))。

fields, such as artificial intelligence, and is used for solving the problem of voiceprint recognition methods in the prior art having low robustness.

(57) 摘要: 一种声纹识别方法及装置, 该方法包括: 采集用户录入的注册语音 (S201); 基于注册语音生成远场条件下的样本语音 (S203); 基于样本语音对声纹识别模型进行训练 (S204); 采集用户录入的验证语音 (S205); 将验证语音输入声纹识别模型进行匹配, 得到匹配结果 (S207); 基于匹配结果确定用户是否为声纹识别模型的注册人 (S208)。该方法涉及人工智能等相关领域, 用以解决现有技术中声纹识别方法鲁棒性低的问题。

一种声纹识别方法及装置

相关申请的交叉引用

本申请要求在2019年07月24日提交中国专利局、申请号为201910673696.7、申请名称为“一种声纹识别方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及计算机技术领域，尤其涉及一种声纹识别方法及装置。

背景技术

声纹识别，是一种通过语音信号来自动辨识和确认说话人身份的技术。声纹识别的基本方案包括声纹注册和声纹验证两个阶段。在声纹注册阶段，将注册人的注册语音信息转换成验证模型；在声纹验证阶段，将验证语音的信息与声纹注册阶段生成的验证模型进行相似度打分，判断该验证语音是否来自于该注册人。

远场声纹识别，相比于近场声纹识别，具有更大的挑战性。其中主要的原因是语音信号在远场条件下的失真，体现为环境噪音的叠加以及房间混响。说话人在房间，或者密闭空间里讲话时，声波在空气中传播，并在墙壁和障碍物上反射；由于材料的吸收，声波中高频衰减，然后再次扩散到房间，从而导致混响。因此，在远场条件中，注册语音与验证语音不相匹配，声纹识别准确性较低。

针对这个问题，一种解决方案为用户在近、远场分别进行声纹注册。具体的，为了使远场条件下的验证语音能够与注册语音相匹配，该方案提出用户在近场条件和远场条件分别进行声纹注册。但是，这种方案需要用户在近场条件、远场条件进行多次的声纹注册，降低了用户体验。

另一种解决方案是前端语音信号增强。具体的，在声纹注册阶段采集近场干净语音作为注册语音，在声纹验证阶段将采集到的远场语音数据通过前端处理，得到增强后语音，再将增强后的语音作为验证语音输入。但是，这种方案中增强后的语音的高频部分较近场干净语音仍有损失，因此增强后语音与注册语音依然不匹配，从而导致声纹识别系统鲁棒性低，识别率提升不明显等问题。

发明内容

本申请提供了一种声纹识别方法及装置，用以解决现有技术中声纹识别方法鲁棒性低的问题。

第一方面，本申请实施例提供的声纹注册方法，包括：电子设备提示用户录入注册语音；电子设备采集用户录入的注册语音；电子设备基于注册语音生成远场条件下的样本语音；电子设备基于样本语音对声纹识别模型进行训练。本申请实施例中电子设备可以根据注册语音仿真生成远场条件下的样本语音，而不需要用户在近场条件、远场条件进行多次的声纹注册，从而可以提高用户体验。并且，电子设备基于远场条件下样本语音对声纹识

别模型进行训练，可以提高声纹识别模型的鲁棒性，从而可以提高声纹识别的准确率。

在一种可能的设计中，电子设备基于注册语音生成远场条件下的样本语音时，可以模拟在远场条件下对声音的混响；并基于远场条件下对声音的混响仿真生成注册语音在远场条件下的样本数据。上述设计中通过模拟声音在远场条件下的混响，从而可以仿真得到注册语音在远场条件下的样本语音。

在一种可能的设计中，电子设备基于注册语音生成远场条件下的样本语音时，可以基于注册语音以及噪声数据生成噪声语音；电子设备模拟远场条件下对声音的混响；电子设备基于远场条件下对声音的混响仿真生成噪声语音在远场条件下的样本数据。上述设计中，通过在仿真注册语音在远场条件下的样本语音时结合噪声语音，从而可以使得样本语音更符合实际场景，从而可以提高声纹识别模型的鲁棒性，进而提高声纹识别的准确性。

在一种可能的设计中，电子设备模拟远场条件下对声音的混响时，可以基于远场条件仿真声音的墙壁反射，得到房间冲击响应（room impulse response, RIR）。上述设计中，通过仿真声音的墙壁反射得到 RIR，可以模拟远场条件下对声音的混响。

在一种可能的设计中，电子设备基于样本语音对声纹识别模型进行训练时，可以对样本语音进行特征提取，得到特征数据；并基于特征数据对声纹识别模型进行训练。上述设计中，通过提取样本语音的特征数据，从而可以提高声纹识别模型的鲁棒性。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景。电子设备基于特征数据对声纹识别模型进行训练时，可以基于特征数据分别对一个或多个子模型进行训练。上述设计中，通过针对不同场景分别训练对应的声纹识别子模型，从而可以解决注册语音场景单一而验证语音场景多变导致数据不匹配问题。

在一种可能的设计中，声纹识别模型可以包括一个融合模型，其中，融合模型对应一个或多个场景。电子设备基于特征数据对声纹识别模型进行训练时，基于特征数据对融合模型进行训练。上述设计中，电子设备通过维护一个融合模型，从而可以节省电子设备的计算资源。

第二方面，本申请实施例提供的声纹识别方法，包括：电子设备提示用户录入验证语音；电子设备采集用户录入的验证语音；电子设备将验证语音输入声纹识别模型进行匹配，得到匹配结果；电子设备基于匹配结果确定用户是否为声纹识别模型的注册人。其中，声纹识别模型可以采用上述第一方面所述的方法训练得到，具体的，电子设备在训练声纹识别模型的过程可以包括：电子设备提示用户录入注册语音；电子设备采集用户录入的注册语音；电子设备基于注册语音生成远场条件下的样本语音；电子设备基于样本语音对声纹识别模型进行训练。本申请实施例中，电子设备通过采用第一方面训练好的声纹识别模型，可以准确的识别验证语音是否来自注册人。

在一种可能的设计中，电子设备在采集用户录入的验证语音之后，可以对验证语音进行场景检测。上述设计中，通过检测验证语音所在的场景，使得电子设备可以结合验证语音所在的场景对验证语音进行声纹识别，从而可以提高声纹识别的准确性。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景；电子设备将验证语音输入声纹识别模型进行匹配，包括：电子设备将验证语音输入第一场景对应的子模型进行匹配，其中，第一场景为场景检测的结果。上述设计中，电子设备通过结合验证语音所在的场景对验证语音进行声纹识别，可以提高声纹识别的准确性。

在一种可能的设计中，若用户是声纹识别模型的注册人，电子设备可以对验证语音进行质量评估，得到质量评估结果。若质量评估结果表示验证语音为高质量语音，电子设备可以基于验证语音对声纹识别模型进行增量学习。上述设计中，通过将高质量的验证语音加入到高质量语音样本库，并进行数据增强和增量学习，更新声纹识别模型库中的模型，从而用户使用过程中，声纹识别模型库中的模型可以越来越适用于用户实际使用场景。

在一种可能的设计中，电子设备基于验证语音对声纹识别模型进行增量学习时，可以对验证语音进行数据增强处理，得到处理后的语音数据。基于处理的语音数据对声纹识别模型进行增量学习。上述设计中，通过基于验证语音进一步进行增量学习，更新声纹识别模型库中的模型，从而用户使用过程中，声纹识别模型库中的模型的鲁棒性越来越高，声纹识别模型库中的模型可以越来越适用于用户实际使用场景。

在一种可能的设计中，在电子设备对验证语音进行数据增强处理之前，电子设备可以确定验证语音所在的第一场景为高频场景。电子设备对验证语音进行数据增强处理时，可以将验证语音进行数据增强处理，得到 j 个不同噪声等级的样本语音。电子设备基于处理的语音数据对声纹识别模型进行增量学习时，可以将 j 个样本语音按照噪声等级进行分组，得到 M 组语音数据， M 为大于 0 且不大于 j 的整数；并基于 M 组语音数据分别对第一场景对应的子模型进行训练，得到 M 个高频子模型。

第三方面，本申请实施例提供的声纹注册装置，包括：第一器件，麦克风以及处理器，其中，第一器件为扬声器或者显示屏。处理器，用于执行：触发第一器件提示用户录入注册语音；通过麦克风采集用户录入的注册语音；基于注册语音生成远场条件下的样本语音；基于样本语音对声纹识别模型进行训练。

在一种可能的设计中，处理器触发第一器件提示用户录入注册语音时，可以触发扬声器播放提示语音，其中，提示语音用于提示用户录入注册语音。或者，处理器也可以触发显示屏显示提示文字，其中，提示文字用于提示用户录入注册语音。

在一种可能的设计中，处理器，在基于注册语音生成远场条件下的样本语音时，可以具体用于：模拟注册语音在远场条件下对声音的混响；基于远场条件下对声音的混响仿真生成注册语音在远场条件下的样本数据。

在一种可能的设计中，处理器，在基于注册语音生成远场条件下的样本语音时，可以具体用于：基于注册语音以及噪声数据生成噪声语音；模拟远场条件下对声音的混响；基于远场条件下对声音的混响仿真生成噪声语音在远场条件下的样本数据。

在一种可能的设计中，处理器，在模拟远场条件下对声音的混响时，可以具体用于：基于远场条件仿真声音的墙壁反射，得到房间冲击响应 RIR。

在一种可能的设计中，处理器，在基于样本语音对声纹识别模型进行训练时，可以具体用于：对样本语音进行特征提取，得到特征数据；基于特征数据对声纹识别模型进行训练。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景。处理器，在基于特征数据对声纹识别模型进行训练时，可以具体用于：基于特征数据分别对一个或多个子模型进行训练。

在一种可能的设计中，声纹识别模型可以包括一个融合模型，其中，融合模型对应一个或多个场景。处理器，在基于特征数据对声纹识别模型进行训练时，可以具体用于：基于特征数据对融合模型进行训练。

第四方面，本申请实施例提供的声纹识别装置，装置包括：第一器件、麦克风以及处理器，其中，第一器件为扬声器或者显示屏。处理器，用于执行：触发第一器件提示用户录入验证语音；通过麦克风采集用户录入的验证语音；将验证语音输入声纹识别模型进行匹配，得到匹配结果；基于匹配结果确定用户是否为声纹识别模型的注册人。其中，声纹识别模型为通过上述第三方面的声纹注册装置的训练得到。

其中，上述声纹注册装置与声纹识别可以是一个装置，这种情况下，在训练声纹识别模型时，声纹识别装置的处理器，还可以用于执行：触发第一器件提示用户录入注册语音；通过麦克风采集用户录入的注册语音；基于注册语音生成远场条件下的样本语音；基于样本语音对声纹识别模型进行训练。

或者，上述声纹注册装置与声纹识别也可以是两个不同装置，这种情况下，声纹注册装置可以包括：第一器件，麦克风以及处理器，其中，第一器件为扬声器或者显示屏。处理器，用于执行：触发第一器件提示用户录入注册语音；通过麦克风采集用户录入的注册语音；基于注册语音生成远场条件下的样本语音；基于样本语音对声纹识别模型进行训练。

在一种可能的设计中，处理器触发第一器件提示用户录入注册语音时，可以触发扬声器播放提示语音，其中，提示语音用于提示用户录入验证语音。或者，处理器也可以触发显示屏显示提示文字，其中，提示文字用于提示用户录入验证语音。

在一种可能的设计中，处理器，还可以用于：在通过麦克风采集用户录入的验证语音之后，对验证语音进行场景检测。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景。处理器，在将验证语音输入声纹识别模型进行匹配时，可以具体用于：将验证语音输入第一场景对应的子模型进行匹配，其中，第一场景为场景检测的结果。

在一种可能的设计中，处理器，还可以用于：若用户是声纹识别模型的注册人，对验证语音进行质量评估，得到质量评估结果；若质量评估结果表示验证语音为高质量语音，基于验证语音对声纹识别模型进行增量学习。

在一种可能的设计中，处理器，在基于验证语音对声纹识别模型进行增量学习时，可以具体用于：对验证语音进行数据增强处理，得到处理后的语音数据；基于处理的语音数据对声纹识别模型进行增量学习。

