

(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
2024 年 12 月 12 日 (12.12.2024)



(10) 国际公布号
WO 2024/251169 A1

- (51) 国际专利分类号:
G10L 15/20 (2006.01) G10L 15/183 (2013.01)
- (21) 国际申请号: PCT/CN2024/097609
- (22) 国际申请日: 2024 年 6 月 5 日 (05.06.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310680032.X 2023年6月8日 (08.06.2023) CN
- (71) 申请人: 阿里巴巴 (中国) 有限公司 (ALIBABA (CHINA) CO., LTD) [CN/CN]; 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。
- (72) 发明人: 史莫晗 (SHI, Mohan); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室,

Zhejiang 310052 (CN)。杜志浩 (DU, Zhihao); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。陈谦 (CHEN, Qian); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。俞帆 (YU, Fan); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。张仕良 (ZHANG, Shiliang); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。李洪泽 (LI, Yangze); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。张结 (ZHANG, Jie); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。戴礼荣 (DAI, Lirong); 中国

(54) Title: SPEECH RECOGNITION METHOD, DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 语音识别方法、设备和存储介质

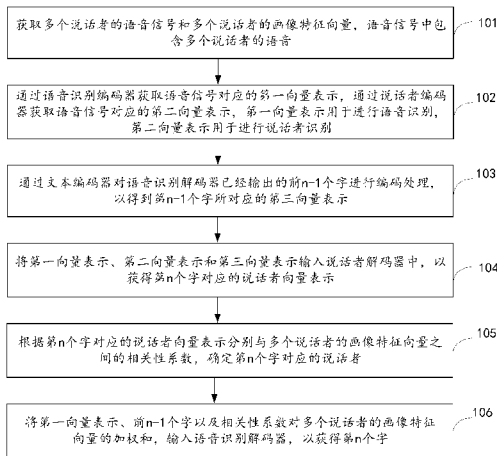


图 1

- 101 Acquire speech signals of a plurality of speakers and portrait feature vectors of the plurality of speakers, the speech signals comprising speeches of the plurality of speakers
- 102 Acquire a first vector representation corresponding to each speech signal by means of a speech recognition encoder, and acquire a second vector representation corresponding to each speech signal by means of a speaker encoder, the first vector representation being used for performing speech recognition, and the second vector representation being used for performing speaker recognition
- 103 By means of a text encoder, encode first n-1 words that have been output by a speech recognition decoder, to obtain a third vector representation corresponding to an (n-1)-th word
- 104 Input the first vector representation, the second vector representation, and the third vector representation into a speaker decoder to obtain a speaker vector representation corresponding to an n-th word
- 105 According to correlation coefficients between the speaker vector representation corresponding to the n-th word and the portrait feature vectors of the plurality of speakers, determine a speaker corresponding to the n-th word
- 106 Input the first vector representation, the first n-1 words, and the weighted sum of the correlation coefficients on the portrait feature vectors of the plurality of speakers into the speech recognition decoder to obtain the n-th word

(57) Abstract: The present disclosure provides a speech recognition method, a device, and a storage medium. The method comprises: acquiring a first vector representation corresponding to a speech signal by means of a speech recognition encoder, and acquiring a second vector representation corresponding to the speech signal by means of a speaker encoder; by means of a text encoder, encoding first n-1 words that have been output by a speech recognition decoder, to obtain a third vector representation corresponding to an (n-1)-th word; inputting the first vector representation, the second vector representation, and the third vector representation into a speaker decoder to obtain a speaker vector representation corresponding to an n-th word; according to correlation coefficients between the speaker vector representation corresponding to the n-th word and portrait feature vectors of a plurality of speakers, determining a speaker corresponding to the n-th word; and inputting the first vector representation, the first n-1 words, and the weighted sum of the correlation coefficients on the portrait feature vectors of the plurality of speakers into the speech recognition decoder to obtain the n-th word.

浙江省杭州市滨江区长河街道网商路699号4
号楼5层508室, Zhejiang 310052 (CN)。

(74) 代理人:北京博浩百睿知识产权代理有限公司
(KSK IP AGENCY); 中国北京市海淀区知春路甲48
号盈都大厦C座1单元10E, Beijing 100098 (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,
MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, 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))。

(57) 摘要: 本公开提供一种语音识别方法、设备和存储介质, 方法包括: 通过语音识别编码器获取语音信号对应的第一向量表示, 通过说话者编码器获取语音信号对应的第二向量表示; 通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理, 得到第 $n-1$ 个字所对应的第三向量表示; 将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中, 获得第 n 个字对应的说话者向量表示; 根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数, 确定第 n 个字对应的说话者; 将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和, 输入语音识别解码器, 获得第 n 个字。

语音识别方法、设备和存储介质

交叉援引

本公开要求于 2023 年 06 月 08 日提交中国专利局、优先权号为 202310680032.X、发明名称为“语音识别方法、设备和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本发明涉及人工智能技术领域，尤其涉及一种语音识别方法、设备和存储介质。

背景技术

在很多场景中都会遇到多个说话者的语音识别问题，例如，在会议场景下，说话的人数可能为两人或者多人，需要同时对多个说话者的语音进行识别，以确定每个说话者说了什么内容。当前语音识别技术已经可以以较高精度识别一个人所讲的内容，但是当说话的人数为两人或者多人时，语音识别率就会极大的降低。

可以采用端到端说话人相关语音识别模型对多个说话者的语音进行识别。但是，现有的端到端说话人相关语音识别模型并不能获得准确性高的语音识别结果。

发明内容

本发明实施例提供一种语音识别方法、设备和存储介质，用以提高多个说话者的语音识别结果的准确性。

第一方面，本发明实施例提供一种语音识别方法，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器，所述方法包括：

获取多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音；

通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别；

通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；

根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者；

将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权和，输入所述语音识别解码器，以获得所述第 n 个字。

第二方面，本发明实施例提供一种语音识别装置，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器，所述装置包括：

获取模块，被设置为获取多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音；

第一编码模块，被设置为通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别；

第二编码模块，被设置为通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

第一解码模块，被设置为将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者；

第二解码模块，被设置为将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权求和，输入所述语音识别解码器，以获得所述第 n 个字。

第三方面，本发明实施例提供一种电子设备，包括：存储器、处理器、通信接口；其中，所述存储器上存储有可执行代码，当所述可执行代码被所述处理器执行时，使所述处理器执行如第一方面所述的语音识别方法。

第四方面，本发明实施例提供了一种非暂时性机器可读存储介质，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器至少可以实现如第一方面所述的语音识别方法。

第五方面，本发明实施例提供一种语音识别方法，该方法包括：

接收终端设备通过调用语音识别服务触发的请求，所述请求中包括多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音；

根据所述模型训练服务对应的计算资源执行如下步骤：

获取语音识别模型，所述语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器；

通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别；

通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；

根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者；

将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权和，输入所述语音识别解码器，以获得所述第 n 个字；

向所述终端设备发送语音识别输出信息，所述语音识别输出信息中包括所述多个说话者各自对应的文字序列。

在上述本发明实施例提供的语音识别方案中，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器。在对多个说话者的语音信号进行语音识别时，可以先通过语音识别模型中的语音识别编码器对该语音信号进行编码处理，以获取被设置为进行语音识别的第一向量表示，并通过说话者编码器对该语音信号进行编码处理，以获取被设置为进行说话者识别的第二向量表示。而后，通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示。之后，将第一向量表示、第二向量表示以及第三向量表示输入说话者解码器中，以获得第 n 个字对应的说话者向量表示，并根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者。最后，将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和，输入语音识别解码器中，以获得第 n 个字。

在上述方案中，在语音识别模型中添加包含注意力层的文本编码器，通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，可以更好地聚合整个句子的上下文信息，以获得第 $n-1$ 个字所对应的包含上下文信息的第三向量表示，即这样可以获得更准确地被设置为进行说话者识别的向量表示，从而提高了各个字对应的说话者识别结果的准确性。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为本发明实施例提供的一种语音识别方法的流程图；

图 2 为本发明实施例提供的一种语音识别模型的结构示意图；

图 3 为本发明实施例提供的语音识别编码器和语音识别解码器的结构示意图；

图 4 为一种通过说话者解码器确定第 n 个字对应的说话者向量表示的流程图；

图 5 为本发明实施例提供的另一种语音识别方法的流程图；

图 6 为一种通过语音识别解码器获得第 n 个字的流程图；

图 7 为本发明实施例提供的一种语音识别方法的应用示意图；

图 8 为本发明实施例提供的另一种语音识别方法的流程图；

图 9 为本发明实施例提供的一种云端服务模式下的语音识别过程的示意图；

图 10 为本发明实施例提供的一种语音识别装置的结构示意图；

图 11 为本实施例提供的一种电子设备的结构示意图。

具体实施方式

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。但是本公开能够以很多不同于在此描述的其它方式来实现，本领域技术人员可以在不违背本公开内涵的情况下做类似推广，因此本公开不受下面公开的具体实施的限制。

为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。另外，下述各方法实施例中的步骤时序为一种举例，而非严格限定。

需要说明的是，本发明实施例中所涉及的用户信息（包括但不限于用户设备信息、用户个人信息等）和数据（包括但不限于用于分析的数据、存储的数据、展示的数据等），均为经用户授权或者经过各方充分授权的信息和数据，并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准，并提供有相应的操作入口，供用户选择授权或者拒绝。

下面先对本发明实施例中涉及到的一些概念进行简单介绍。

说话人相关语音识别（Speaker-Attributed Automatic Speech Recognition，简称 SA-ASR）：是一种用于解决“谁说了什么内容”的语音识别任务。

序列化输出训练（Serialized Output Training，简称 SOT）：在对语音进行转录（即语音转为文字）时，采用特殊的分隔符号将不同说话者所讲的文字连接起来，并且按照不同说话者开始说话的时间顺序对转录文字进行排列，以生成文字序列，并输出文字序列。例如，在一段语音中包括两个说话者，说话者 A 说了“今天天气怎么样”，说话者 B 说了“今天天气不错”。那么在对该段语音进行序列化输出训练时，将输出：今天天气怎么样<sc>今天天气不错，并且每个字或者说每句话关联对应的说话者标签。

说话者画像（Speaker Profile）：是指一组特征向量，其包含有多个说话者的画像特征向量。可以从各个说话者的一段语音信号中提取出各个说话者的画像特征向量。

端到端（End to End，简称 E2E）：区别于对目标任务进行模块化处理，而由一个完整的模型对目标任务进行处理，即输入是原始数据，输出是最后的处理结果。在学习训练时可以直接优化整体的目标函数。

现有的端到端说话人相关语音识别模型中主要包括语音识别编码器、说话者编码

器、语音识别解码器、说话者解码器，该模型中缺乏对上下文信息感知模块，在通过这些编码器和解码器对多个说话者的语音进行识别时，没有充分考虑各个字所对应的上下文信息，使得预测出的各个字对应的说话者的结果不准确，从而影响了最终的语音识别结果，为了解决多说话者的语音识别结果不准确的问题，本发明实施例提供了一种新的语音识别方案。在该语音识别方案中，在语音识别模型中添加包含注意力层的文本编码器，在确定第 n 个字所对应的说话者向量表示时，通过该文本编码器可以对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以整合句子的上下文信息，使得第 n 个字所对应的说话者向量表示更精准，从而使得说话者编码器可以更准确地预测出各个字对应的说话者。

下面结合附图，对本发明的一些实施方式作详细说明。在各实施例之间不冲突的情况下，下述的实施例及实施例中的特征可以相互组合。

图 1 为本发明实施例提供的一种语音识别方法的流程图，如图 1 所示，该方法包括如下步骤：

101、获取多个说话者的语音信号和多个说话者的画像特征向量，语音信号中包含多个说话者的语音。

102、通过语音识别编码器获取语音信号对应的第一向量表示，通过说话者编码器获取语音信号对应的第二向量表示，第一向量表示被设置为进行语音识别，第二向量表示被设置为进行说话者识别。