在一种可能的设计中，处理器，还可以用于：在对验证语音进行数据增强处理之前，确定验证语音所在的第一场景为高频场景。处理器，在对验证语音进行数据增强处理时，可以具体用于：将验证语音进行数据增强处理，得到 j 个不同噪声等级的样本语音。处理器，在基于处理的语音数据对声纹识别模型进行增量学习时，可以具体用于：将 j 个样本语音按照噪声等级进行分组，得到 M 组语音数据， M 为大于 0 且不大于 j 的整数；基于 M 组语音数据分别对第一场景对应的子模型进行训练，得到 M 个高频子模型。

第五方面，本申请实施例提供的芯片，该芯片包括处理器和通信接口，所述通信接口用于接收代码指令，并传输到处理器。处理器，用于调用通信接口传输的代码指令以执行：触发电子设备的扬声器或者显示屏提示用户录入注册语音；触发电子设备的麦克风采集用户录入的注册语音；基于注册语音生成远场条件下的样本语音；基于样本语音对声纹识别模型进行训练。

在一种可能的设计中，处理器触发扬声器提示用户录入注册语音时，可以触发扬声器播放提示语音，其中，提示语音用于提示用户录入注册语音。

在一种可能的设计中，处理器触发显示屏提示用户录入注册语音时，可以触发显示屏显示提示文字，其中，提示文字用于提示用户录入注册语音。

在一种可能的设计中，处理器，在基于注册语音生成远场条件下的样本语音时，可以具体用于：模拟注册语音在远场条件下对声音的混响；基于远场条件下对声音的混响仿真生成注册语音在远场条件下的样本数据。

在一种可能的设计中，处理器，在基于注册语音生成远场条件下的样本语音时，可以具体用于：基于注册语音以及噪声数据生成噪声语音；模拟远场条件下对声音的混响；基于远场条件下对声音的混响仿真生成噪声语音在远场条件下的样本数据。

在一种可能的设计中，处理器，在模拟远场条件下对声音的混响时，可以具体用于：基于远场条件仿真声音的墙壁反射，得到房间冲击响应 RIR。

在一种可能的设计中，处理器，在基于样本语音对声纹识别模型进行训练时，可以具体用于：对样本语音进行特征提取，得到特征数据；基于特征数据对声纹识别模型进行训练。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景。处理器，在基于特征数据对声纹识别模型进行训练时，可以具体用于：基于特征数据分别对一个或多个子模型进行训练。

在一种可能的设计中，声纹识别模型可以包括一个融合模型，其中，融合模型对应一个或多个场景。处理器，在基于特征数据对声纹识别模型进行训练时，可以具体用于：基于特征数据对融合模型进行训练。

第六方面，本申请实施例提供的芯片，该芯片包括处理器和通信接口，所述通信接口用于接收代码指令，并传输到处理器。处理器，用于调用通信接口传输的代码指令以执行：触发电子设备的扬声器或者电子设备的显示屏提示用户录入验证语音；通过麦克风采集用户录入的验证语音；将验证语音输入声纹识别模型进行匹配，得到匹配结果，其中，声纹识别模型为权利要求 13~18 任一项的装置训练得到；基于匹配结果确定用户是否为声纹识别模型的注册人。

在一种可能的设计中，处理器触发电子设备的扬声器提示用户录入注册语音时，可以触发扬声器播放提示语音，其中，提示语音用于提示用户录入验证语音。

在一种可能的设计中，处理器触发电子设备的显示屏提示用户录入注册语音时，可以触发显示屏显示提示文字，其中，提示文字用于提示用户录入验证语音。

在一种可能的设计中，处理器，还可以调用通信接口传输的代码指令执行：在触发电子设备的麦克风采集用户录入的验证语音之后，对验证语音进行场景检测。

在一种可能的设计中，声纹识别模型可以包括一个或多个子模型，其中，一个子模型对应一个场景。处理器，在将验证语音输入声纹识别模型进行匹配时，可以具体用于：将验证语音输入第一场景对应的子模型进行匹配，其中，第一场景为场景检测的结果。

在一种可能的设计中，处理器，还可以用于：若用户是声纹识别模型的注册人，对验证语音进行质量评估，得到质量评估结果；若质量评估结果表示验证语音为高质量语音，基于验证语音对声纹识别模型进行增量学习。

在一种可能的设计中，处理器，在基于验证语音对声纹识别模型进行增量学习时，可以具体用于：对验证语音进行数据增强处理，得到处理后的语音数据；基于处理的语音数据对声纹识别模型进行增量学习。

在一种可能的设计中，处理器，还可以用于：在对验证语音进行数据增强处理之前，确定验证语音所在的第一场景为高频场景。处理器，在对验证语音进行数据增强处理时，可以具体用于：将验证语音进行数据增强处理，得到 j 个不同噪声等级的样本语音。处理器，在基于处理的语音数据对声纹识别模型进行增量学习时，可以具体用于：将 j 个样本语音按照噪声等级进行分组，得到 M 组语音数据， M 为大于 0 且不大于 j 的整数；基于 M 组语音数据分别对第一场景对应的子模型进行训练，得到 M 个高频子模型。

第七方面，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中包括指令，当其在计算机上运行时，使得计算机执行上述各方面所述的方法。

第八方面，本申请还提供一种包括指令的计算机程序产品，当其被运行时，使得上述各方面所述的方法被执行。

附图说明

图 1 为本申请提供的一种电子设备的硬件结构示意图；

图 2 为本申请提供的一种声纹识别方法的流程示意图；

图 3 为本申请提供的一种显示屏提示用户录入注册语音的示意图；

图 4 为本申请提供的一种用户触发电子设备进行声纹验证的示意图；

图 5 为本申请提供的一种显示屏输出识别结果的示意图；

图 6 为本申请提供的一种声纹识别过程的示意图；

图 7 为本申请提供的一种显示屏输出识别结果的示意图；

图 8 为本申请提供的另一种声纹识别过程的示意图。

具体实施方式

应理解，在本申请中除非另有说明，“/”表示或的意思，例如， A/B 可以表示 A 或 B ；本申请中的“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如， A 和/或 B ，可以表示：单独存在 A ，同时存在 A 和 B ，单独存在 B 这三种情况。“至少一个”是指一个或者多个，“多个”是指两个或两个以上。

在本申请中，“示例的”、“在一些实施例中”、“在另一些实施例中”等用于表示作例子、例证或说明。本申请中被描述为“示例”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用示例的一词旨在以具体方式呈现概念。

另外，本申请中涉及的“第一”、“第二”等词汇，仅用于区分描述的目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量，也不能理解为指示或暗示顺序。

本申请实施例中的电子设备为具有声纹识别功能的电子设备。声纹识别，是一种通过语音信号来自动辨识和确认说话人身份的技术。本申请实施例中的电子设备可以采集用户的语音数据，并对语音数据进行声纹识别以判断该用户是否为注册人。

以下介绍电子设备、用于这样的电子设备的图形用户界面（graphical user interface, GUI）、和用于使用这样的电子设备的实施例。为描述方便，以下将 GUI 简称为用户界面。

本申请实施例中的电子设备可以为便携式电子设备，诸如手机、平板电脑、人工智能（artificial intelligence, AI）智能语音终端、可穿戴设备、增强现实（augmented reality,

AR)/虚拟现实(virtual reality, VR)设备等。便携式电子设备的示例性实施例包括但不限于搭载IOS®、Android®、Microsoft®或者其它操作系统的便携式电子设备。上述便携式电子设备也可以是车载终端、膝上型计算机(Laptop)等。还应当理解的是,本申请实施例的电子设备还可以台式计算机、智能家居设备(例如智能电视、智能音箱)等,对此不作限定。

示例的,如图1所示,为本申请实施例的一种电子设备的硬件结构示意图。具体的如图1所示,电子设备包括处理器110、内部存储器121、外部存储器接口122、摄像头131、显示屏132、传感器模块140、用户标识模块(subscriber identification module, SIM)卡接口151、按键152、音频模块160、扬声器161、受话器162、麦克风163、耳机接口164、通用串行总线(universal serial bus, USB)接口170、充电管理模块180、电源管理模块181、电池182、移动通信模块191和无线通信模块192。在另一些实施例中,电子设备还可以包括马达、指示器、按键等。

应理解,图1所示的硬件结构仅是一个示例。本申请实施例的电子设备可以具有比图1所示电子设备更多的或者更少的部件,可以组合两个或更多的部件,或者可以具有不同的部件配置。图1所示出的各种部件可以在包括一个或多个信号处理和/或专用集成电路在内的硬件、软件、或硬件和软件的组合中实现。

其中,处理器110可以包括一个或多个处理单元。例如:处理器110可以包括应用处理器(application processor, AP)、调制解调器、图形处理器(graphics processing unit, GPU)、图像信号处理器(image signal processor, ISP)、控制器、视频编解码器、数字信号处理器(digital signal processor, DSP)、基带处理器、和/或神经网络处理器(neural-network processing unit, NPU)等。其中,不同的处理单元可以是独立的器件,也可以集成在一个或多个处理器中。

在一些实施例中,处理器110中还可以设置缓存器,用于存储指令和/或数据。示例的,处理器110中的缓存器可以为高速缓冲存储器。该缓存器可以用于保存处理器110刚用过的、生成的、或循环使用的指令和/或数据。如果处理器110需要使用该指令或数据,可从该缓存器中直接调用。有助于减少了处理器110获取指令或数据的时间,从而有助于提高系统的效率。

内部存储器121可以用于存储程序和/或数据。在一些实施例中,内部存储器121包括存储程序区和存储数据区。其中,存储程序区可以用于存储操作系统(如Android、IOS等操作系统)、至少一个功能所需的计算机程序(比如声纹识别功能、声音播放功能)等。存储数据区可以用于存储电子设备使用过程中所创建、和/或采集的数据(比如音频数据)等。示例的,处理器110可以通过调用内部存储器121中存储的程序和/或数据,使得电子设备执行相应的方法,从而实现一种或多种功能。例如,处理器110调用内部存储器中的某些程序和/或数据,使得电子设备执行本申请实施例中所提供的声纹识别方法、从而实现声纹识别功能。其中,内部存储器121可以采用高速随机存取存储器、和/或非易失性存储器等。例如,非易失性存储器可以包括一个或多个磁盘存储器件、闪存器件、和/或通用闪存存储器(universal flash storage, UFS)等中的至少一个。

外部存储器接口122可以用于连接外部存储卡(例如, Micro SD卡),实现扩展电子设备的存储能力。外部存储卡通过外部存储器接口122与处理器110通信,实现数据存储功能。例如电子设备可以通过外部存储器接口122将图像、音乐、视频等文件保存在外部

存储卡中。

摄像头 131 可以用于捕获动、静态图像等。通常情况下，摄像头 131 包括镜头和图像传感器。其中，物体通过镜头生成的光学图像投射到图像传感器上，然后转换为电信号，在进行后续处理。示例的，图像传感器可以是电荷耦合器件 (charge coupled device, CCD) 或互补金属氧化物半导体 (complementary metal-oxide-semiconductor, CMOS) 光电晶体管。图像传感器把光信号转换成电信号，之后将电信号传递给 ISP 转换成数字图像信号。需要说明的是，电子设备可以包括 1 个或 N 个摄像头 131，其中，N 为大于 1 的正整数。

显示屏 132 可以包括显示面板，用于显示用户界面。显示面板可以采用液晶显示屏 (liquid crystal display, LCD)、有机发光二极管 (organic light-emitting diode, OLED)、有源矩阵有机发光二极体或主动矩阵有机发光二极体 (active-matrix organic light emitting diode, AMOLED)、柔性发光二极管 (flex light-emitting diode, FLED)、Miniled、MicroLed、Micro-oLed、量子点发光二极管 (quantum dot light emitting diodes, QLED) 等。需要说明的是，电子设备可以包括 1 个或 M 个显示屏 132，M 为大于 1 的正整数。示例的，电子设备可以通过 GPU、显示屏 132、应用处理器等实现显示功能。

传感器模块 140 可以包括一个或多个传感器。例如，触摸传感器 140A、陀螺仪 140B、加速度传感器 140C、指纹传感器 140D、压力传感器 140E 等。在一些实施例中，传感器模块 140 还可以包括环境光传感器、距离传感器、接近光传感器、骨传导传感器、温度传感器等。

其中，触摸传感器 140A，也可称为“触控面板”。触摸传感器 140A 可以设置于显示屏 132，由触摸传感器 140A 与显示屏 132 组成触摸屏，也称“触控屏”。触摸传感器 140A 用于检测作用于其上或附近的触摸操作。触摸传感器 140A 可以将检测到的触摸操作传递给应用处理器，以确定触摸事件类型。电子设备可以通过显示屏 132 提供与触摸操作相关的视觉输出等。在另一些实施例中，触摸传感器 140A 也可以设置于电子设备的表面，与显示屏 132 所处的位置不同。

陀螺仪 140B 可以用于确定电子设备的运动姿态。在一些实施例中，可以通过陀螺仪 140B 确定电子设备围绕三个轴 (即，x、y 和 z 轴) 的角速度。陀螺仪 140B 可以用于拍摄防抖。示例性的，当按下快门，陀螺仪 140B 检测电子设备抖动的角度，根据角度计算出镜头模组需要补偿的距离，让镜头通过反向运动抵消电子设备的抖动，从而实现防抖。陀螺仪传感器 140B 还可以用于导航、体感游戏场景。

加速度传感器 140C 可检测电子设备在各个方向上 (一般为三轴) 加速度的大小。当电子设备静止时可检测出重力的大小及方向。加速度传感器 140C 还可以用于识别电子设备的姿态，应用于横竖屏切换、计步器等应用。

指纹传感器 140D 用于采集指纹。电子设备可以利用采集的指纹特性实现指纹解锁、访问应用锁、指纹拍照、指纹接听来电等。

压力传感器 140E 用于感受压力信号，可以将压力信号转换成电信号。示例的，压力传感器 140E 可以设置于显示屏 132。其中，作用于相同触摸位置，但不同触摸操作强度的触摸操作，可以对应不同的操作指令。

SIM 卡接口 151 用于连接 SIM 卡。SIM 卡可以通过插入 SIM 卡接口 151，或从 SIM 卡接口 151 拔出，实现和电子设备的接触和分离。电子设备可以支持 1 个或 K 个 SIM 卡接口 151，K 为大于 1 的正整数。SIM 卡接口 151 可以支持 Nano SIM 卡、Micro SIM 卡、

和/或 SIM 卡等。同一个 SIM 卡接口 151 可以同时插入多张卡。所述多张卡的类型可以相同，也可以不同。SIM 卡接口 151 也可以兼容不同类型的 SIM 卡。SIM 卡接口 151 也可以兼容外部存储卡。电子设备通过 SIM 卡和网络交互，实现通话以及数据通信等功能。在一些实施例中，电子设备还可以采用 eSIM，即：嵌入式 SIM 卡。eSIM 卡可以嵌在电子设备中，不能和电子设备分离。

按键 152 可以包括开机键、音量键等。按键 152 可以是机械按键，也可以是触摸式按键。电子设备可以接收按键输入，产生与电子设备的用户设置以及功能控制有关的键信号输入。

电子设备可以通过音频模块 160、扬声器 161、受话器 162、麦克风 163、耳机接口 164 以及应用处理器等实现音频功能。例如，音频播放功能、录音功能、声纹注册功能、声纹验证功能、声纹识别功能等。

音频模块 160 可以用于对音频数据进行数模转换、和/或模数转换，还可以用于对音频数据进行编码和/或解码。示例的，音频模块 160 可以独立于处理器设置，也可以设置于处理器 110 中，或将音频模块 160 的部分功能模块设置于处理器 110 中。

扬声器 161，也称“喇叭”，用于将音频数据转换为声音，并播放声音。例如，电子设备 100 可以通过扬声器 161 收听音乐、接听免提电话、或者发出语音提示等。

受话器 162，也称“听筒”，用于将音频数据转换成声音，并播放声音。例如，当电子设备 100 接听电话时，可以通过将受话器 162 靠近人耳进行接听。

麦克风 163，也称“话筒”、“传声器”，用于采集声音（例如周围环境声音，包括人发出的声音、设备发出的声音等），并将声音转换为音频电数据。当拨打电话或发送语音时，用户可以通过人嘴靠近麦克风 163 发出声音，麦克风 163 采集用户发出的声音。当电子设备的声纹识别功能已开启的情况下，麦克风 163 可以实时采集周围环境声音，获取音频数据。其中，麦克风 163 采集声音的情况与所处的环境相关。例如，当周围环境较为嘈杂时，用户说出验证话语时，则麦克风 163 采集的声音包括周围环境噪声和用户发出验证话语的声音。再例如，当周围环境较为安静时，用户说出验证话语，则麦克风 163 采集的声音为用户发出验证话语的声音。再例如，当周围环境为远场条件时，用户说出验证话语，则麦克风 163 采集的声音为周围环境噪声的叠加以及混响用户发出验证话语的混响。又例如，当周围环境较为嘈杂时，电子设备的声纹识别功能已开启，但是用户并未说出验证话语，则麦克风 163 采集的声音仅为周围环境噪声。

需要说明的是，电子设备可以设置至少一个麦克风 163。例如，电子设备中设置两个麦克风 163，除了采集声音，还可以实现降噪功能。又示例如，电子设备中还可以设置三个、四个或更多个麦克风 163，从而可以在实现声音采集、降噪的基础上，还可以实现声音来源的识别、或定向录音功能等。

耳机接口 164 用于连接有线耳机。耳机接口 164 可以是 USB 接口 170，也可以是 3.5mm 的开放移动电子设备平台（open mobile terminal platform, OMTF）标准接口、美国蜂窝电信工业协会（cellular telecommunications industry association of the USA, CTIA）标准接口等。

USB 接口 170 是符合 USB 标准规范的接口，具体可以是 Mini USB 接口、Micro USB 接口、USB Type C 接口等。USB 接口 170 可以用于连接充电器为电子设备充电，也可以用于电子设备与外围设备之间传输数据。也可以用于连接耳机，通过耳机播放音频。示例

的, USB 接口 170 除了可以为耳机接口 164 以外, 还可以用于连接其他电子设备, 例如 AR 设备、计算机等。

充电管理模块 180 用于从充电器接收充电输入。其中, 充电器可以是无线充电器, 也可以是有线充电器。在一些有线充电的实施例中, 充电管理模块 180 可以通过 USB 接口 170 接收有线充电器的充电输入。在一些无线充电的实施例中, 充电管理模块 180 可以通过电子设备的无线充电线圈接收无线充电输入。充电管理模块 180 为电池 182 充电的同时, 还可以通过电源管理模块 180 为电子设备供电。

电源管理模块 181 用于连接电池 182、充电管理模块 180 与处理器 110。电源管理模块 181 接收电池 182 和/或充电管理模块 180 的输入, 为处理器 110、内部存储器 121、显示屏 132、摄像头 131 等供电。电源管理模块 181 还可以用于监测电池容量、电池循环次数、电池健康状态 (漏电、阻抗) 等参数。在其他一些实施例中, 电源管理模块 181 也可以设置于处理器 110 中。在另一些实施例中, 电源管理模块 181 和充电管理模块 180 也可以设置于同一个器件中。

移动通信模块 191 可以提供应用在电子设备上的包括 2G/3G/4G/5G 等无线通信的解决方案。移动通信模块 191 可以包括滤波器、开关、功率放大器、低噪声放大器 (low noise amplifier, LNA) 等。

无线通信模块 192 可以提供应用在电子设备上的包括 WLAN (如 Wi-Fi 网络)、蓝牙 (Bluetooth, BT)、全球导航卫星系统 (global navigation satellite system, GNSS)、调频 (frequency modulation, FM)、近距离无线通信技术 (near field communication, NFC)、红外技术 (infrared, IR) 等无线通信的解决方案。无线通信模块 192 可以是集成至少一个通信处理模块的一个或多个器件。

在一些实施例中, 电子设备的天线 1 和移动通信模块 191 耦合, 天线 2 和无线通信模块 192 耦合, 使得电子设备可以与其他设备通信。具体的, 移动通信模块 191 可以通过天线 1 与其它设备通信, 无线通信模块 193 可以通过天线 2 与其它设备通信。

本申请实施例以下将结合附图和应用场景, 对本申请实施例提供的声纹识别方法进行详细介绍。以下实施例均可以在具有上述硬件结构的电子设备 100 中实现。

为了很好的理解本申请实施例提供的声纹识别方法, 下面对本申请实施例中涉及的词语进行解释说明。

近场条件: 声源与麦克风 (mic) 之间的距离较近, 例如声源距离 mic 在 1 米之内。

近场语音: 可以理解为在近场条件下采集的语音数据, 例如声源和 mic 之间的距离为 1 米以下时 mic 针对该声源采集的语音数据为近场语音。近场语音可以包括近场干净语音以及近场带噪语音, 其中, 近场干净语音可以理解为在近场条件下采集到的没有噪声的语音数据, 近场带噪语音可以理解为在近场条件下采集到的带噪声的语音数据。

远场条件: 声源与麦克风 (mic) 之间的距离较远, 例如声源距离 mic 在 1 米~10 米之内等等。

远场语音: 可以理解为在远场条件下采集的语音数据, 例如声源和 mic 之间的距离为 5 米以下时 mic 针对该声源采集的语音数据为远场语音。远场语音可以包括远场干净语音以及远场带噪语音, 其中, 远场干净语音可以理解为在远场条件下采集到的没有噪声的语音数据, 远场带噪语音可以理解为在远场条件下采集到的带噪声的语音数据。

声纹识别模型: 可以是电子设备基于高斯混合模型 (gaussian mixture model, GMM)

-背景模型(universal background model, UBM)、支持向量机(support vector machine, SVM)、联合因子分析(joint factor analysis, JFA)、身份向量(identity vector, I-vector)、X-vector 等方法建立的数据模型, 电子设备在建立初始的声纹识别模型后采用样本数据对该初始的声纹识别模型进行训练, 训练好的声纹识别模型可以用于进行声纹识别。

5 多场景融合模型: 可以理解为采用多个场景的样本数据对初始的声纹识别模型进行训练, 该声纹识别模型训练好后即可以看做是一个多场景融合模型。

单场景模型: 采用一个场景的样本数据对初始的声纹识别模型进行训练, 即可得到一个单场景模型, 本申请实施例中, 每个场景对应的模型分别为一个单场景模型。具体来说, 采用一个场景的样本数据对初始的声纹识别模型进行训练, 该声纹识别模型训练好后即可以看做是该场景对应的模型, 例如, 采用居家场景的样本数据对初始的声纹识别模型进行训练, 得到居家场景对应的模型(或者可以简称为居家模型), 采用车载场景的样本数据对初始的声纹识别模型进行训练, 得到车载场景对应的模型(或者也可以称为车载模型)。因此, 通过采用不同场景的样本数据分别对初始的声纹识别模型进行训练, 即可得到每个场景对应的单场景模型。

15 增量学习: 每当新增样本数据时, 并不需要重建声纹识别模型, 而是在原有声纹识别模型的基础上, 对由于新增样本数据所引起的变化进行更新, 也就是在之前训练好的声纹识别模型的基础上采用新增的样本数据进一步训练, 从而不断更新声纹识别模型。

特征提取: 对数据进行变换, 以突出该数据具有代表性特征的一种方法, 本申请实施例中可以指对语音数据进行变换提取语音数据中属于特征性的信息的方法及过程。

20 场景检测: 通过提取语音数据的背景数据, 从而判断该语音数据所处的场景。

参见图 2, 示例性的示出了本申请实施例提供的一种声纹识别方法的流程, 该方法由电子设备执行。声纹识别的基本方案包括声纹注册和声纹验证两个阶段。其中, 声纹注册可以通过步骤 S201~S204 实现。声纹验证可以通过步骤 S205~S209 实现。