103、通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示。

104、将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中，以获得第 n 个字对应的说话者向量表示。

105、根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者。

106、将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和，输入语音识别解码器，以获得第 n 个字。

本发明实施例提供的语音识别方案可以利用预先训练好的语音识别模型对待识别的多个说话者的语音信号进行处理，以获得与该语音信号相对应的各个字以及各个字所对应的说话者。由于语音识别模型的训练过程与使用语音识别模型对待识别的多个说话者的语音信号进行识别的过程有相似之处，这里描述使用语音识别模型的过程。

其中，语音识别模型的结构可以参考附图 2 所示。语音识别模型中主要包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器。语音识别编码器主要被设置为将多个说话者的语音信号转换为被设置为进行语音识别的第一向量表示。说话者编码器主要被设置为将多个说话者的语音信号转换为被设置为进行说话者识别的第二向量表示。语音识别解码器主要被设置为完成语

音到文字的识别，以识别出语音信号中各个说话者所说的全部文字，最终输出各个说话者对应的语音识别结果。说话者解码器主要被设置为确定各个文字与说话者的对应关系。文本编码器主要被设置为整合文字的上下文信息，根据各个字所对应的上下文信息确定各个字所对应的说话者向量表示，以使得获得的各个字对应的说话者向量表示更加准确。

在一可选实施例中，语音识别编码器和语音识别解码器的具体结构可以参考附图3所示。语音识别编码器中可以包括级联的多个编码器，每个编码器中可以包括两个子层：注意力层和前馈神经网络层。语音识别解码器中可以包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层。同样地，说话编码器与语音识别编码器具体相似的结构，可以参考图3所示的具体结构。其中，说话者编码器中可以包括级联的多个编码器，并且每个编码器中包括注意力层和前馈神经网络层。同样地，说话解码器与语音识别解码器具体相似的结构，说话者解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层。文本编码器中包括级联的至少一个编码器，每个编码器中包括注意力层和前馈神经网络层。其中，语音识别编码器、说话者编码器中所包含的编码器的数量可以根据实际需求进行设定，在此不做限定。同样地，语音识别解码器、说话者解码器中所包含的解码器的数量可以根据实际需求进行设定。另外，这里的解码器中的注意力层可以包括自注意力层和源-目标注意力层。

由上述描述可以获知语音识别模型主要包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器。那么，利用语音识别模型对多个说话者的语音信号进行语音识别时，语音识别模型中的各个编码器和解码器的具体执行过程如下：首先，获取多个说话者的语音信号和多个说话者的画像特征向量。其中，语音信号中包含多个说话者的语音。说话者的画像特征向量是指说话者的特征向量，可以从该说话者的一段语音信号中提取出该说话者对应的画像特征向量。

在对多个说话者的语音信号进行处理之前，可以预先对多个说话者的画像特征进行提取，以获得各个说话者的画像特征向量。具体地，可以通过一个与说话者编码器结构相同的编码器对各个说话者的语音信号进行编码处理，以获得各个说话者的画像特征向量。

在获取到包含多个说话者的语音信号之后，可以对该语音信号进行处理，以提取有效的语音信息，获得该语音信号对应的声学特征序列。其具体实现过程可以包括：对获取到的包含多个说话者的语音信号进行分割处理，将语音信号切成一帧一帧，然后对每帧语音信号提取声学特征，并根据各帧语音信号对应的声学特征，生成该语音信号对应的声学特征序列。

接着，将语音信号对应的声学特征序列输入至语音识别编码器中，以通过语音识别编码器获取该语音信号对应的第一向量表示。其中，第一向量表示被设置为进行语音识别，也就是说该第一向量表示可以表征该语音信号的语音特征。并将语音信号对

应的声学特征序列输入至说话者编码器中，以通过说话者编码器获取该语音信号对应的第二向量表示。其中，第二向量表示被设置为进行说话者识别，也就是说该第二向量表示可以被设置为表征该语音信号的说话者的音色等特征。

在本实施例中，语音识别编码器可以将语音信号对应的声学特征序列转换为可以进行语音识别的向量表示，以便于后续直接基于第一向量表示进行语音识别处理，以获得该语音信号所对应的各个字。说话者编码器可以将语音信号对应的声学特征序列转换为进行说话者识别的向量表示，以便于后续直接基于第二向量表示进行各个字所对应的说话者识别处理。

对语音信号进行识别转录处理时，采用逐个字逐个字的形式进行预测，例如语音识别解码器确定出该语音信号对应的第1个字之后，第1个字自回归到语音识别解码器的输入端，被设置为确定该语音信号对应的第2个字，依次确定该语音信号对应的各个字，直到遇到结束符为止。由于对语音信号中的各个字进行预测时，其具体实现过程基本一致，那么这里以第 n 个字的预测过程进行举例说明，其中，假设语音识别解码器已经输出了前 $n-1$ 个字。需要说明的是，识别出的第1个字是基于设定的起始字符或者说空字符作为语音识别解码器的输入来预测出的。

在确定第 n 个字时，首先要获得第 $n-1$ 个字所对应的包含上下文信息的第三向量表示，即结合前 $n-1$ 个字的信息，确定第 n 个字，这样确定出第 n 个字更准确。由于文本编码器主要是被设置为整合各个字所对应的上下文信息，以获得包含上下文信息的向量表示，那么在获取到语音识别解码器输出的前 $n-1$ 个字之后，通过文本编码器对语音识别解码器已经输出的这前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示。

需要注意的是：文本编码器在确定第 $n-1$ 个字所对应的第三向量表示时，首先获取语音识别编码器经过各种处理之后最终输出的前 $n-1$ 个字，并对这已经确定出的前 $n-1$ 个字进行编码处理，以从编码结果中获得第 $n-1$ 个字所对应的编码结果作为该第 $n-1$ 个字对应的第三向量表示。例如，当要获得第4个字所对应的第三向量表示时，首先获取语音识别编码器已经输出的前3个字，对这前3个字进行编码处理，获得第4个字对应的第三向量表示。

在一可选实施例中，通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示的具体实现过程可以包括：获取语音识别解码器中的自注意力层分别对前 $n-1$ 个字进行注意力计算后得到的前 $n-1$ 个字分别对应的注意力系数加权后的向量表示；并将前 $n-1$ 个字分别对应的注意力系数加权后的向量表示输入文本编码器中，以获得文本编码器输出的第 $n-1$ 个字所对应的第三向量表示。也就是说，输入到文本编辑器的是语音识别解码器中自注意力层输出的前 $n-1$ 个字各自对应的加权后编码向量。

如上文所述，语音识别解码器中可以包括至少一个解码器，每个解码器中都可以

包括注意力层,而且具有可以包括几种不同的注意力层,比如自注意力层和源-目标注意力层。为了简化计算复杂度以及保留每个字的浅层次的语义信息,可选地,上述语音识别解码器中的自注意力层可以是语音识别解码器的第一层即第一个解码器中包含的自注意力层。此时,就是要收集语音识别编码器的第一层中自注意力层针对前 $n-1$ 个字分别输出的结果,形成一个序列,送入文本编码器中。实际上,文本编码器会输出前 $n-1$ 个字每个字对应的融合上下文信息后的第三向量表示,这里只是从中这前 $n-1$ 个字对应的第三向量表示中取出第 $n-1$ 个字对应的第三向量表示。之所以文本编码器可以进行文字上下文信息的融合是因为文本编码器中包括注意力层,比如可以是采用自注意力机制或其他注意力机制的注意力层,本文中自注意力层为例。

可选地,针对语音识别解码器来说,第一层中的自注意力层分别对前 $n-1$ 个字进行注意力计算可以表示为如下公式: $\overline{Z}_{n-1,1}^{asr} = Z_{n-1,1}^{asr} + MHA_{[1]}^{asr-self}(Z_{n-1,1}^{asr}, Z_{[1:n-1],1}^{asr}, Z_{[1:n-1],1}^{asr})$, 其中, $\overline{Z}_{n-1,1}^{asr}$ 表示语音识别解码器的第一层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示, $Z_{n-1,1}^{asr}$ 表示输入语音识别解码器的第一层中的第 $n-1$ 个字对应的嵌入向量表示, $MHA_{[1]}^{asr-self}$ 表示语音识别解码器的第一层中自注意力层所执行的多头注意力计算, $Z_{[1:n-1],1}^{asr}$ 表示输入语音识别解码器的第一层中的前 $n-1$ 个字分别对应的嵌入向量表示。其中,嵌入向量表示被设置为表示嵌入编码处理后的向量表示,并且该嵌入向量表示可以通过对前 $n-1$ 个字分别进行嵌入编码或者嵌入编码和位置编码,以得到前 $n-1$ 个字各自对应的嵌入向量表示。具体的,可以采用公式

$Z_{[1:n-1],1}^{asr} = PosEnc(Embed(y_{[1:n-1]}))$, 获得前 $n-1$ 个字各自对应的嵌入向量表示。其中, $PosEnc$ 表示位置编码, $Embed$ 表示嵌入编码, $y_{[1:n-1]}$ 表示语音识别解码器输出的前 $n-1$ 个字。

在获得前 $n-1$ 个字分别对应的注意力系数加权后的向量表示之后,文本编辑器对前 $n-1$ 个字分别对应的注意力系数加权后的向量表示进行处理,以获得第 $n-1$ 个字所对应的第三向量表示。在一可选实施例中,文本编辑器可以根据公式 $\overline{Z}_{n-1}^{context} = Context-Enc(\overline{Z}_{[1:n-1],1}^{asr})$, 计算第 $n-1$ 个字所对应的第三向量表示。其中, $\overline{Z}_{n-1}^{context}$ 表示第 $n-1$ 个字所对应的第三向量表示, $Context-Enc$ 表示上下文文本编码, $\overline{Z}_{[1:n-1],1}^{asr}$ 表示前 $n-1$ 个字分别对应的注意力系数加权后的向量。

在获得第三向量表示之后,将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中,以获得第 n 个字对应的说话者向量表示。其中,说话者向量表示可以用于表征说话者的特征。在确定的第 n 个字对应的说话者向量表示时结合了语音信号的语音识别特征、说话者识别特征、以及上下文信息特征,这样确定出的第 n 个字对应的说话者向量表示质量更高,从而可以提高各个字对应的预测说话者的准确性,

使得语音信号所对应的转录结果更准确。

通过说话者解码器确定出第 n 个字对应的说话者向量表示之后，接着根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者。其中，相关性系数被设置为表征各个字对应的说话者向量表示与多个说话者的画像特征向量之间的相关性，也就是说该相关性系数被设置为表示第 n 个字与语音信号中包含的各个说话者之间的相关性。那么，在确定第 n 个字对应的说话者时，首先获取第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，然后根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者。

例如，假设第一段语音信号中包含 3 个说话者，第 n 个字对应的说话向量表示为 q_n ，第一个说话者的画像特征向量为 d_1 ，第二个说话者的画像特征向量为 d_2 ，第三个说话者的画像特征向量为 d_3 ，获取第 n 个字对应的说话者向量表示 q_n 与 d_1 之间的相关性系数 $\beta_{n,1}$ ，获取第 n 个字对应的说话者向量表示 q_n 与 d_2 之间的相关性系数 $\beta_{n,2}$ ，获取第 n 个字对应的说话者向量表示 q_n 与 d_3 之间的相关性系数 $\beta_{n,3}$ ，然后根据这三个相关性系数，确定第 n 个字对应的说话者。在一可选实施例中，可以从多个相关性系数中选取最大相关性系数对应的说话者确定为第 n 个字对应的说话者。

然后，计算多个相关性系数对多个说话者的画像特征向量的加权求和，其中加权求和可以看作是加权后的第 n 个字对应的说话者向量表示。可选地，可以根据公式 $\bar{d}_n = \sum_{k=1}^K \beta_{n,k} d_k$ 计算多个相关性系数对多个说话者的画像特征向量的加权求和 $\bar{d}_n$ ，其中， $\bar{d}_n$ 表示加权求和， K 表示语音信号中包含的说话者总数， $\beta_{n,k}$ 表示第 n 个字所对应的说话者向量与第 k 个说话者的画像特征向量之间的相关系数， d_k 表示第 K 个说话者的画像特征向量。