25 S201, 电子设备采集用户录入的注册语音。其中, 用户录入的注册语音可以是近场干净语音。

具体来说, 电子设备可以通过麦克风 163 采集周围环境声音, 获取用户录入的注册语音。

具体实施中, 用户可以在电子设备的提示下说出注册语音, 例如, 如图 3 所示, 电子设备可以在显示屏 132 上显示文字提示用户说出注册语“1234567”。又例如, 电子设备也可以通过扬声器 161 进行语音提示, 等等。其中, 可以是用户首次启动电子设备的声纹识别功能时电子设备自动提示用户说出注册语音, 或者, 也可以是用户首次启动电子设备的声纹识别功能时由用户操作电子设备提示用户说出注册语音, 或者, 也可以是用户在后续启动声纹识别功能时用户根据需求触发电子设备提示用户说出注册语音。

35 作为一种可能的实施方式, 用户在进行声纹注册时可以多次输入注册语音, 从而提高声纹识别的准确性。

S202, 电子设备在采集到注册语音后, 可以将注册语音存储到高质量语音样本库, 其中, 高质量语音样本库用于存储语音质量得分大于或等于质量阈值的语音。

40 S203, 电子设备将高质量语音样本库中包括的注册语音进行数据增强处理, 得到多个样本语音。该样本语音可以但不限于为: 由注册语音生成的带噪语音, 由注册语音生成的远场语音, 由注册语音生成的远场带噪语音等。

上述方式中, 电子设备可以基于注册语音生成带噪语音、远场语音等, 而不需要用户分别在近场、远场等场景分别进行注册, 从而可以提升用户体验。

电子设备基于注册语音生成带噪语音时, 可以通过如下方式实现: 在仿真房间中加入注册语音和噪声源, 将注册语音和噪声源进行处理得到带噪语音, 其中, 噪声源可以有一个或多个。具体的, 电子设备基于注册语音可以生成不同噪声等级的带噪语音。例如, 不同场景可能对应不同的噪声等级, 因此电子设备可以针对每个场景将注册语音仿真生成该场景对应的带噪语音。

电子设备基于注册语音生成远场语音时, 可以通过如下方式实现: 使用镜像源模型(image source model, ISM)算法, ISM 算法以虚拟声音源仿真声音的墙壁反射, 根据信号延迟与衰减参数计算房间冲击响应(room impulse response, RIR), 其中, 语音到墙壁后会有一些有损耗的反射, 做 RIR 是为了仿真模拟远场条件下对声音的混响。根据 RIR 仿真生成注册语音对应的远场语音。具体的, 电子设备基于注册语音可以生成不同远场等级的远场语音。例如, 不同场景可能对应不同的远场距离, 因此电子设备可以针对每个场景将注册语音仿真生成该场景对应的远场语音。

此外, 电子设备还可以采用其他方式仿真远场条件下的混响, 例如, 对声音进行脉冲响应卷积来模拟远场条件下对声音的混响, 等。

电子设备基于注册语音生成远场带噪语音时, 可以通过如下方式实现: 在仿真房间中加入注册语音和噪声源, 将注册语音和噪声源进行处理得到带噪语音, 其中, 噪声源可以有一个或多个; 使用 ISM 算法计算 RIR, 根据 RIR 仿真生成带噪语音对应的远场带噪语音。具体的, 电子设备基于注册语音可以生成不同远场等级、不同噪声等级的远场带噪语音。例如, 电子设备可以针对具体场景的噪声特点、远场特点等将注册语音仿真生成该场景对应的远场带噪语音。

上述过程中, 噪声等级可以理解为噪声强度等级, 远场等级可以理解为远场距离等级。

S204, 电子设备对样本语音进行特征提取, 并基于提取的特征训练声纹识别模型库中的模型, 得到训练好的模型。

示例性的, 声纹识别模型库中的模型可以但不限于采用 GMM-UBM、SVM、JFA、I-vector、X-vector 等方法建立得到的。

在具体实施中, 声纹识别模型库可以包括多场景融合模型, 因此, 电子设备可以使用多个场景的样本语音对多场景融合模型进行训练。或者, 声纹识别模型库中也可以包括多个场景分别对应的模型, 因此, 针对每个场景对应的模型, 电子设备可以使用该场景对应的样本语音进行训练。或者, 声纹识别模型也可以包括多场景融合模型以及多个场景分别对应的模型, 因此, 电子设备可以使用多个场景的样本语音对多场景融合模型进行训练, 并针对每个场景对应的模型, 电子设备可以使用该场景对应的样本语音进行训练。

若声纹识别模型为一个多场景融合模型, 则验证语音输入该多场景融合模型可以得到一个唯一的匹配得分, 在学习到高质量语音样本库的数据后, 该多场景融合模型将与实际使用场景越来越匹配。若声纹识别模型为多个场景分别对应的模型, 则可以通过对录入的验证语音进行场景检测, 将验证语音输入对应场景的模型进行声纹识别。进一步的, 若该验证语音已通过质量评估进入高质量语音样本库, 则将其进行数据增强, 增量学习更新该场景对应的模型, 可以使对应场景的模型与实际场景越来越匹配。

电子设备对样本语音进行特征提取时可以但不限于采用滤波器组(filter bank, FBank)、

梅尔频率倒谱系数 (mel-frequency cepstral coefficients, MFCC)、D-vector 等方法。

S205, 电子设备采集用户录入的验证语音。

具体实施中, 用户可以在电子设备的提示下说出验证语音。其中, 电子设备提示用户说出验证语的方法与电子设备提示用户说出注册语的方法类似, 重复之处不再一一赘述。

其中, 电子设备可以是在用户的操作触发下采集用户录入的验证语音, 例如, 用户通过操作电子设备触发验证指令, 从而电子设备在收到验证指令后采集提示用户录入验证语音, 并采集用户录入的验证语音。例如, 用户可以通过点击电子设备的触摸屏上声纹识别功能对应图标的相应位置触发验证指令, 从而电子设备提示用户说出验证语音; 又例如, 用户可以通过操作物理实体 (如物理键、鼠标、摇杆等) 进行触发; 又例如, 用户可以通过特定手势 (如双击电子设备的触摸屏等等) 进行触发验证指令, 从而电子设备提示用户说出验证语音。又例如, 用户可以向电子设备 (如智能手机、车载装置等等) 说出关键词 “声纹识别”, 电子设备通过麦克风 163 采集到用户发出的关键词 “声纹识别” 后触发验证指令, 并提示用户说出验证语音。

或者, 用户也可以在向电子设备说出用于控制电子设备的控制命令时, 电子设备可以采集该控制命令, 并将该控制命令作为验证语音进行声纹识别。即, 电子设备在接收到控制命令时触发验证指令, 并将该控制指令作为验证语音进行声纹识别。例如, 如图 4 所示, 用户可以向电子设备 (如智能手机、车载装置等等) 发出控制命令 “打开音乐”, 电子设备通过麦克风 163 采集到用户发出的语音 “打开音乐” 后, 将该语音作为验证语音进行声纹识别。又例如, 用户可以向电子设备 (如智能空调) 发出控制命令 “调到 27℃”, 电子设备通过麦克风 163 采集到用户发出的语音 “调到 27℃” 后, 将该语音作为验证语音进行声纹识别。

S206, 电子设备对验证语音进行特征提取以及场景检测。

电子设备对验证语音进行特征提取时, 可以但不限于采用 FBank、MFCC、D-vector 等方法。

进一步的, 电子设备在对验证语音进行场景检测之后可以给该验证语音加上场景标签。例如, 电子设备对验证语音进行场景检测之后确定该验证语音在车载场景中录入的, 则可以给该验证语音加上车载场景对应的场景标签。

示例性的, 场景检测的方法可以但不限于包括 GMM、深度神经网络 (deep neural network, DNN) 等。场景标签可以根据应用场景来选取, 比如居家场景、车载场景、背景音乐场景、嘈杂人声环境、以及远场场景、近场场景等。

一些实施方式中, 电子设备可以预先针对每个场景训练一个检测模型 (该检测模型可以是基于 GMM 算法的, 也可以是基于 DNN 算法的), 从而电子设备可以将验证语音依次输入各个场景对应的检测模型匹配得分, 根据各个场景对应模型的匹配得分确定该验证语音对应的场景。

另一些实施方式中, 电子设备也可以预先训练一个分类模型 (该分类模型可以是基于 DNN 算法的), 从而电子设备可以将验证语音输入该分类模型, 该分类模型可以输出分类结果, 该分类结果即为验证语音对应的场景。

S207, 电子设备将验证语音输入声纹注册阶段训练好的声纹识别模型进行匹配得分。如果匹配得分大于匹配阈值则可以判断该验证语音来自于注册人, 否则不是来自注册人。

其中, 匹配得分的方法可以但不限于包括: 余弦距离 (cosine distance, CDS)、线性

判别分析 (linear discriminant analysis, LDA)、概率线性判别分析 (probabilistic linear discriminant analysis, PLDA) 等算法。

具体来说, 若声纹识别模型为一个多场景融合模型, 则通过该多场景融合模型匹配得分可以得到一个得分, 若声纹识别模型包括多个场景分别对应的模型, 则可以通过多个场景对应的模型分别进行匹配得分, 得到多个得分, 然后结合步骤 S206 中得到的场景标签采用加权的方式得到一个融合得分。

进一步的, 电子设备在确定验证语音不是来自注册人时可以向用户输出识别结果。具体的, 电子设备可以在显示屏 132 上输出识别结果, 如图 5 所示, 电子设备可以在显示屏 132 上显示文字 “不是注册人! ”。又例如, 电子设备也可以通过扬声器 161 播报语音 “不是注册人”, 等等。

S208, 电子设备在确定验证语音来自注册人时, 可以结合该验证语音的场景标签, 对该验证语音进行质量评估。若该验证语音的质量得分大于质量阈值, 则可以将该验证语音加入高质量语音样本库。

示例性的, 对该验证语音进行质量评估的方法可以为: 确定验证语音的表征语音质量的参数的值来确定该验证语音是否为高质量的语音, 其中, 表征语音质量的参数可以但不限于包括如下参数中一项或多项: 信噪比 (signal-to-noise ratio, SNR)、分段信噪比 (segment signal-to-noise ratio, SegSNR)、语音质量的感知评价 (perceptual evaluation of speech quality, PESQ)、对数似然比测度 (log likelihood ratio measure, LLR) 等。

或者, 也可以将验证语音输入用于进行质量评估的模型中来确定验证语音是否为高质量的语音, 其中, 用于进行质量评估的模型可以是基于 GMM 算法的, 也可以是基于 DNN 算法的。具体的, 可以将验证语音输入该模型后得到一个质量得分, 然后根据该质量得分的高低确定验证语音是否为高质量的语音。

具体实施中, 声纹识别模型可以包括多个场景分别对应的模型, 高质量语音样本库也可以按照多个场景进行分类存储, 即高质量语音样本库可以包括该多个场景分别对应的样本库, 其中, 一个场景对应的样本库可以用于训练该场景对应的模型。基于此, 一种可能的实现方式为, 若验证语音的质量得分大于质量阈值, 电子设备可以将该验证语音加入到步骤 S206 中检测到的场景所对应的样本库中。

以声纹识别模型包括 A 场景对应的模型、B 场景对应的模型、C 场景对应的模型以及 D 场景对应的模型, 高质量语音样本库可以包括 A 场景对应的样本库、B 场景对应的样本库、C 场景对应的样本库以及 D 场景对应的样本库为例进行说明, 假设步骤 S206 中确定验证语音来自 A 场景, 则电子设备可以在验证语音的质量得分大于质量阈值时, 将该验证语音加入到 A 场景对应的样本库中。

S209, 电子设备将高质量语音样本库的语音进行数据增强处理, 将处理后的语音数据用于增量学习, 更新声纹识别模型。

其中, 增量学习的算法可以但不限于包括: 方法 1, 将增强后的语音数据通过加权的方式与原始注册语音相加, 将相加得到的语音数据用于训练声纹识别模型; 方法 2, 单独基于增强后的语音数据对上一次训练得到声纹识别模型进行训练, 得到新的声纹识别模型, 将新的声纹识别模型与上一次训练得到声纹识别模型进行加权相加完成模型更新。

步骤 S209 中结合场景标签, 将高质量语音库内语音进行数据增强, 可以在用户使用过程中得到更加丰富的数据。比如, 近场干净的语音可以增强得到远场干净的语音, 居家