最后，将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权求和，输入语音识别解码器，以获得第 n 个字。其中，加权求和为第 n 个字对应的加权后的说话者向量表示，在确定第 n 个字时引入第 n 个字对应的加权后的说话者向量表示，这样获得的第 n 个字不包括该字的具体内容还包括该字所对应的说话者。例如输出的第 n 个字的具体内容是“我”，对应的说话者是“小李”。

依据上述方法可以依次确定语音信号中各个字以及每个字对应的说话者，最终可以确定出每个说话者对应的 SOT 格式的识别结果。例如，在一段语音信号中识别出包括两个说话者，且识别出的文字“今天天气怎么样”所对应的说话者为说话者 a，识别出的文字“今天天气不错”所对应的说话者为说话者 b，输出的内容为：今天天气怎么样<sc>今天天气不错。

在本发明实施例中，在语音识别模型中添加包含注意力层的文本编码器，通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，可以更好地聚合整个句子的上下文信息，以获得第 $n-1$ 个字所对应的包含上下文信息的第三向量表示，

即这样基于第一向量表示、第二向量表示以及第三向量表示，可以获得更准确地被设置为进行说话者识别的说话者向量表示，从而提高了各个字对应的说话者识别结果的准确性。

在实际应用中，一个说话者解码器中包括级联的多个解码器。为了便于理解语音识别模型中的说话者解码器中各个层所对应的编码器的具体处理过程，结合图4对各个层所对应的编码器的处理过程进行示例性说明。

图4为一种通过说话者解码器确定第 n 个字对应的说话者向量表示的流程图；如图4所示，说话者解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层，该方法包括如下步骤：

401、将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中的第一层解码器中，以获得第一层解码器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示。

402、将第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示输入第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的第 $n-1$ 个字对应的说话者向量表示。

403、确定说话者解码器中最后一层解码器输出的第 $n-1$ 个字对应的说话者向量表示。

404、根据最后一层解码器输出的第 $n-1$ 个字对应的说话者向量表示和第一层解码器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示，确定第 n 个字对应的说话者向量表示。

本发明实施例中的说话者解码器中包括多个解码器，各个解码器分别设置于不同的层级，各个解码器按照层级逐层进行解码处理，并且上一层解码器的输出作为下一层解码器的输入。另外，各个解码器均包括注意力层和前馈神经网络层，分别由注意力层和前馈神经网络层两个子层进行解码处理，首先将输入到解码器中的向量表示传输至注意力层中进行处理，并将处理结果输入至前馈神经网络层，前馈神经网络层对注意力层输出的处理结果进行处理，并将处理后的结果输入至下一层的解码器中的注意力层中，由下一层的解码器中的注意力层进行处理，即说话者解码器中的各个解码器均按照该方法依次进行处理。

具体的，各个解码器中的处理过程可以包括：首先将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中，以通过第一层解码器中的注意力层，获得第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示。然后，将第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示输入至第一层解码器中的前馈神经网络层，以通过第一层解码器中的前馈神经网络层获得输入到第二层解码器的第 $n-1$ 个字对应的说话者向量表示。也就是说，第一层解码器中的前馈神经网络层输出结果（第 $n-1$ 个字对应的说话者向量表示）作为第二层解码器中的注意力层的输入。

在一可选实施例中，具体的，第一层解码器中的注意力层可以根据公式

$\overline{Z}_{n-1,1}^{spk} = MHA_{n-1,1}^{spk-src}(\overline{Z}_{n-1}^{context}, H^{asr}, H^{spk})$ ，计算第 n-1 个字对应的注意力系数加权后的说话者向量表示。其中， $\overline{Z}_{n-1,1}^{spk}$ 表示第一层解码器注意力层输出的第 n-1 个字对应的注意力系数加权后的说话者向量表示， $MHA_{n-1,1}^{spk-src}$ 表示第一层编码器的注意力层的多头注意力， H^{asr} 表示第一向量表示， H^{spk} 表示第二向量表示。在计算出第 n-1 个字对应的注意力系数加权后的说话者向量表示之后，第一层解码器中的前馈神经网络层可以根据公式 $Z_{n-1,2}^{spk} = \overline{Z}_{n-1,1}^{spk} + FF_1^{spk}(\overline{Z}_{n-1,1}^{spk})$ ，计算第 n-1 个字对应的说话者向量表示。其中， $Z_{n-1,2}^{spk}$ 表示输入到第二层解码器的第 n-1 个字对应的说话者向量表示， $\overline{Z}_{n-1,1}^{spk}$ 表示第一层解码器注意力层输出的第 n-1 个字对应的注意力系数加权后的说话者向量表示， FF_1^{spk} 表示说话者编码器中的第一层前馈神经网络。

在一可选实施例中，第二层解码器以及第二层以上的解码器中的注意力层包括自注意力层和源-目标注意力层。那么，在通过第一层解码器中的前馈神经网络层获得输入到第二层解码器的第 n-1 个字对应的说话者向量表示之后，将其输出至第二层解码器中的自注意力层，第二层解码器中的自注意力层对第 n-1 个字对应的说话者向量表示、前 n-1 个字对应的说话者向量表示进行自注意力计算，获得第二层解码器中的自注意力层对应的第 n-1 个字对应的自注意力系数加权后的向量表示，并将第 n-1 个字对应的自注意力系数加权后的向量表示输入至第二层解码器中的源-目标注意力层。第二层解码器中的源-目标注意力层对第 n-1 个字对应的自注意力系数加权后的向量表示和第二向量表示进行注意力计算，以获得第二层解码器中的源-目标注意力层对应的第 n-1 个字对应的注意力系数加权后的向量表示，并将第 n-1 个字对应的注意力系数加权后的向量表示输入至第二层解码器中的前馈神经网络层。第二层解码器中的前馈神经网络层对第 n-1 个字对应的注意力系数加权后的向量表示进行处理，以获得输入至第三层编码器的第 n-1 个字对应的说话者向量表示。

第二层以上解码器按照上述方法分别进行计算处理，直至确定出说话者解码器中最后一层解码器输出的所述第 n-1 个字对应的说话者向量表示，并根据最后一层解码器输出的第 n-1 个字对应的说话者向量表示和第一层解码器中的注意力层输出的第 n-1 个字对应的注意力系数加权后的说话者向量表示，确定第 n 个字对应的说话者向量表示，其详细处理过程参考上述描述在此不再赘述。

在一可选实施例中，第二层解码器以及第二层以上的解码器中的自注意力层可以根据公式 $\overline{Z}_{n-1,l}^{spk} = Z_{n-1,l}^{spk} + MHA_l^{spk-self}(Z_{n-1,l}^{spk}, Z_{[1:n-1],l}^{spk}, Z_{[1:n-1],l}^{spk})$ ，计算输入至该层解码器中的源-目标注意力层的第 n-1 个字对应的自注意力系数加权后的向量表示。其中， l 表示编码器所对应的层数，且 $l > 1$ ， $\overline{Z}_{n-1,l}^{spk}$ 表示第 l 层编码器中的自注意力层输出的第 n-1

个字对应的自注意力系数加权后的向量表示（即输入至第 l 层编码器中的源-目标注意力层的第 $n-1$ 个字对应的自注意力系数加权后的向量表示）， $\bar{Z}_{n-1,l}^{spk}$ 表示第 $l-1$ 层编码器中的前馈神经网络层输出的第 $n-1$ 个字对应的说话者向量表示， $MHA_l^{spk-self}$ 表示第 l 层编码器中的自注意力层的多头注意力， $\bar{Z}_{[1:n-1],l}^{spk}$ 表示第 $l-1$ 层编码器中的前馈神经网络层输出的前 $n-1$ 个字对应的说话者向量表示。

在一可选实施例中，第二层解码器以及第二层以上的解码器中的源-目标注意力层可以根据公式 $\bar{\bar{Z}}_{n-1,l}^{spk} = \bar{Z}_{n-1,l}^{spk} + MHA_l^{spk-src}(\bar{Z}_{n-1,l}^{spk}, H^{spk}, H^{spk})$ ，计算输入至该层解码器中的前馈神经网络层的第 $n-1$ 个字对应的注意力系数加权后的向量表示。其中， l 表示编码器所对应的层数，且 $l > 1$ ， $\bar{\bar{Z}}_{n-1,l}^{spk}$ 表示第 l 层编码器中的源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示（即输入至第 l 层编码器中的前馈神经网络层的第 $n-1$ 个字对应的注意力系数加权后的向量表示）， $\bar{Z}_{n-1,l}^{spk}$ 表示第 l 层编码器中的自注意力层输出的第 $n-1$ 个字对应的自注意力系数加权后的向量表示， $MHA_l^{spk-src}$ 表示第 l 层编码器中的源-目标注意力层的多头注意力， H^{spk} 表示第二向量表示。

在一可选实施例中，第二层解码器以及第二层以上的解码器中的前馈神经网络层可以根据公式 $\bar{Z}_{n-1,l+1}^{spk} = \bar{\bar{Z}}_{n-1,l}^{spk} + FF_l^{spk}(\bar{\bar{Z}}_{n-1,l}^{spk})$ ，计算输入至下一层编码器的第 $n-1$ 个字对应的说话者向量表示。其中， l 表示编码器所对应的层数，且 $l > 1$ ， $\bar{Z}_{n-1,l+1}^{spk}$ 表示第 l 层编码器中的前馈神经网络输出的第 $n-1$ 个字对应的说话者向量表示（即输入至第 $l+1$ 层编码器中的第 $n-1$ 个字对应的说话者向量表示）， $\bar{\bar{Z}}_{n-1,l}^{spk}$ 表示第 l 层编码器中的源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示， FF_l^{spk} 表示第 l 层编码器中的前馈神经网络。

依次确定各个层解码器中的前馈神经网络层输出，直到确定出说话者解码器中最后一层解码器输出的第 $n-1$ 个字对应的说话者向量表示。最后，说话者解码器对最后一层解码器输出的第 $n-1$ 个字对应的说话者向量表示和第一层解码器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示进行跳接处理

(skip-connect)，根据跳接结果来预测第 n 个字对应的说话者向量表示。具体的，在一可选实施例中，假设说话者解码器中最后一层解码器输出的第 $n-1$ 个字对应的说话者向量表示为 $\bar{Z}_{n-1,L}^{spk}$ ，第一层解码器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示为 $\bar{\bar{Z}}_{n-1,1}^{spk}$ ，根据公式 $q_n = W^q \cdot (\bar{Z}_{n-1,L}^{spk} + \bar{\bar{Z}}_{n-1,1}^{spk})$ ，确定第 n 个字对应的说话者向量表示 q_n 。其中， W^q 表示说话者编码器中设定的权重。

本发明实施例，说话者解码器中包含多层解码器，第一层解码器通过对第一向量表示、第二向量表示和包含上下文信息的第三向量表示进行处理，以获得第一层解码

器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示，即在第一层解码器在进行处理时结合了上下文信息，使得获得第 $n-1$ 个字对应的注意力系数加权后的说话者表示更准确。然后，第一层解码器中的前馈神经网络层对结合上下文信息处理后获得的第 $n-1$ 个字对应的注意力系数加权后的说话者向量进行处理，这样输入到第二层解码器的第 $n-1$ 个字对应的说话者向量表示具有较高的质量，可以更好地反映第 $n-1$ 个字对应的说话者特征，再依次通过多层解码器的处理可以获得的更高质量的各个字对应的说话者向量表示，从而使得各个字所预测的说话者结果的准确性。

上述实施例介绍了确定各个字对应的说话者向量表示的实施方式，然而在实际应用中，在确定出了各个字对应的说话者向量表示之后，可以进一步根据各个字对应的说话者向量表示确定各个字对应的说话者。具体的，在一可选实施例中，根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者。那么在确定第 n 个字对应的说话者时，首先确定第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数。在下述实施例中，结合附图 5 对确定第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数的具体过程进行示例性说明。

图 5 为本发明实施例提供的另一种语音识别方法的流程图；如图 5 所示，为了提高该方法语音识别结果的准确性，在上述实施例的基础上，该方法还可以包括如下步骤：

501、针对多个说话者中任一说话者的画像特征向量，根据设定的相似度算法，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度。

502、获取前 n 个字各自对应的说话者向量表示。

503、分别将前 n 个字各自对应的说话者向量表示与任一说话者的画像特征向量进行拼接。

504、将多个拼接后的向量表示输入设定的包含注意力模块的打分模型，以获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度。

505、根据第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度以及第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第二相似度，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数。

在本发明实施例中，采用设定的相似度算法和设定的包含注意力模块的打分模型联合打分的方式确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数。其中，相关性系数可以被设置为表示第 n 个字对应各个说话者的后验概率。

具体的，针对多个说话者中任一说话者的画像特征向量，根据设定的相似度算法，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度。也就是说，根据设定的相似度算法，可以分别确定出第 n 个字对应的说话者向量表示与各个说话者的画像特征向量之间的第一相似度。其中，可选地，设定的相似度算法可以为余弦相似度算法，例如，根据公式 $score_{n,k}^{CI} = \cos(q_n, d_k)$ ，可以确定第 n 个字对应的说话者向量表示与各个说话者的画像特征向量之间的第一相似度。其中， $score_{n,k}^{CI}$ 表示第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的第一相似度， $\cos$ 表示余弦函数， q_n 表示第 n 个字对应的说话者向量表示， d_k 表示第 k 个说话者的画像特征向量， k 是语音信号中包含的 K 个说话者中的第 k 个说话者。也就是说，第 n 个字对应的说话者向量表示分别与各个说话者的画像特征向量进行余弦相似度分数计算，得到第 n 个字分别对应于各个说话者的概率。

由上述描述可知：第一相似度的确定只考虑了第 n 个字对应的说话者向量表示，未考虑前 n 个字对应的说话者向量表示，因此，第一相似度的确定是其它说话者无关的。

为了提高确定出的各个字对应各个说话者的后验概率的准确度，在确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数时，充分结合了其他说话者的说话者向量表示。具体的，通过在语音识别模型中添加设定的包含注意力模块的打分模型，以通过设定的包含注意力模块的打分模型，获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度。

其中，确定第二相似度的具体实现过程可以为：获取前 n 个字各自对应的说话者向量表示，分别将前 n 个字各自对应的说话者向量表示与任一说话者的画像特征向量进行拼接。将多个拼接后的向量表示输入设定的包含注意力模块的打分模型，以获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度。

具体的，在一可选实施例中，可以根据公式 $score_{n,k}^{CD} = \tanh(CD\text{-}scorer(q_{[1:n]}, d_k))$ ，确定前 n 个字各自对应的说话者向量表示分别与各个说话者的画像特征向量之间的第二相似度。其中， $score_{n,k}^{CD}$ 表示第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的第二相似度， $\tanh$ 为双曲正切函数， $CD\text{-}scorer$ 表示设定的包含注意力模块的打分模型。 $q_{[1:n]}$ 表示前 n 个字对应的说话者向量表示， d_k 表示第 k 个说话者的画像特征向量。可以按照该公式分别生成第 n 个字对应的说话者向量表示与各个说话者的画像特征向量之间的第二相似度。由该公式可知，确定第二相似度的实现方式可以为：获取前 n 个字中每个字对应的说话者向量表示，将前 n 个字中每个字对应的说话者向量表示分别与第 k 个说话人特征向量拼接，得到第一向量序列，长度为 n 。并将第一向量序列输入至打分模型（比如采用 transformer 模型的结

构)中,获得第二向量序列。其中第二向量序列是经过 tanh 变换后得到的取值在[-1,1]的一个序列,取出其中最后一个确定为第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的第二相似度。Tanh 为一种映射函数的举例,不以此为限。

在确定出第一相似度和第二相似度之后,根据第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度以及第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第二相似度,确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数。也就是说,对于第 n 个字来说,针对每个说话者均计算第一相似度和第二相似度,将第 n 个字与每个说话者之间的第一相似度和第二相似度相加,获得第一和值,并将第一和值除以第 n 个字与各个说话者之间第一相似度和第二相似度总和,获得第 n 个字对应的说话者向量表示与每个说话者的画像特征向量之间的相关性系数。

具体的,在一可选实施例中,可以根据公式 $\beta_{n,k} = \frac{\exp(score_{n,k}^{CI} + score_{n,k}^{CD})}{\sum_j^K \exp(score_{n,j}^{CI} + score_{n,j}^{CD})}$, 确定第 n 个字对应的说话者向量表示与各个说话者的画像特征向量之间的相关性系数。其中, $\beta_{n,k}$ 表示第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的相关性系数, $\exp$ 表示以 e 为底的指数函数, $score_{n,k}^{CI}$ 表示第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的第一相似度, $score_{n,k}^{CD}$ 表示第 n 个字对应的说话者向量表示与第 k 个说话者的画像特征向量之间的第二相似度, K 为语音信号中所包含的说话者总数, j 为第 j 个说话者。另外,上述中的 k 可以为 K 个说话者中的任意一个,那么根据上述公式可以获得第 n 个字对应的说话者向量表示与各个说话者的画像特征向量之间的相关性系数。

在确定出第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数之后,可以根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数,确定第 n 个字对应的说话者。

本发明实施例中,通过根据设定的相似度算法,确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度,通过设定的包含注意力模块的打分模型,获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度,再根据第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度以及第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第二相似度,确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数,即采用设定的相似度算法和设定的包含注意力模块的打分模型结合的方式,确定出的相关性系数更准确,使得语音识别模型具有更好的性能,从而可以进一步提高语音识别结果的准确性。

上述实施例介绍了确定第 n 个字对应的说话者向量表示分别与所述多个说话者的

画像特征向量之间的相关性系数的具体实现方式，在确定出第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数之后，计算相关性系数对多个说话者的画像特征向量的加权和。具体的，获取多个相关性系数分别与各个说话者的画像特征向量的乘积值，然后对多个乘积值进行求和运算，以获得第 n 个字对应的加权后的说话人特征向量。

在获取到相关性系数对多个说话者的画像特征向量的加权和之后，将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和，输入语音识别解码器，以获得第 n 个字。在实际应用中，一个语音识别解码器中包括级联的多个解码器。为了便于理解语音识别模型中的语音识别解码器中各个层所对应的编码器对第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和的具体处理过程，结合图 6 对各个层所对应的编码器的处理过程进行示例性说明。

图 6 为一种通过语音识别解码器获得第 n 个字的流程图；如图 6 所示，语音识别解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层，该方法包括如下步骤：

601、对前 $n-1$ 个字分别进行嵌入编码，得到前 $n-1$ 个字各自对应的嵌入向量表示。

602、将前 $n-1$ 个字各自对应的嵌入向量表示和第一向量表示输入语音识别解码器的第一层解码器中，以获得第一层解码器中的注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示。

603、将第 $n-1$ 个字对应的注意力系数加权后的向量表示和加权和输入第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的第 $n-1$ 个字对应的向量表示。

604、根据语音识别解码器中最后一层解码器输出的第 $n-1$ 个字对应的向量表示，确定第 n 个字。

本发明实施例中的语音识别解码器中包括多层解码器，各层解码器按照层级逐层进行解码处理，并且上一层解码器的输出作为下一层解码器的输入。其中，每个解码器中包括注意力层和前馈神经网络层，由两个子层分别对输入至编码器中的向量表示进行处理。语音解码器的作用是采用自回归的方式逐个预测各个字。例如，在进行第 n 次迭代时，被设置为预测输入语音信号对应的第 n 个字，根据输入的前 $n-1$ 个字、第 $n-1$ 个字对应的注意力系数加权后的向量表示以及相关性系数对多个说话者的画像特征向量的加权和，预测第 n 个字。其中，预测出的第 n 个字是根据第 n 个字对应字典 V 中各个字的概率得到的。假设字典中一个 4950 个字，确定第 n 个字对应于字典 V 中各个字的概率 on ，即 on 为 4950 个概率，在获得的 4950 个概率中选取其中概率最大的字作为第 n 个字的预测结果。

具体的，对前 $n-1$ 个字分别进行嵌入编码，得到前 $n-1$ 个字各自对应的嵌入向量表示。可选地，可以先对前 $n-1$ 个字分别进行嵌入处理，获得处理后的向量表示，对该处理后的向量表示进行位置编码处理，获得前 $n-1$ 个字各自对应的嵌入向量表示。例

如,根据公式 $Z_{[1:n-1],l}^{asr} = PosEnc(Embed(y_{[1:n-1]}))$, 获得前 $n-1$ 个字各自对应的嵌入向量表示。其中, $PosEnc$ 表示位置编码, $Embed$ 表示嵌入函数, $y_{[1:n-1]}$ 表示语音识别解码器已经输出的前 $n-1$ 个字。

在得到前 $n-1$ 个字各自对应的嵌入向量表示之后,将前 $n-1$ 个字各自对应的嵌入向量表示和第一向量表示输入至第一层解码器中的注意力层,第一层解码器中的注意力层对前 $n-1$ 个字各自对应的嵌入向量表示和第一向量表示进行处理,以获得第一层解码器中的注意力层对应的第 $n-1$ 个字对应的注意力系数加权后的向量表示,并将其输入至第一层解码器中的前馈神经网络层。

在一可选实施例中,语音识别编码器中的各个解码器中的注意力层可以包括自注意力层和源-目标注意力层。那么,当说话者编码器接收到前 $n-1$ 个字各自对应的嵌入向量表示和第一向量表示之后,将前 $n-1$ 个字各自对应的嵌入向量表示输入至第一层解码器中的自注意力层,以获得第一层解码器中的自注意力层输出的第 $n-1$ 个字对应的自注意力系数加权后的向量表示,并将该向量表示输入至第一层解码器中的源-目标注意力层中。第一层解码器中的源-目标注意力层对第 $n-1$ 个字对应的自注意力系数加权后的向量表示和第一向量表示进行计算,获得源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示。接着,将第 $n-1$ 个字对应的注意力系数加权后的向量表示和相关性系数对多个说话者的画像特征向量的加权和输入第一层解码器中的前馈神经网络层,以获得输入到第二层解码器的第 $n-1$ 个字对应的向量表示。按照上述方法分别完成第二层以及第二层以上解码器的处理操作,直至获得语音识别解码器中最后一层解码器输出的第 $n-1$ 个字对应的向量表示。

具体的,在一可选实施例中,第一层及第一层以上的解码器中的自注意力层可以根据公式 $\overline{Z}_{n-1,l}^{asr} = Z_{n-1,l}^{asr} + MHA_l^{asr-self}(Z_{n-1,l}^{asr}, Z_{[1:n-1],l}^{asr}, Z_{[1:n-1],l}^{asr})$, 计算输入至该层解码器中的源-目标注意力层的第 $n-1$ 个字对应的自注意力系数加权后的向量表示。其中, l 表示编码器所对应的层数, $\overline{Z}_{n-1,l}^{asr}$ 表示第 l 层编码器中的自注意力层输出的第 $n-1$ 个字对应的自注意力系数加权后的向量表示 (即输入至第 l 层编码器中的源-目标注意力层的第 $n-1$ 个字对应的自注意力系数加权后的向量表示), $Z_{n-1,l}^{asr}$ 表示第 $l-1$ 层编码器中的前馈神经网络层输出的第 $n-1$ 个字对应的向量表示, $MHA_l^{asr-self}$ 表示第 l 层编码器中的自注意力层的多头注意力, $Z_{[1:n-1],l}^{asr}$ 表示第 $l-1$ 层编码器中的前馈神经网络层输出的前 $n-1$ 个字对应的向量表示 (输入至第 l 层编码器中的前 $n-1$ 个字对应的向量表示)。

在一可选实施例中,第一层及第一层以上的解码器中的源-目标注意力层可以根据公式 $\overline{\overline{Z}}_{n-1,l}^{asr} = \overline{Z}_{n-1,l}^{asr} + MHA_l^{asr-src}(\overline{Z}_{n-1,l}^{asr}, H^{asr}, H^{asr})$, 计算输入至该层解码器中的前馈神经网络层的第 $n-1$ 个字对应的注意力系数加权后的向量表示。其中, l 表示编码器所对应的层