场景低噪语音可以增强得到居家场景中噪的语音。并且，在声纹识别模型包括多个场景对应的模型，高质量语音样本库按照多个场景进行分类存储的情况下，可以通过增量学习，更新对应场景的模型。通过将用户日常验证语音数据中的高质量语音进行数据增强后用于增量学习，更新声纹识别模型，可以使声纹识别模型与实际使用场景越来越匹配，从而可以提升声纹识别系统的鲁棒性。

为了更好地理解本申请实施例，以下结合具体应用场景，对声纹识别过程进行具体详细描述。

场景一：对于使用场景经常变化的情况，比如手机、耳机、手环等随身携带的电子设备，由于手机、耳机、耳环等随身携带的电子设备会随着用户的移动处于不同的场景，例如，用户从家里出来，开车到达商场，这种情况下，这些随身携带的电子设备经历了从家居场景移动到车载场景，然后又进入商场场景。用户设备在这些随身携带的电子设备上

进行声纹识别时，可以使用如下步骤 S601~S614 来实现声纹识别。

如图 6 所示，声纹识别过程具体可以包括：

S601，电子设备采集用户的 k 条注册语音。其中，k 可以为大于或等于 1 的整数。执行步骤 S602。

其中，用户可以在电子设备的提示下多次录入注册语音，提示方法具体可以参阅上述步骤 S201 所述方法，这里不再重复赘述。从而，电子设备可以通过麦克风 163 采集用户的 k 条注册语音。

S602，电子设备将 k 条注册语音加入高质量语音样本库。执行步骤 S603。

S603，电子设备对 k 条注册语音进行数据增强处理，得到样本语音。数据增强处理的方法具体可以参阅步骤 S203 所述方法，这里不再重复赘述。其中，一条注册语音可以生成多条不同噪声等级、不同远场等级的样本语音。执行步骤 S604。

具体来说，高质量语音样本库可以针对不同场景分类存储。因此电子设备可以针对不同场景对 k 条注册语音分别进行数据增强处理，从而针对不同场景可以生成该场景对应的样本语音。

例如，电子设备可以针对 A 场景对 k 条注册语音进行数据增强处理，具体的，电子设备可以基于一条注册语音生成 s1 条噪声等级不同、远场等级不同的样本数据，从而得到 A 场景对应的 $k \times s1$ 条样本语音；针对 B 场景对 k 条注册语音进行数据增强处理，具体的，电子设备可以基于一条注册语音生成 s2 条噪声等级不同、远场等级不同的样本数据，从而得到 B 场景对应的 $k \times s2$ 条样本语音；针对 C 场景对 k 条注册语音进行数据增强处理，具体的，电子设备可以基于一条注册语音生成 s3 条噪声等级不同、远场等级不同的样本数据，从而得到 C 场景对应的 $k \times s3$ 条样本语音。

进一步的，电子设备可以针对每个场景，将该场景的样本语音存储到该场景对应的样本库中，如，将 A 场景的 $k \times s1$ 条样本语音存储到 A 场景对应的样本库 1，将 B 场景的 $k \times s2$ 条样本语音存储到 B 场景对应的样本库 2，C 场景的 $k \times s3$ 条样本语音存储到 C 场景对应的样本库 3。

一些实施例中，针对每个场景，电子设备可以采用该场景对应的噪声源对 k 条注册语音进行数据增强处理，得到该场景对应的样本语音，其中，该场景对应的噪声源可以是在该场景采集的噪声数据，或者，也可以是针对该场景仿真生成的噪声数据，等等。例如，针对 A 场景，电子设备可以采用 A 场景的噪声源对注册数据进行数据增强处理，针对 B

场景，电子设备可以采用 B 场景的噪声源对注册数据进行数据增强处理。

S604，电子设备对样本语音进行特征提取，并基于提取的特征训练声纹识别模型库中的模型，得到训练好的模型。其中，特征提取的方法具体参阅步骤 S204 所述方法，重复之处不再赘述。执行步骤 S605。

其中，电子设备可以建立一个多场景融合模型，即声纹识别模型库中包括一个多场景融合模型。电子设备可以采用步骤 S603 中获取的样本数据对该多场景融合模型进行训练。

或者，电子设备也可以针对不同场景分别建立模型，即声纹识别模型可以包括多个模型，如近场安静模型、近场居家模型、远场居家模型、车载模型等，并可以采用各个场景的样本库对该场景对应的模型进行训练，例如，采用 A 场景的样本库 1 训练 A 场景对应的模型，示例性的，如，近场安静场景的样本库训练近场安静模型，采用近场居家场景的样本库训练近场居家模型，采用远场居家场景的样本库训练远场居家模型，采用车载场景的样本库训练车载模型等。

示例性的，电子设备可以针对居家场景、车载场景、商场场景、工作场景分别建立模型，并且电子设备在采集到用户录入的注册语音后，基于该注册语音分别针对居家场景、车载场景、商场场景、工作场景进行语音数据增强，从而分别得到居家场景、车载场景、商场场景、工作场景下的样本语音，然后可以采用居家场景的样本数据对居家场景的模型进行训练、采用车载场景的样本数据对车载场景的模型进行训练、采用商场场景的样本数据对商场场景的模型进行训练、采用工作场景的样本数据对工作场景的模型进行训练。从而，电子设备在采集用户录入的验证语音后，可以结合对该验证语音场景检测的结果来选择对应场景的模型，例如，假设对该验证语音进行场景检测得到的结果为居家场景，则电子设备可以将该验证语音输入居家场景的模型进行匹配。

当然，电子设备也可以针对不同场景分别建立模型，同时建立一个多场景融合模型，即声纹识别模型可以包括一个多场景融合模型以及多个场景分别对应的模型。S605，电子设备采集用户录入的验证语音。

S605，电子设备采集用户录入的验证语音。执行步骤 S606。

步骤 S605 具体可以参阅步骤 S205，这里不再重复赘述。

S606，电子设备对验证语音进行特征提取以及场景检测。执行步骤 S607。

步骤 S606 具体可以参阅步骤 S206，这里不再重复赘述。

S607，电子设备将验证语音输入步骤 S604 中训练好的模型进行匹配得分，得到第一分值。执行步骤 S608。

进行匹配得分的方法可以有多种，其中，一种可能的方法为：声纹识别模型包括多个场景分别对应的模型，电子设备可以选择步骤 S606 中所检测到的场景（假设为 A 场景）对应的模型进行匹配得分，即将验证语音输入 A 场景对应的模型进行匹配得分，从而得到第一分值。

另一种可能的方法为：声纹识别模型包括多场景融合模型，电子设备可以选择多场景融合模型来进行匹配得分，即将验证语音输入多场景融合模型中进行匹配得分，从而得到第一分值。

又一种可能的方法为：声纹识别模型包括多个场景分别对应的模型，电子设备可以将验证语音分别输入各个场景对应的模型进行匹配得分，得到多个得分，将多个得分进行融合，从而得到第一分值。示例性的，第一分值可以但不限于为：多个得分的平均值、多个

得分的加权值等等。

具体实施中还可以采用其他方法进行匹配得分，这里不再一一列举。

需要补充说明的是，如果在实际实施过程中由于某种原因不想在声纹识别模型库中同时维护多个模型，则步骤 S604 中可以只建立并训练一个多场景融合模型。

5 S608, 电子设备判断第一分值是否大于第一阈值。若是, 执行步骤 S609 以及步骤 S611; 若否, 执行步骤 S610。

S609, 电子设备输出声纹识别结果: 是注册人。

具体的, 电子设备可以通过在显示屏 132 中呈现文字“是注册人”, 示例性的, 呈现界面可以如图 7 所示。

10 或者, 电子设备也可以通过扬声器 163 播报语音“是注册人”。

S610, 电子设备输出声纹识别结果: 不是注册人。

具体的, 电子设备可以通过在显示屏 132 中呈现文字“不是注册人”, 示例性的, 呈现界面可以如图 5 所示。

或者, 电子设备也可以通过扬声器 163 播报语音“不是注册人”。

15 S611, 电子设备对该验证语音进行质量评估, 得到第二分值。执行步骤 S612。

具体的, 电子设备可以结合步骤 S606 中检测得到的场景对该验证语音进行质量评估。

在一些实施例中, 电子设备可以根据步骤 S606 中检测得到的场景对应的模型对验证语音进行打分, 如果分数高于质量评估的阈值, 则加入高质量语音样本库。

20 或者, 也可以采用质量评估的方法确定验证语音的质量评估得分, 并将该质量评估得分与步骤 S606 所检测得到的场景对应的阈值进行比较, 来确定是不是该场景的高质量语音。

S612, 电子设备判断第二分值是否大于第二阈值。若是, 执行步骤 S613; 若否, 则结束。

S613, 电子设备将验证语音存储到高质量语音样本库。执行步骤 S614。

25 具体的, 若高质量语音样本库按照多个场景进行分类存储, 即高质量语音样本库可以包括该多个场景分别对应的样本库, 则电子设备可以将验证语音存储到步骤 S606 所检测到的场景对应的样本库, 如, 步骤 S606 检测到验证语音来源于居家场景, 则电子设备可以将验证语音存储到居家场景对应的样本库中。

30 电子设备还可以将验证语音进行数据增强, 并数据增强后的验证语音存储到高质量语音样本库。例如, 如验证语音为居家场景语音, 则可以将验证语音进行数据增强后得到远场居家语音, 或者, 也可以将验证语音进行数据增强后得到其他噪声等级的居家场景语音, 其中, 数据增强后得到的居家场景语音的噪声等级可以大于验证语音。

S614, 电子设备基于高质量语音样本库新增的语音数据进行增量学习, 更新声纹识别模型库中的模型。

35 具体来说, 若声纹识别模型库中包括一个多场景融合模型, 则电子设备可以基于高质量语音样本库新增的语音数据对上一次训练得到多场景融合模型进行训练, 得到新的多场景融合模型。电子设备可以将该新的多场景融合模型与上一次训练得到的多场景融合模型进行加权相加完成模型更新。或者, 电子设备也可以将高质量语音样本库新增的语音数据加权的方式与高质量语音样本库原来存储的语音数据相加, 并基于相加得到的语音数据训练上一次训练得到的多场景融合模型完成模型更新。

40

若声纹识别模型库中包括多个场景分别对应的模型，以步骤 S606 检测到的场景为车载场景，步骤 S613 中将验证语音存储到车载场景的样本库为例，电子设备可以基于车载场景的样本库新增的语音数据对上一次训练得到车载场景模型进行训练，得到新的车载场景模型。电子设备可以将该新的车载场景模型与上一次训练得到的车载场景模型进行加权相加完成模型更新。或者，电子设备也可以将车载场景的样本库新增的语音数据加权的方式与车载场景的样本库原来存储的语音数据相加，并基于相加得到的语音数据训练上一次训练得到的车载场景模型完成模型更新。

上述声纹识别过程通过对原始注册语音进行多场景标签的数据增强，可以解决注册语音场景单一而验证语音场景多变导致数据不匹配问题。并且，通过将高质量的验证语音加入到高质量语音样本库，并进行数据增强和增量学习，更新声纹识别模型库中的模型，从而用户使用过程中，声纹识别模型库中的模型可以越来越适用于用户实际使用场景。因此，通过上述声纹识别方法可以提高声纹识别算法对多场景和变场景的鲁棒性。

场景二：对于使用场景常常为某一种场景的情况，比如用于智能音箱、智能家居、车载装置等设备上可以使用如下步骤 S801~S817 来实现声纹识别。

如图 8 所示，声纹识别过程具体可以包括：

S801~S813，具体可以参阅步骤 S601~S613，这里不再重复赘述。

其中，在步骤 S813 之后可以执行步骤 S814。

S814，电子设备判断验证语音是否为高频场景下的高质量语音。若是，执行步骤 S815。若否，执行步骤 S817。其中，步骤 S817 可以参阅步骤 S614，这里不再重复赘述。