数, 且 $\overline{\overline{Z}}_{n-1,l}^{asr}$ 表示第 l 层编码器中的源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示 (即输入至第 l 层编码器中的前馈神经网络层的第 $n-1$ 个字对应的注意力系数加权后的向量表示), $\overline{\overline{Z}}_{n-1,l}^{asr}$ 表示第 l 层编码器中的自注意力层输出的第 $n-1$ 个字对应的自注意力系数加权后的向量表示, $MHA_i^{asr-stc}$ 表示第 l 层编码器中的源-目标注意力层的多头注意力, H^{asr} 表示第一向量表示。

在一可选实施例中, 第一层解码器中的前馈神经网络层可以根据公式 $Z_{n-1,l+1}^{asr} = \overline{\overline{Z}}_{n-1,l}^{asr} + FF_l^{asr}(\overline{\overline{Z}}_{n-1,l}^{asr} + W^{spk} \cdot \overline{d}_n)$, 计算输入至第二层编码器的第 $n-1$ 个字对应的向量表示。其中, l 表示编码器所对应的层数, 且 $l=1$, $Z_{n-1,l+1}^{asr}$ 表示第 l 层编码器中的前馈神经网络输出的第 $n-1$ 个字对应的向量表示 (即输入至第 $l+1$ 层编码器中的第 $n-1$ 个字对应的向量表示), $\overline{\overline{Z}}_{n-1,l}^{asr}$ 表示第 l 层编码器中的源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示, FF_l^{asr} 表示第 l 层编码器中的前馈神经网络, W^{spk} 为说话者编码器对应的权重, $\overline{d}_n$ 为相关性系数对多个说话者的画像特征向量的加权和。对于第二层及第二层以上解码器中的前馈神经网络层可以根据公式 $Z_{n-1,l+1}^{asr} = \overline{\overline{Z}}_{n-1,l}^{asr} + FF_l^{asr}(\overline{\overline{Z}}_{n-1,l}^{asr})$, 计算输入至下一层编码器的第 $n-1$ 个字对应的向量表示 (即前馈神经网络层的输出作为下一层编码器的输入)。其中, l 表示编码器所对应的层数, 且 $l>1$ 。

最后, 根据语音识别解码器中最后一层解码器输出的第 $n-1$ 个字对应的向量表示, 确定第 n 个字。具体的, 利用 softmax 操作对其最后一层解码器输出的第 $n-1$ 个字对应的向量表示进行处理, 获得第 n 个字对应的后验概率 O_n , 再根据第 n 个字对应的后验概率 O_n , 预测第 n 个字。例如, 假设最后一层解码器输出的第 $n-1$ 个字对应的向量表示为 $Z_{n-1,L}^{asr}$, 根据公式 $O_n = \text{soft max}(W^o \cdot Z_{n-1,L}^{asr} + b^o)$, 确定第 n 个字对应的后验概率 O_n , 其中 soft max 为激活函数, W^o 、 b^o 分别为语音识别编码器对应的权重。

由上述描述可知: 对于语音识别解码器中的第 1 层解码器, 对已经输出的前 $n-1$ 个字进行嵌入编码处理, 在得到输入至第 1 层解码器的 $n-1$ 个字分别对应的嵌入向量表示之后, 取出第 $n-1$ 个字的嵌入向量表示作为自注意力层中的 query, 这 $n-1$ 个字分别对应的嵌入向量表示作为 key、value, 进行自注意力处理, 得到第 $n-1$ 个字对这前 $n-1$ 个字进行注意力加权后的结果, 即第 $n-1$ 个字对应的自注意力系数加权后的编码向量, 再与加权前的第 $n-1$ 个字的嵌入向量表示相加, 获得的相加结果即为第 $n-1$ 个字对应的自注意力系数加权后的向量表示, 并将其作为源-目标注意力层的其中一个输入, 另外两个输入为两个第一向量表示, 最终得到第 1 层中源-目标注意力层输出的第 $n-1$ 个字对应的注意力系数加权后的向量表示。之后, 将第 $n-1$ 个字对应的注意力系数加权后的向量表示输入到第一层编码器中的前馈神经网络层中, 以便得到第 1 层编

码器的输出结果亦即为第 2 层编码器的输入。

需要注意的是，针对第 1 层编码器中的前馈神经网络层，再计算其输出的时候会加上 $\overline{d_n}$ 这一项，其他层编码器中的前馈神经网络层进行处理时不加此项。其中，在第一层编码器中的前馈神经网络层进行计算时引入的 $\overline{d_n}$ 是为了在语音识别第 n 个字的过程中引入第 n 个字对应于各个说话者的加权后的说话向量表示，从而识别出的第 n 个字不包括该字对应的具体内容是什么还包括该字对应的说话者是什么。这样语音识别编码器最后就可以直接输出 SOT 格式的识别结果，例如一个说话者的文字序列 <sc>另一个说话者的文字序列。

综上，在本发明实施例中，语音解码器采用自回归的方式逐个预测各个字，在对各个字进行预测处理时，语音编码器中的各层解码器按照层级逐层对输入的前 n-1 个字、第 n-1 个字对应的注意力系数加权后的向量表示以及相关性系数对多个说话者的画像特征向量的加权和进行解码处理，并且上一层解码器的输出作为下一层解码器的输入，最后根据语音识别解码器中最后一层解码器输出的第 n-1 个字对应的向量表示，确定第 n 个字，这样获得的第 n 个字不包含该字对应的内容和该字对应的说话者，以便后续直接基于预测出的各个字生成 SOT 格式的文字内容，而且确定出的各个字时综合考虑了上下文信息，可以使得预测出的各个字更准确。

近年来，随着越来越多的研究集中在更真实场景下的语音识别处理，例如多人会议自动记录、多方人机交互、音/视频自动标注等。具体应用时，参考附图 7 所示。语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器、包含注意力层的文本编码器、以及联合打分器。其中，联合打分器包括说话者无关打分器和说话者有关打分器，说话者无关打分器可以根据设定的相似度算法，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度，说话者有关打分器中包括设定的包含注意力模块的打分模型，以通过设定的包含注意力模块的打分模型，获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度。

具体的，在进行语音识别时，按照迭代的方式逐个预测语音信号对应的各个字。各个字的具体识别预测过程基本一致，这里以第 n 个字的识别预测过程进行示例性说明。首先获取多个说话者的语音信号和多个说话者的画像特征向量，其中，语音信号中包含多个说话者的语音。接着，将多个说话者的语音信号输入至语音识别编码器和说话者编码器中，通过语音识别编码器获取语音信号对应的第一向量表示，通过说话者编码器获取语音信号对应的第二向量表示。其中，第一向量表示被设置为进行语音识别，第二向量表示被设置为进行说话者识别。获取语音识别解码器已经输出的前 n-1 个字，通过文字编码器对语音识别解码器已经输出的前 n-1 个字进行编码处理，以得到第 n-1 个字所对应的第三向量表示。将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中，以获得第 n 个字对应的说话者向量表示。

然后，分别将第 n 个字对应的说话者向量表示与任一说话者的画像特征向量进行拼接，将拼接后的向量表示输入至说话者无关打分器中，针对多个说话者中任一说话者的画像特征向量，根据设定的相似度算法，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度。接着，分别将前 n 个字各自对应的说话者向量表示与任一说话者的画像特征向量进行拼接，并将多个拼接后的向量表示输入至说话者相关打分器中设定的包含注意力模块的打分模型，以获得前 n 个字各自对应的说话者向量表示分别与任一说话者的画像特征向量之间的第二相似度。根据第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第一相似度以及第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的第二相似度，确定第 n 个字对应的说话者向量表示与任一说话者的画像特征向量之间的相关性系数。而后，计算该相关性系数对多个说话者的画像特征向量的加权和。

接着，将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和，输入语音识别解码器，获得第 n 个字对应的后验概率。再根据第 n 个字对应的后验概率，预测第 n 个字。最后，根据语音识别器输出的各个字，生成 SOT 格式的文字序列。

其中，本发明实施例中涉及到的具体的实施过程可以参考上述实施例中内容，在此不再赘述。

为了使得上述实施例中文本编码器、说话者无关打分器、以及说话者相关打分器可以更好地获得语音信号对应地全局上下文信息，进行双阶段解码处理。第一阶段解码就是按照上述各个实施例中的方法逐个预测语音信号对应的各个字。当获得语音信号对应的所有 N 个字之后，进行第二阶段解码处理，第二阶段解码是通过第一阶段解码处理后得到所有 N 个字的预测结果，将这 N 个字的预测结果作为文本编码器的输入，再逐个计算出一个第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，即获得 N 个字中每个字对应的说话者的概率分布，根据 N 个字中每个字对应的说话者的概率分布，以重新获得各个字对应的说话者，这样可以使得最后获得的 N 个字的结果更加准确。其具体的实现过程可以参考附图 8 所示。

图 8 为本发明实施例提供的另一种语音识别方法的流程图；如图 8 所示，为了提高该方法语音识别结果的准确性，使得该语音识别方法可以更好地结合全局的上下文信息，在上述实施例的基础上，该方法还可以包括如下步骤：

801、确定语音识别解码器输出的完整的文字序列。

802、通过文本编码器对文字序列进行编码处理，以得到文字序列中各个字对应的第四向量表示。

803、将第一向量表示、第二向量表示和第四向量表示输入说话者解码器中，以获得各个字对应的说话者向量表示。

804、根据各个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的

相关性系数，重新确定各个字对应的说话者。

在本实施例中，在进行第二阶段解码时，首先确定第一阶段语音识别解码器输出的完整的文字序列。然后将完整的文字序列输入至文本编码器，以通过文本编码器对文字序列进行编码处理，以得到文字序列中各个字对应的第四向量表示。而后，将第一向量表示、第二向量表示和第四向量表示输入说话者解码器中，以获得各个字对应的说话者向量表示。根据各个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，重新确定各个字对应的说话者。

由上述描述可知：第一阶段解码是基于前 $n-1$ 个字预测第 n 个字，只能获取到该字对应的上文信息，没有下文信息。第二阶段解码是在所有 N 个字预测完一遍之后，基于所有 N 个字再进行预测时，这样针对其中每个字就可以看到上文信息和下文信息了，可以使得再计算一遍获得的各个字对应的说话者的预测结果会更准确。

其中，本发明实施例中涉及到的具体的实施过程可以参考上述各个实施例中的内容，在此不再赘述。

本发明实施例提供的语音识别方法可以在云端来执行，在云端可以部署有若干计算节点（云服务器），每个计算节点中都具有计算、存储等处理资源。在云端，可以组织由多个计算节点来提供某种服务，当然，一个计算节点也可以提供一种或多种服务。云端提供该服务的方式可以是对外提供服务接口，用户调用该服务接口以使用相应的服务。

针对本发明实施例提供的方案，云端可以提供有语音识别服务的接口，用户通过终端设备调用该服务接口，以向云端触发语音识别服务请求，该请求中包括多个说话者的语音信号和多个说话者的画像特征向量，语音信号中包含多个说话者的语音，云端确定响应该请求的计算节点，利用该计算节点中的处理资源执行如下步骤：

获取语音识别模型，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器；

通过语音识别编码器获取语音信号对应的第一向量表示，通过说话者编码器获取语音信号对应的第二向量表示，第一向量表示被设置为进行语音识别，第二向量表示被设置为进行说话者识别；

通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

将第一向量表示、第二向量表示和第三向量表示输入说话者解码器中，以获得第 n 个字对应的说话者向量表示；

根据第 n 个字对应的说话者向量表示分别与多个说话者的画像特征向量之间的相关性系数，确定第 n 个字对应的说话者；

将第一向量表示、前 $n-1$ 个字以及相关性系数对多个说话者的画像特征向量的加权和，输入语音识别解码器，以获得第 n 个字；

向终端设备发送语音识别输出信息，语音识别输出信息中包括多个说话者各自对应的文字序列。

上述执行过程可以参考前述其他实施例中的相关说明，在此不赘述。

为便于理解，结合图 9 来示例性说明。用户可以通过图 9 中示意的终端设备 E1 调用语音识别服务，以对目标语音信号进行语音识别，以获得目标语音信号对应的文本序列预测。用户调用该服务的服务接口包括软件开发工具包 (Software Development Kit, 简称 SDK)、应用程序接口 (Application Programming Interface, 简称 API) 等形式。图 9 中示意的是 API 接口的情形。在云端，如图中所示，假设由服务集群 E2 提供语音识别服务，服务集群 E2 中包括至少一个计算节点。服务集群 E2 收到该请求后，执行前述实施例中的步骤，以得到目标语音信号对应的包括多个说话者各自对应的文字序列，并反馈给终端设备 E1。

以下将详细描述本发明的一个或多个实施例的语音识别装置。本领域技术人员可以理解，这些装置均可使用市售的硬件组件通过本方案所教导的步骤进行配置来构成。

图 10 为本发明实施例提供的一种语音识别装置的结构示意图，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器，如图 10 所示，该装置包括：获取模块 11、第一编码模块 12、第二编码模块 13、第一解码模块 14、第二解码模块 15。

获取模块 11，被设置为获取多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音。

第一编码模块 12，被设置为通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别。

第二编码模块 13，被设置为通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示。

第一解码模块 14，被设置为将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者。

第二解码模块 15，被设置为将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权求和，输入所述语音识别解码器，以获得所述第 n 个字。

可选地，所述第一解码模块 14 还被设置为：针对所述多个说话者中任一说话者的画像特征向量，根据设定的相似度算法，确定所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第一相似度；获取前 n 个字各自对应的说话者向量表示；分别将所述前 n 个字各自对应的说话者向量表示与所述任一说话者的画像

特征向量进行拼接；将多个拼接后的向量表示输入设定的包含注意力模块的打分模型，以获得所述前 n 个字各自对应的说话者向量表示分别与所述任一说话者的画像特征向量之间的第二相似度；根据所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第一相似度以及所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第二相似度，确定所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的相关性系数。

可选地，所述第二编码模块 13 还被设置为：获取所述语音识别解码器中的自注意力层分别对所述前 $n-1$ 个字进行注意力计算后得到的所述前 $n-1$ 个字分别对应的注意力系数加权后的向量表示；将所述前 $n-1$ 个字分别对应的注意力系数加权后的向量表示输入所述文本编码器中，以获得所述文本编码器输出的第 $n-1$ 个字所对应的第三向量表示。

可选地，所述文本编码器中包括级联的至少一个编码器，每个所述编码器中包括注意力层和前馈神经网络层。

可选地，所述语音识别解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层；基于此，所述第二解码模块 15 还被设置为：对所述前 $n-1$ 个字分别进行嵌入编码，得到所述前 $n-1$ 个字各自对应的嵌入向量表示；将所述前 $n-1$ 个字各自对应的嵌入向量表示和所述第一向量表示输入所述语音识别解码器的第一层解码器中，以获得所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的向量表示；将所述第 $n-1$ 个字对应的注意力系数加权后的向量表示和所述加权和输入所述第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的所述第 $n-1$ 个字对应的向量表示；根据所述语音识别解码器中最后一层解码器输出的所述第 $n-1$ 个字对应的向量表示，确定所述第 n 个字。

可选地，所述说话者解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层；基于此，所述第一解码模块 14 还被设置为：将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中的第一层解码器中，以获得所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示；将所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示输入所述第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的所述第 $n-1$ 个字对应的说话者向量表示；确定所述说话者解码器中最后一层解码器输出的所述第 $n-1$ 个字对应的说话者向量表示；根据所述最后一层解码器输出的所述第 $n-1$ 个字对应的说话者向量表示和所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示，确定所述第 n 个字对应的说话者向量表示。

可选地，所述第二解码模块 15 还被设置为：确定所述语音识别解码器输出的完整的文字序列；通过所述文本编码器对所述文字序列进行编码处理，以得到所述文字序

列中各个字对应的第四向量表示；将所述第一向量表示、所述第二向量表示和所述第四向量表示输入所述说话者解码器中，以获得所述各个字对应的说话者向量表示；根据所述各个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，重新确定所述各个字对应的说话者。

图 10 所示装置可以执行前述实施例中语音识别方法中的步骤，详细的执行过程和技术效果参见前述实施例中的描述，在此不再赘述。

本发明实施例还提供一电子设备，如图 11 所示，该电子设备可以包括：处理器 21、存储器 22、通信接口 23。其中，存储器 22 上存储有可执行代码，当所述可执行代码被处理器 21 执行时，使处理器 21 实现如前述实施例中语音识别方法。

另外，本发明实施例提供了一种非暂时性机器可读存储介质，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器至少可以实现如前述实施例中提供的语音识别方法。

以上所描述的装置实施例是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助加必需的通用硬件平台的方式来实现，当然也可以通过硬件和软件结合的方式来实现。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以计算机产品的形式体现出来，本发明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。

最后应说明的是：以上实施例用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

工业实用性

本公开实施例提供了一种语音识别方法、设备和存储介质，具体实施方案为在语音识别模型中添加包含注意力层的文本编码器，通过文本编码器对语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，可以更好地聚合整个句子的上下文信息，以获得第 $n-1$ 个字所对应的包含上下文信息的第三向量表示，即这样可以获得更准确地被设置为进行说话者识别的向量表示，从而提高了各个字对应的说话者识别结果的准确性。

权 利 要 求 书

1.一种语音识别方法，其特征在于，语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器；所述方法包括：

获取多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音；

通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别；

通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；

根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者；

将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权和，输入所述语音识别解码器，以获得所述第 n 个字。

2.根据权利要求1所述的方法，其特征在于，所述方法还包括：

针对所述多个说话者中任一说话者的画像特征向量，根据设定的相似度算法，确定所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第一相似度；

获取前 n 个字各自对应的说话者向量表示；

分别将所述前 n 个字各自对应的说话者向量表示与所述任一说话者的画像特征向量进行拼接；

将多个拼接后的向量表示输入设定的包含注意力模块的打分模型，以获得所述前 n 个字各自对应的说话者向量表示分别与所述任一说话者的画像特征向量之间的第二相似度；

根据所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第一相似度以及所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的第二相似度，确定所述第 n 个字对应的说话者向量表示与所述任一说话者的画像特征向量之间的相关性系数。

3.根据权利要求1所述的方法，其特征在于，所述通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示，包括：

获取所述语音识别解码器中的自注意力层分别对所述前 $n-1$ 个字进行注意力计算后得到的所述前 $n-1$ 个字分别对应的注意力系数加权后的向量表示；

将所述前 $n-1$ 个字分别对应的注意力系数加权后的向量表示输入所述文本编码器中，以获得所述文本编码器输出的第 $n-1$ 个字所对应的第三向量表示。。

4.根据权利要求 3 所述的方法，其特征在于，所述文本编码器中包括级联的至少一个编码器，每个所述编码器中包括注意力层和前馈神经网络层。

5.根据权利要求 1 所述的方法，其特征在于，所述语音识别解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层；

所述将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权和，输入所述语音识别解码器，以获得所述第 n 个字，包括：

对所述前 $n-1$ 个字分别进行嵌入编码，得到所述前 $n-1$ 个字各自对应的嵌入向量表示；

将所述前 $n-1$ 个字各自对应的嵌入向量表示和所述第一向量表示输入所述语音识别解码器的第一层解码器中，以获得所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的向量表示；

将所述第 $n-1$ 个字对应的注意力系数加权后的向量表示和所述加权和输入所述第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的所述第 $n-1$ 个字对应的向量表示；

根据所述语音识别解码器中最后一层解码器输出的所述第 $n-1$ 个字对应的向量表示，确定所述第 n 个字。

6.根据权利要求 1 所述的方法，其特征在于，所述说话者解码器中包括级联的多个解码器，每个解码器中包括注意力层和前馈神经网络层；

所述将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示，包括：

将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中的第一层解码器中，以获得所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示；

将所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示输入所述第一层解码器中的前馈神经网络层，以获得输入到第二层解码器的所述第 $n-1$ 个字对应的说话者向量表示；

确定所述说话者解码器中最后一层解码器输出的所述第 $n-1$ 个字对应的说话者向量表示；

根据所述最后一层解码器输出的所述第 $n-1$ 个字对应的说话者向量表示和所述第一层解码器中的注意力层输出的所述第 $n-1$ 个字对应的注意力系数加权后的说话者向量表示，确定所述第 n 个字对应的说话者向量表示。

7.根据权利要求1至6中任一项所述的方法，其特征在于，所述方法还包括：
确定所述语音识别解码器输出的完整的文字序列；

通过所述文本编码器对所述文字序列进行编码处理，以得到所述文字序列中各个字对应的第四向量表示；

将所述第一向量表示、所述第二向量表示和所述第四向量表示输入所述说话者解码器中，以获得所述各个字对应的说话者向量表示；

根据所述各个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，重新确定所述各个字对应的说话者。

8.一种语音识别方法，其特征在于，所述方法包括：

接收终端设备通过调用语音识别服务触发的请求，所述请求中包括多个说话者的语音信号和所述多个说话者的画像特征向量，所述语音信号中包含所述多个说话者的语音；

根据所述模型训练服务对应的计算资源执行如下步骤：

获取语音识别模型，所述语音识别模型中包括语音识别编码器、说话者编码器、语音识别解码器、说话者解码器以及包含注意力层的文本编码器；

通过所述语音识别编码器获取所述语音信号对应的第一向量表示，通过所述说话者编码器获取所述语音信号对应的第二向量表示，所述第一向量表示被设置为进行语音识别，所述第二向量表示被设置为进行说话者识别；

通过所述文本编码器对所述语音识别解码器已经输出的前 $n-1$ 个字进行编码处理，以得到第 $n-1$ 个字所对应的第三向量表示；

将所述第一向量表示、所述第二向量表示和所述第三向量表示输入所述说话者解码器中，以获得第 n 个字对应的说话者向量表示；

根据所述第 n 个字对应的说话者向量表示分别与所述多个说话者的画像特征向量之间的相关性系数，确定所述第 n 个字对应的说话者；

将所述第一向量表示、所述前 $n-1$ 个字以及所述相关性系数对所述多个说话者的画像特征向量的加权和，输入所述语音识别解码器，以获得所述第 n 个字；

向所述终端设备发送语音识别输出信息，所述语音识别输出信息中包括所述多个说话者各自对应的文字序列。

9.一种电子设备，其特征在于，包括：存储器、处理器、通信接口；其中，所述存储器上存储有可执行代码，当所述可执行代码被所述处理器执行时，使所述处理器执行如权利要求1至7中任一项所述的语音识别方法。

10.一种非暂时性机器可读存储介质，其特征在于，所述非暂时性机器可读存储介质上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如权利要求1至7中任一项所述的语音识别方法。