其中，电子设备进行声音识别时所采集的验证语音大部分是来自某场景，则可以认为该场景为高频场景。

具体实施中，电子设备可以通过如下方式判断验证语音是否为高频场景下的高质量语音：针对步骤 S806 中检测到的场景（假设为 A 场景），电子设备可以统计最近 N 次声纹识别过程中验证语音的场景检测结果为 A 场景的次数 n，若 n 大于第三阈值（或 n/N 大于第四阈值），则电子设备可以确定 A 场景为高频场景，则验证语音为高频场景下的高质量语音，反之，则验证语音不是高频场景下的高质量语音。

例如，假设步骤 S806 中针对验证语音进行场景检测得到的结果是居家场景，则电子设备可以通过最近 10 次声纹识别过程中验证语音的场景检测结果为居家场景的次数 n，若 n 大于 5（即第三阈值），则可以判断居家场景为高频场景，即验证语音为高频场景下的高质量语音；若 n 小于或等于 5，则可以判断居家场景不是高频场景，即验证语音不是高频场景下的高质量语音。

又例如，假设步骤 S806 中针对验证语音进行场景检测得到的结果是车载场景，则电子设备可以通过最近 20 次声纹识别过程中验证语音的场景检测结果为居家场景的次数 n，若 $n/20$ 大于 50%（即第四阈值），则可以判断车载场景为高频场景，即验证语音是高频场景下的高质量语音；若 $n/20$ 小于或等于 50%，则可以判断车载场景不是高频场景，即验证语音不是高频场景下的高质量语音。

S815，电子设备将第一场景的样本库中的 i 条样本语音进行数据增强，其中，第一场景为步骤 S806 中检测到的场景。i 条语音可以是第一场景的样本库的所有样本语音，也可以是第一场景的样本库的部分样本语音。执行步骤 S816。

具体的，针对 i 条样本语音中的每一条样本语音，电子设备可以该样本语音进行数据增强，得到 j 个不同噪声等级的噪声语音，其中，该 j 个噪声语音的噪声等级均大于该样本语音。或者，针对 i 条样本语音中的每一条样本语音，电子设备可以该样本语音进行数据增强，得到 k 个不同远场等级的远场语音，其中，该 k 个远场语音的远场等级均大于该样本语音。或者，针对 i 条样本语音中的每一条样本语音，电子设备可以该样本语音进行数据增强，得到 j 个不同噪声等级的噪声语音，再针对每个噪声语音进行数据增强，得到 $j \times k$ 个远场噪声语音。

S816，电子设备基于步骤 S815 中得到的语音数据进行增量学习，得到高频场景下的子模型。

具体的，电子设备可以将步骤 S815 中得到的语音数据按照噪声等级分为 M 个组，其中，同一组内的语音数据的噪声等级相同，或者，同一组内的语音数据的噪声等级在同一个噪声等级范围内。然后，针对每一组，电子设备使用该组的语音数据对上一次训练得到的第一场景的模型进行训练，得到该组对应的子模型，并加入到声纹识别模型库中。具体来说，针对每一组，电子设备可以使用该组的语音数据对上一次训练得到的第一场景下该组对应的子模型进行训练。

上述声纹识别过程中，通过对原始注册语音进行多场景标签的数据增强，解决注册语音场景与验证语音场景不匹配问题。并且，通过增加高频场景判断，将高频场景下的高质量的验证语音加入到高质量语音样本库，并进行数据增强和增量学习，精细化高频场景下的模型，使得电子设备在该高频场景的不同噪声等级下或远场等级下可以更加准确地进行声纹识别。比如，在车载场景中，可精准匹配到时速 30km/h、60km/h、90km/h、120km/h 等对应的子模型，而不是一个粗略的车载场景模型。再比如，在远场居家环境中，可以精准匹配到远场 3m、4m、5m 等对应的子模型，而不是一个粗略的远场居家环境说话人模型。从而用户使用过程中，可以根据场景检测结果匹配到高频场景下的子模型，使得声纹识别更加准确，并且随着使用数据的增加，声纹识别模型库中的模型通过的增量学习不断更新，可以越来越准确。

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。

- 5 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

权 利 要 求

1、一种声纹注册方法，其特征在于，包括：

电子设备提示用户录入注册语音；

所述电子设备采集所述用户录入的所述注册语音；

5 所述电子设备基于所述注册语音生成远场条件下的样本语音；

所述电子设备基于所述样本语音对声纹识别模型进行训练。

2、如权利要求 1 所述的方法，其特征在于，所述电子设备基于所述注册语音生成远场条件下的样本语音，包括：

所述电子设备模拟在远场条件下对声音的混响；

10 所述电子设备基于所述远场条件下对声音的混响仿真生成所述注册语音在所述远场条件下的样本数据。

3、如权利要求 1 所述的方法，其特征在于，所述电子设备基于所述注册语音生成远场条件下的样本语音，包括：

所述电子设备基于所述注册语音以及噪声数据生成噪声语音；

15 所述电子设备模拟远场条件下对声音的混响；

所述电子设备基于所述远场条件下对声音的混响仿真生成所述噪声语音在所述远场条件下的样本数据。

4、如权利要求 2 或 3 所述的方法，其特征在于，所述电子设备模拟远场条件下对声音的混响，包括：

20 所述电子设备基于远场条件仿真声音的墙壁反射，得到房间冲击响应 RIR。

5、如权利要求 1 至 4 任一项所述的方法，其特征在于，所述电子设备基于所述样本语音对声纹识别模型进行训练，包括：

所述电子设备对所述样本语音进行特征提取，得到特征数据；

所述电子设备基于所述特征数据对声纹识别模型进行训练。

25 6、如权利要求 5 所述的方法，其特征在于，所述声纹识别模型包括一个或多个子模型，其中，一个子模型对应一个场景；

所述电子设备基于所述特征数据对声纹识别模型进行训练，包括：

所述电子设备基于所述特征数据分别对所述一个或多个子模型进行训练。

30 7、如权利要求 5 所述的方法，其特征在于，所述声纹识别模型包括一个融合模型，其中，所述融合模型对应一个或多个场景；

所述电子设备基于所述特征数据对声纹识别模型进行训练，包括：

所述电子设备基于所述特征数据对所述融合模型进行训练。

8、一种声纹识别方法，其特征在于，所述方法包括：

电子设备提示用户录入验证语音；

35 所述电子设备采集所述用户录入的所述验证语音；

所述电子设备将所述验证语音输入声纹识别模型进行匹配，得到匹配结果，其中，所述声纹识别模型基于权利要求 1~7 任一项所述的方法训练得到；

所述电子设备基于所述匹配结果确定所述用户是否为所述声纹识别模型的注册人。

9、如权利要求 8 所述的方法，其特征在于，在所述电子设备采集所述用户录入的所

述验证语音之后,还包括:

所述电子设备对所述验证语音进行场景检测。

10、如权利要求 9 所述的方法,其特征在于,所述声纹识别模型包括一个或多个子模型,其中,一个子模型对应一个场景;

5 所述电子设备将所述验证语音输入声纹识别模型进行匹配,包括:

所述电子设备将所述验证语音输入第一场景对应的子模型进行匹配,其中,所述第一场景为所述场景检测的结果。

11、如权利要求 8 至 10 任一项所述的方法,其特征在于,所述方法还包括:

10 若所述用户是所述声纹识别模型的注册人,所述电子设备对所述验证语音进行质量评估,得到质量评估结果;

若所述质量评估结果表示所述验证语音为高质量语音,所述电子设备基于所述验证语音对所述声纹识别模型进行增量学习。

12、如权利要求 11 所述的方法,其特征在于,所述电子设备基于所述验证语音对所述声纹识别模型进行增量学习,包括:

15 所述电子设备对所述验证语音进行数据增强处理,得到处理后的语音数据;

所述电子设备基于所述处理的语音数据对所述声纹识别模型进行增量学习。

13、如权利要求 12 所述的方法,其特征在于,在所述电子设备对所述验证语音进行数据增强处理之前,所述方法还包括:

所述电子设备确定所述验证语音所在的第一场景为高频场景;

20 所述电子设备对所述验证语音进行数据增强处理,包括:

所述电子设备将所述验证语音进行数据增强处理,得到 j 个不同噪声等级的样本语音;

所述电子设备基于所述处理的语音数据对所述声纹识别模型进行增量学习,包括:

所述电子设备将所述 j 个样本语音按照噪声等级进行分组,得到 M 组语音数据,所述 M 为大于 0 且不大于 j 的整数;

25 所述电子设备基于所述 M 组语音数据分别对所述第一场景对应的子模型进行训练,得到 M 个高频子模型。

14、一种声纹注册装置,其特征在于,包括:

第一器件,麦克风以及处理器,其中,所述第一器件为扬声器或者显示屏;

所述处理器,用于执行:

30 触发所述第一器件提示用户录入注册语音;

通过所述麦克风采集所述用户录入的所述注册语音;

基于所述注册语音生成远场条件下的样本语音;

基于所述样本语音对声纹识别模型进行训练。

15、如权利要求 14 所述的装置,其特征在于,所述处理器,在基于所述注册语音生成远场条件下的样本语音时,具体用于:

模拟所述注册语音在远场条件下对声音的混响;

基于所述远场条件下对声音的混响仿真生成所述注册语音在所述远场条件下的样本数据。

16、如权利要求 14 所述的装置,其特征在于,所述处理器,在基于所述注册语音生成远场条件下的样本语音时,具体用于:

基于所述注册语音以及噪声数据生成噪声语音；

模拟远场条件下对声音的混响；

基于所述远场条件下对声音的混响仿真生成所述噪声语音在所述远场条件下的样本数据。

17、如权利要求 15 或 16 所述的装置，其特征在于，所述处理器，在模拟远场条件下对声音的混响时，具体用于：

基于远场条件仿真声音的墙壁反射，得到房间冲击响应 RIR。

18、如权利要求 14 至 17 任一项所述的装置，其特征在于，所述处理器，在基于所述样本语音对声纹识别模型进行训练时，具体用于：

对所述样本语音进行特征提取，得到特征数据；

基于所述特征数据对声纹识别模型进行训练。

19、如权利要求 18 所述的装置，其特征在于，所述声纹识别模型包括一个或多个子模型，其中，一个子模型对应一个场景；

所述处理器，在基于所述特征数据对声纹识别模型进行训练时，具体用于：

基于所述特征数据分别对所述一个或多个子模型进行训练。

20、如权利要求 18 所述的装置，其特征在于，所述声纹识别模型包括一个融合模型，其中，所述融合模型对应一个或多个场景；

所述处理器，在基于所述特征数据对声纹识别模型进行训练时，具体用于：

基于所述特征数据对所述融合模型进行训练。

21、一种声纹识别装置，其特征在于，所述装置包括：

第一器件、麦克风以及处理器，其中，所述第一器件为扬声器或者显示屏；

所述处理器，用于执行：

触发所述第一器件提示用户录入验证语音；

通过所述麦克风采集所述用户录入的所述验证语音；

将所述验证语音输入声纹识别模型进行匹配，得到匹配结果，其中，所述声纹识别模型为权利要求 14~18 任一项所述的装置训练得到；

基于所述匹配结果确定所述用户是否为所述声纹识别模型的注册人。

22、如权利要求 21 所述的装置，其特征在于，所述处理器，还用于：

在通过所述麦克风采集所述用户录入的所述验证语音之后，对所述验证语音进行场景检测。

23、如权利要求 22 所述的装置，其特征在于，所述声纹识别模型包括一个或多个子模型，其中，一个子模型对应一个场景；

所述处理器，在将所述验证语音输入声纹识别模型进行匹配时，具体用于：

将所述验证语音输入第一场景对应的子模型进行匹配，其中，所述第一场景为所述场景检测的结果。

24、如权利要求 21~23 任一项所述的装置，其特征在于，所述处理器，还用于：

若所述用户是所述声纹识别模型的注册人，对所述验证语音进行质量评估，得到质量评估结果；

若所述质量评估结果表示所述验证语音为高质量语音，基于所述验证语音对所述声纹识别模型进行增量学习。

25、如权利要求 24 所述的装置，其特征在于，所述处理器，在基于所述验证语音对所述声纹识别模型进行增量学习时，具体用于：

对所述验证语音进行数据增强处理，得到处理后的语音数据；
基于所述处理的语音数据对所述声纹识别模型进行增量学习。

5 26、如权利要求 25 所述的装置，其特征在于，所述处理器，还用于：

在对所述验证语音进行数据增强处理之前，确定所述验证语音所在的第一场景为高频场景；

所述处理器，在对所述验证语音进行数据增强处理时，具体用于：

将所述验证语音进行数据增强处理，得到 j 个不同噪声等级的样本语音；

10 所述处理器，在基于所述处理的语音数据对所述声纹识别模型进行增量学习时，具体用于：

将所述 j 个样本语音按照噪声等级进行分组，得到 M 组语音数据，所述 M 为大于 0 且不大于 j 的整数；

15 基于所述 M 组语音数据分别对所述第一场景对应的子模型进行训练，得到 M 个高频子模型。

27、一种芯片，其特征在于，所述芯片与电子设备中的存储器耦合，执行如权利要求 1 至 13 任一所述的方法。

28、一种计算机存储介质，其特征在于，所述计算机存储介质中存储计算机指令，该计算机指令在被一个或多个处理器执行时实现权利要求 1 至 13 中任一项所述的方法。

20

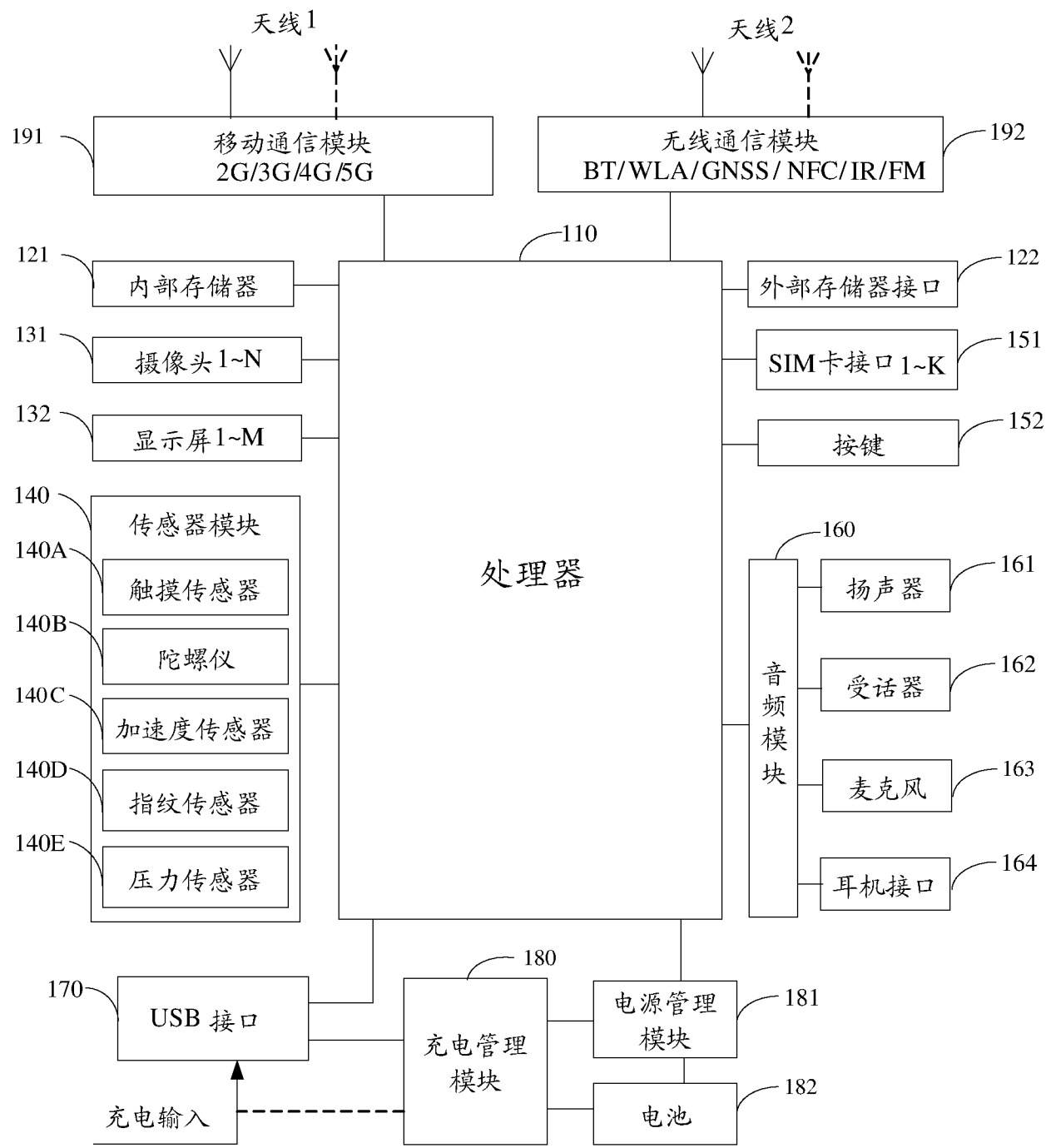


图 1

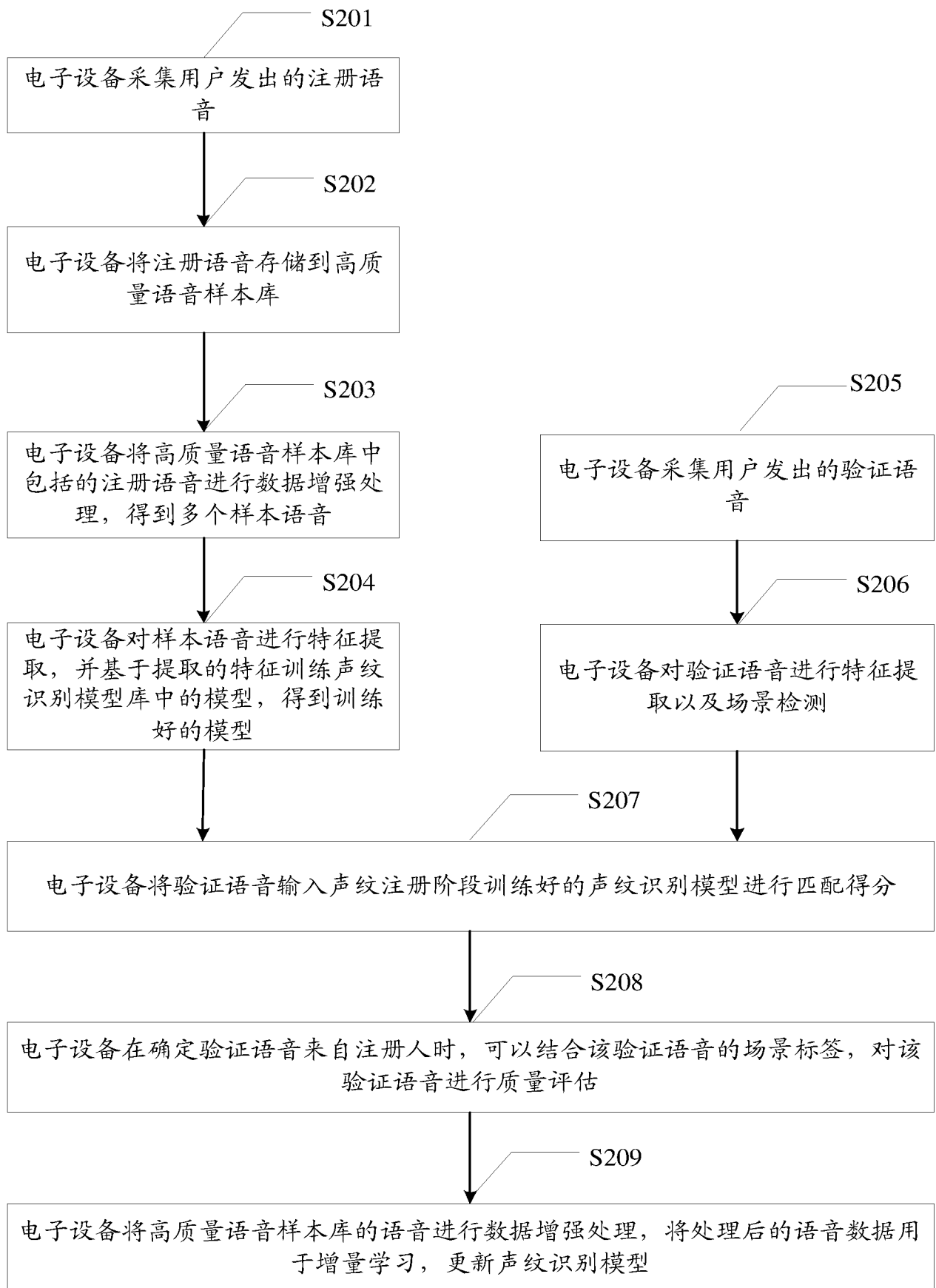


图 2



图 3

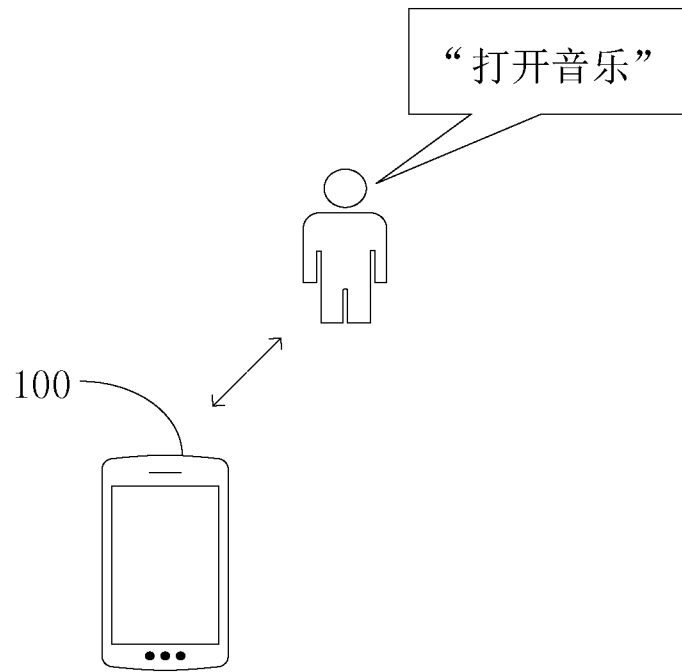


图 4



图 5

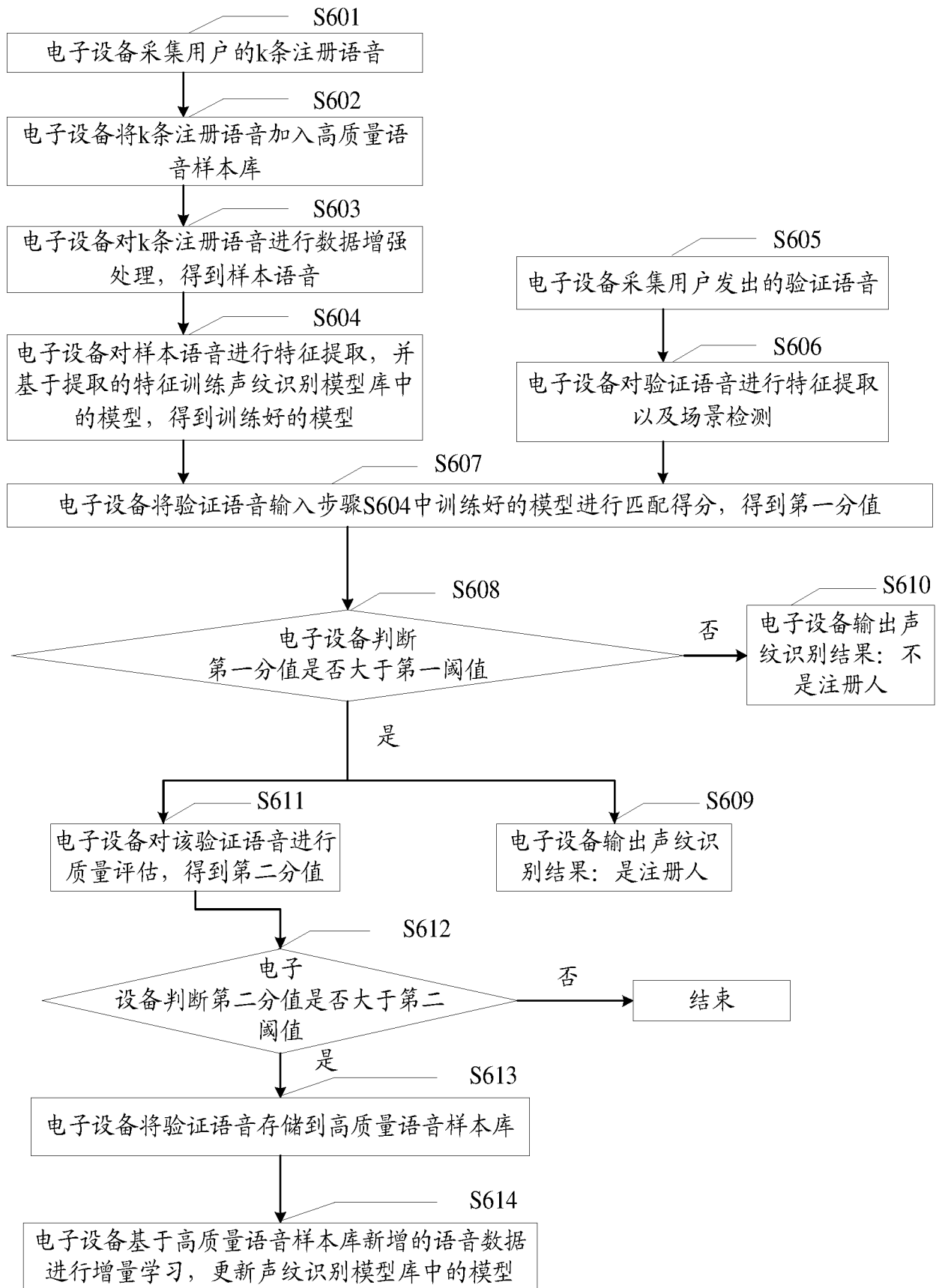


图 6



图 7

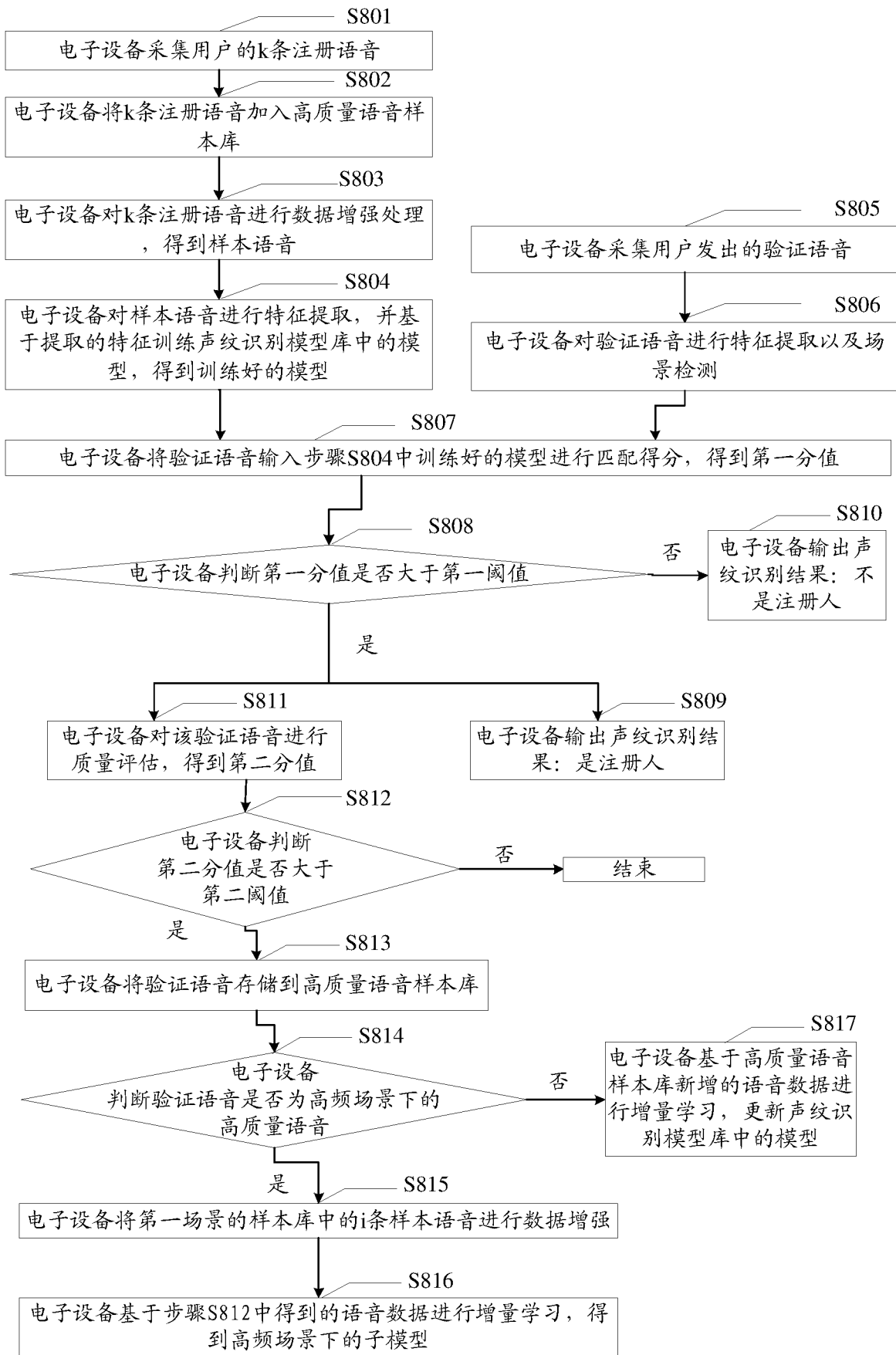


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/104545

A. CLASSIFICATION OF SUBJECT MATTER

G10L 17/20(2013.01)i; G10L 17/04(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)

G10L 17

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)

CNABS, CNTXT, VEN, USTXT, WOTXT; 语音, 音频, 近场, 增强, 噪声, 远场, 样本, 训练, 模型, 声纹, 注册, 识别, speech, sound, voice, enhanc+, noise, near, far, sample, train+, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109841218 A (BEIJING SOUNDAL TECHNOLOGY CO., LTD.) 04 June 2019 (2019-06-04) description, paragraphs [0047]-[0081]	1-7, 14-20, 27, 28
X	CN 107481731 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 15 December 2017 (2017-12-15) description, paragraphs [0056]-[0130]	1-7, 14-20, 27, 28
X	CN 108269567 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 10 July 2018 (2018-07-10) description, paragraphs [0028]-[0035] and [0050]-[0055]	1-7, 14-20, 27, 28
Y	CN 109841218 A (BEIJING SOUNDAL TECHNOLOGY CO., LTD.) 04 June 2019 (2019-06-04) description, paragraphs [0047]-[0081]	8-13, 21-26
Y	CN 104952450 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 30 September 2015 (2015-09-30) description, paragraphs [0031]-[0056]	8-13, 21-26



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

“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

24 October 2020

Date of mailing of the international search report

28 October 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/104545**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107680586 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 09 February 2018 (2018-02-09) entire document	1-28
A	JP 2007279349 A (TOSHIBA CORP.) 25 October 2007 (2007-10-25) entire document	1-28

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/104545

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109841218	A	04 June 2019	None			