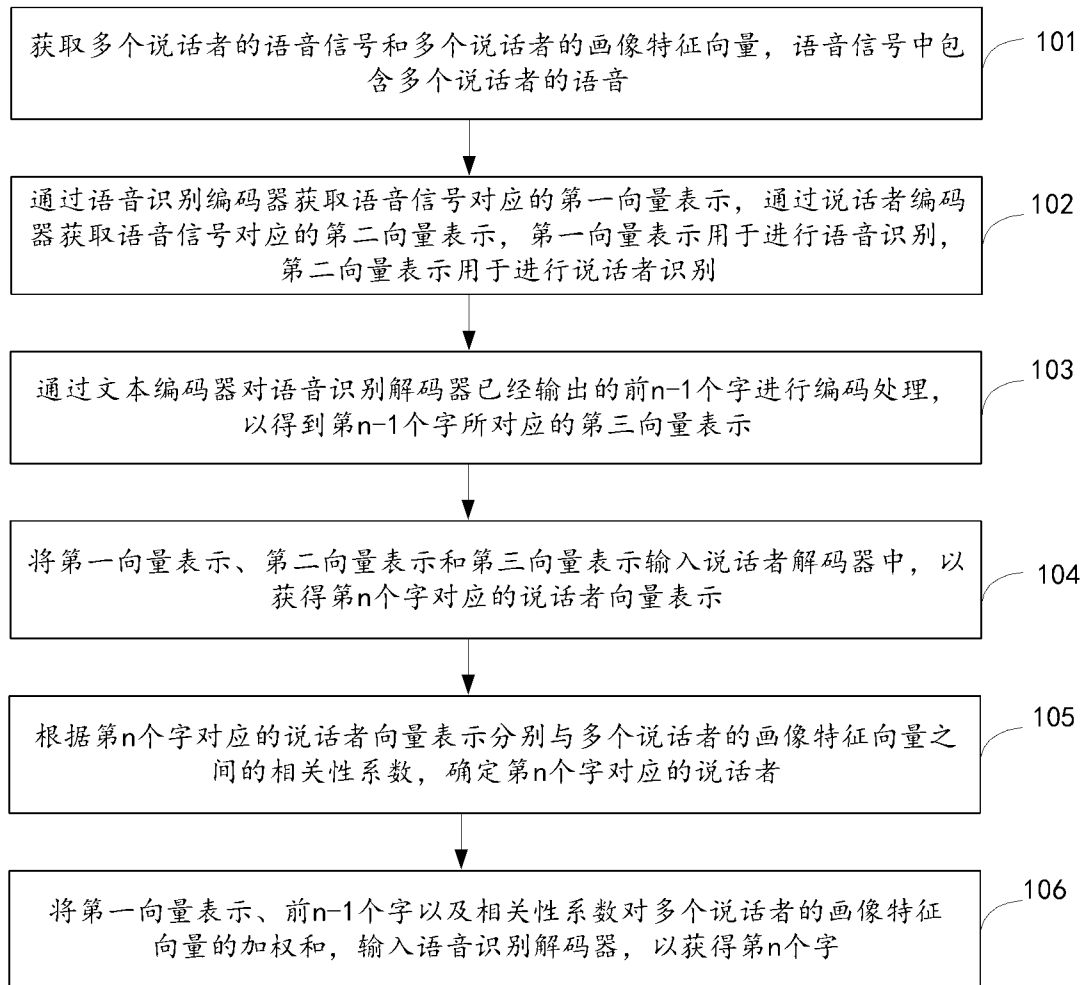


图 1

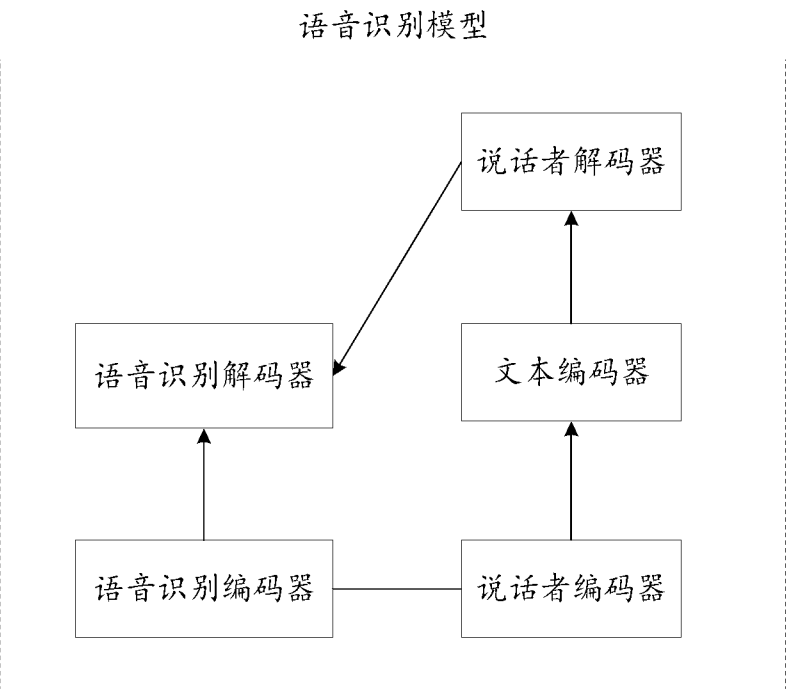


图 2

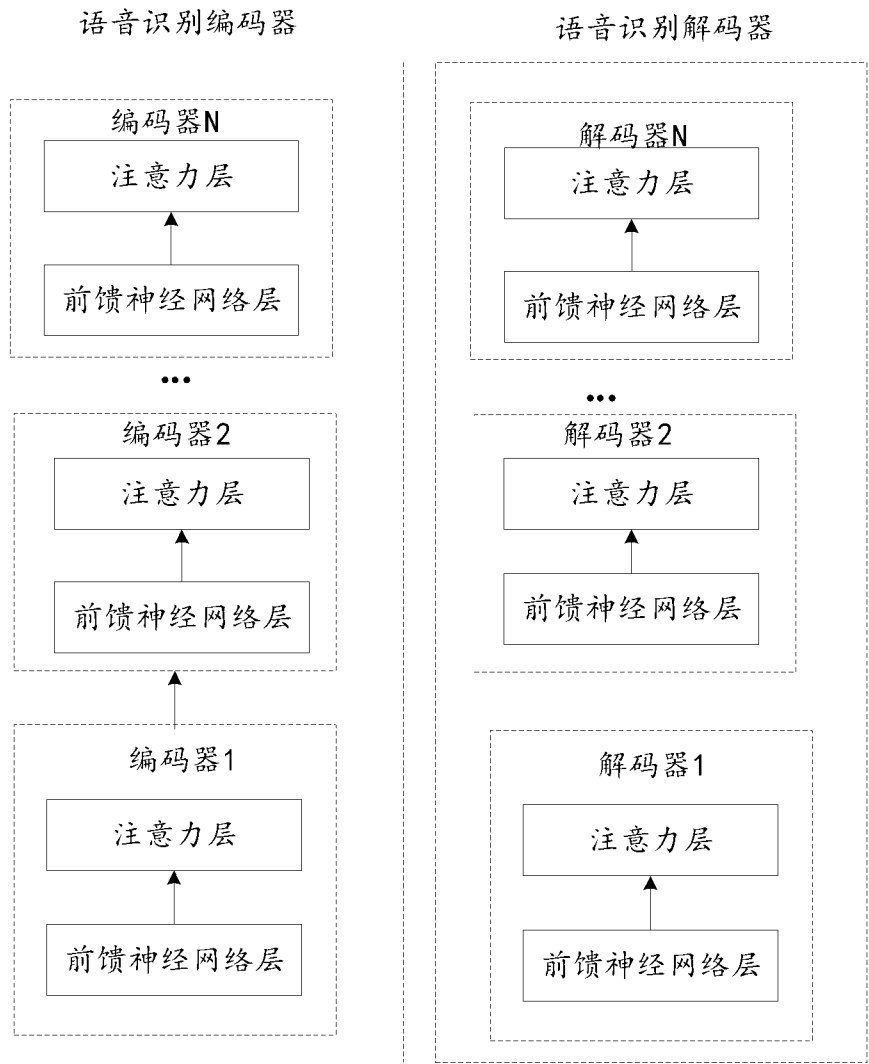


图 3

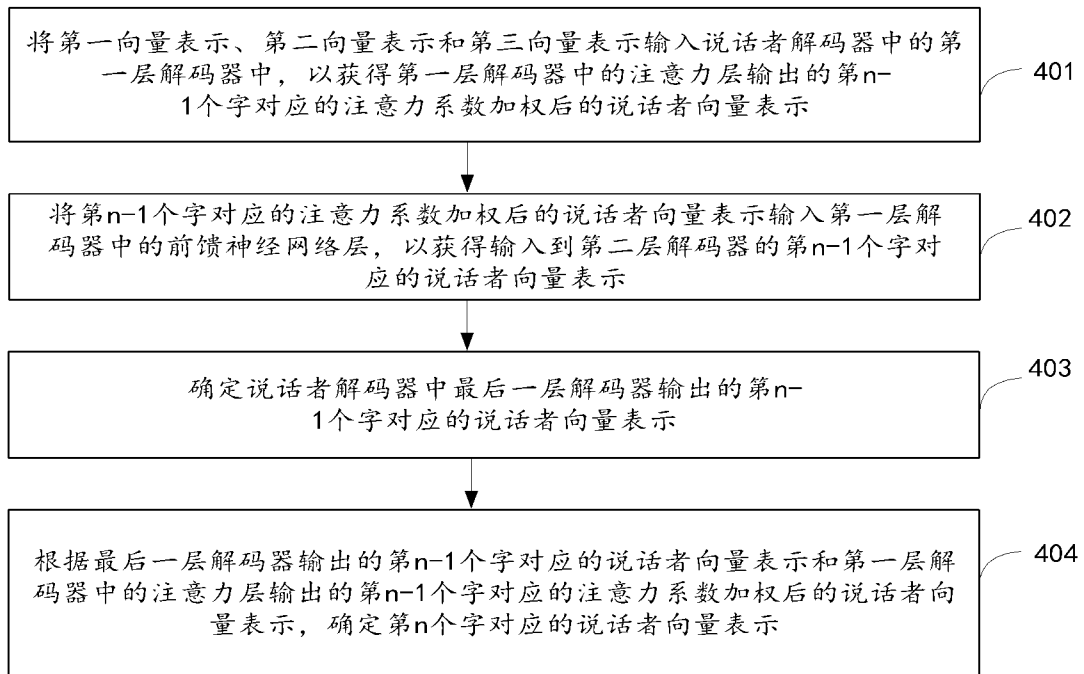


图 4

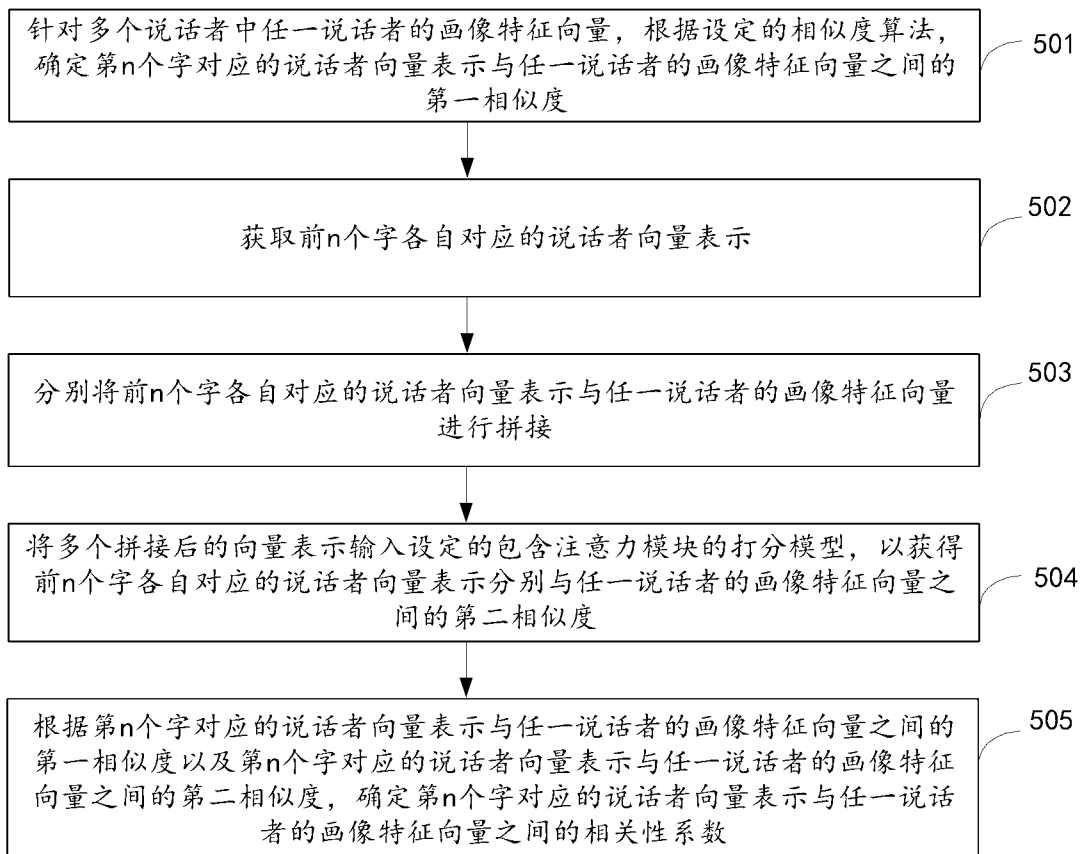


图 5

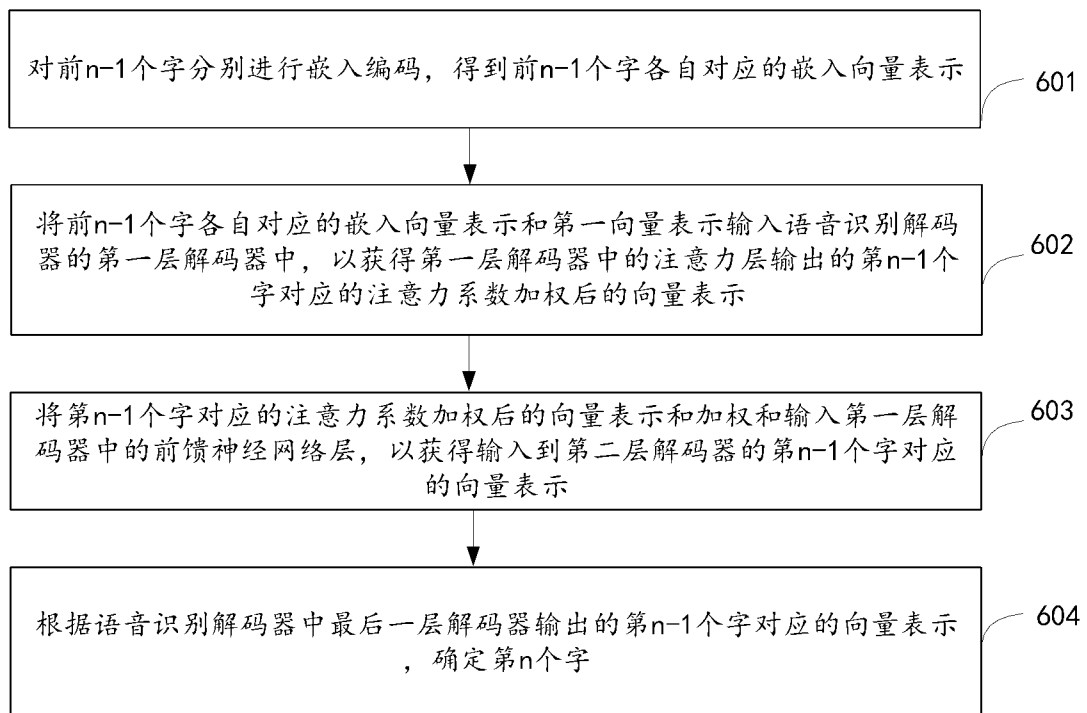


图 6

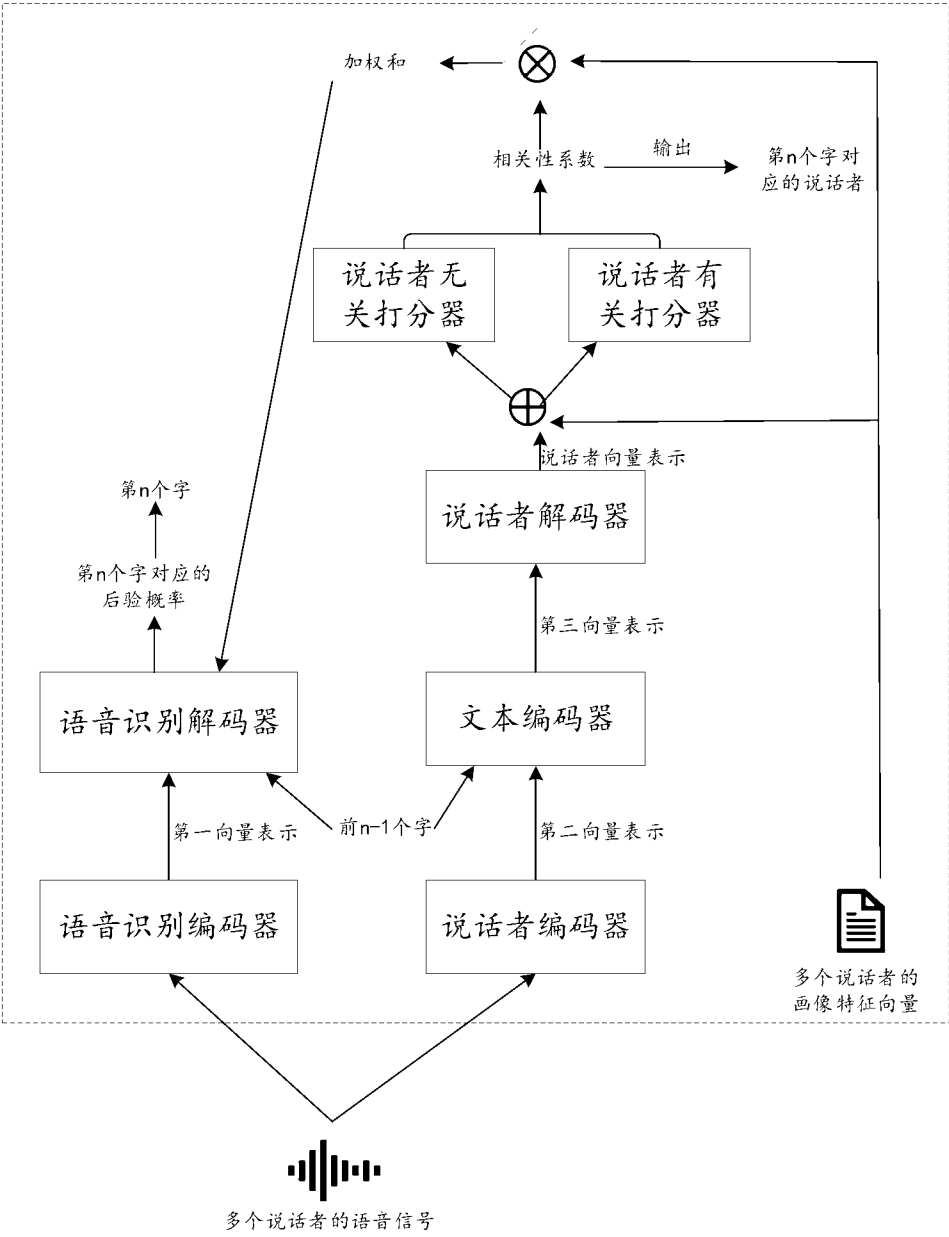


图 7

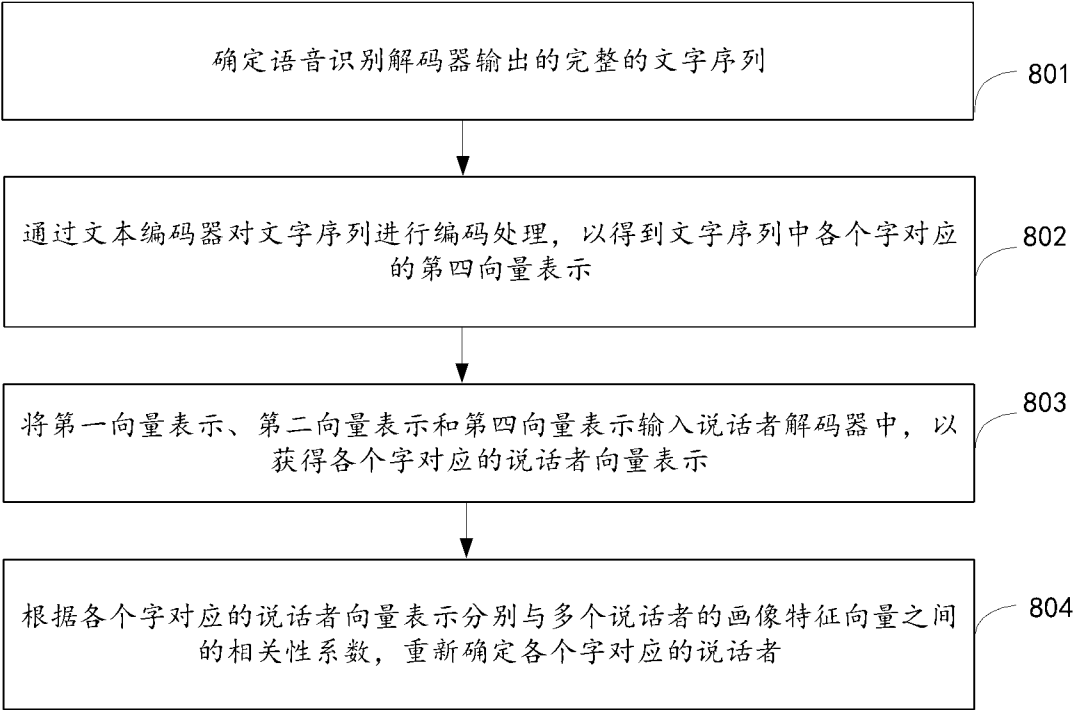


图 8

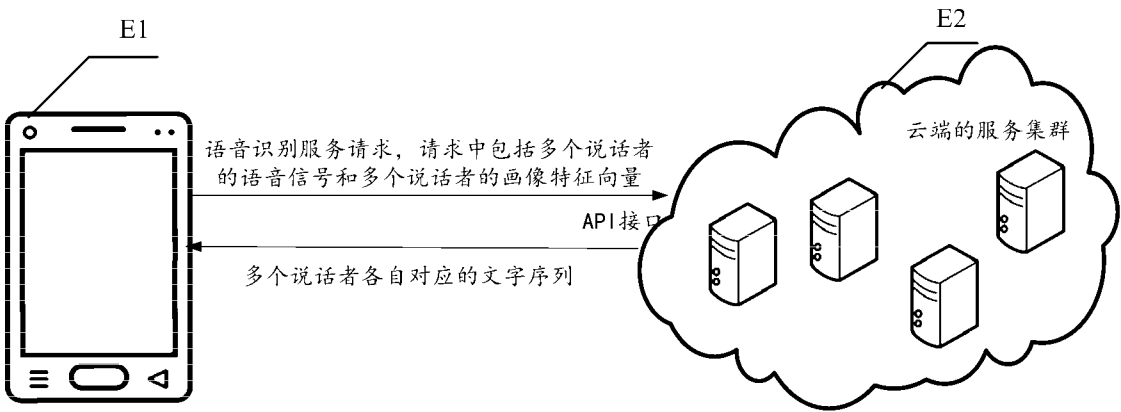


图 9