CN	107481731	A	15 December 2017	None			
CN	108269567	A	10 July 2018	US	2019228791	A1	25 July 2019
CN	104952450	A	30 September 2015	CN	104952450	B	17 November 2017
CN	107680586	A	09 February 2018	US	2019043482	A1	07 February 2019
JP	2007279349	A	25 October 2007	JP	4245617	B2	25 March 2009
				US	2007276662	A1	29 November 2007
				CN	101051461	A	10 October 2007

国际检索报告

国际申请号

PCT/CN2020/104545

A. 主题的分类 G10L 17/20(2013.01)i; G10L 17/04(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) G10L 17 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)) CNABS, CNTXT, VEN, USTXT, WOTXT; 语音, 音频, 近场, 增强, 噪声, 远场, 样本, 训练, 模型, 声纹, 注册, 识别, speech, sound, voice, enhanc+, noise, near, far, sample, train+, model																										
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段</td> <td>1-7、14-20、27-28</td> </tr> <tr> <td>X</td> <td>CN 107481731 A (百度在线网络技术北京有限公司) 2017年 12月 15日 (2017 - 12 - 15) 说明书第0056-0130段</td> <td>1-7、14-20、27-28</td> </tr> <tr> <td>X</td> <td>CN 108269567 A (北京百度网讯科技有限公司) 2018年 7月 10日 (2018 - 07 - 10) 说明书第0028-0035、0050-0055段</td> <td>1-7、14-20、27-28</td> </tr> <tr> <td>Y</td> <td>CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段</td> <td>8-13、21-26</td> </tr> <tr> <td>Y</td> <td>CN 104952450 A (百度在线网络技术北京有限公司) 2015年 9月 30日 (2015 - 09 - 30) 说明书第0031-0056段</td> <td>8-13、21-26</td> </tr> <tr> <td>A</td> <td>CN 107680586 A (百度在线网络技术北京有限公司) 2018年 2月 9日 (2018 - 02 - 09) 全文</td> <td>1-28</td> </tr> <tr> <td>A</td> <td>JP 2007279349 A (TOSHIBA CORP.) 2007年 10月 25日 (2007 - 10 - 25) 全文</td> <td>1-28</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段	1-7、14-20、27-28	X	CN 107481731 A (百度在线网络技术北京有限公司) 2017年 12月 15日 (2017 - 12 - 15) 说明书第0056-0130段	1-7、14-20、27-28	X	CN 108269567 A (北京百度网讯科技有限公司) 2018年 7月 10日 (2018 - 07 - 10) 说明书第0028-0035、0050-0055段	1-7、14-20、27-28	Y	CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段	8-13、21-26	Y	CN 104952450 A (百度在线网络技术北京有限公司) 2015年 9月 30日 (2015 - 09 - 30) 说明书第0031-0056段	8-13、21-26	A	CN 107680586 A (百度在线网络技术北京有限公司) 2018年 2月 9日 (2018 - 02 - 09) 全文	1-28	A	JP 2007279349 A (TOSHIBA CORP.) 2007年 10月 25日 (2007 - 10 - 25) 全文	1-28
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																								
X	CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段	1-7、14-20、27-28																								
X	CN 107481731 A (百度在线网络技术北京有限公司) 2017年 12月 15日 (2017 - 12 - 15) 说明书第0056-0130段	1-7、14-20、27-28																								
X	CN 108269567 A (北京百度网讯科技有限公司) 2018年 7月 10日 (2018 - 07 - 10) 说明书第0028-0035、0050-0055段	1-7、14-20、27-28																								
Y	CN 109841218 A (北京声智科技有限公司) 2019年 6月 4日 (2019 - 06 - 04) 说明书第0047-0081段	8-13、21-26																								
Y	CN 104952450 A (百度在线网络技术北京有限公司) 2015年 9月 30日 (2015 - 09 - 30) 说明书第0031-0056段	8-13、21-26																								
A	CN 107680586 A (百度在线网络技术北京有限公司) 2018年 2月 9日 (2018 - 02 - 09) 全文	1-28																								
A	JP 2007279349 A (TOSHIBA CORP.) 2007年 10月 25日 (2007 - 10 - 25) 全文	1-28																								
☐ 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 “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 2020年 10月 24日		国际检索报告邮寄日期 2020年 10月 28日																								
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 房倩 电话号码 86-(010)-62089621																								

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/104545

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	109841218	A	2019年 6月 4日	无			
CN	107481731	A	2017年 12月 15日	无			
CN	108269567	A	2018年 7月 10日	US	2019228791	A1	2019年 7月 25日
CN	104952450	A	2015年 9月 30日	CN	104952450	B	2017年 11月 17日
CN	107680586	A	2018年 2月 9日	US	2019043482	A1	2019年 2月 7日
JP	2007279349	A	2007年 10月 25日	JP	4245617	B2	2009年 3月 25日
				US	2007276662	A1	2007年 11月 29日
				CN	101051461	A	2007年 10月 10日