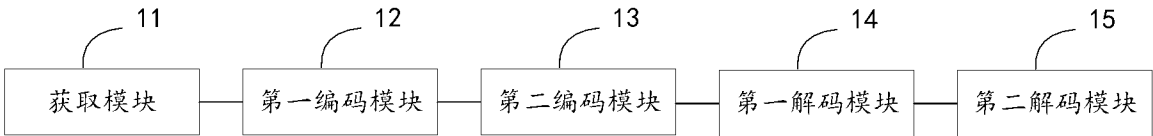


图 10

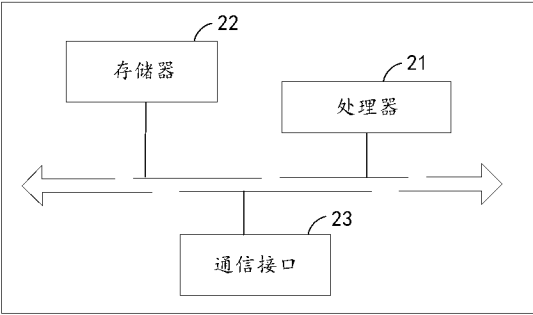


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/097609

A. CLASSIFICATION OF SUBJECT MATTER G10L 15/20(2006.01)i; G10L 15/183(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) IPC: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) CNABS, CNTXT, CNKI: 多, 说话人, 话者, 上下文, 语音, 识别, 向量, 序列, 特征, 字, 词, 前一帧; VEN, ENTXT, USTXT, WOTXT, ELSEVIER: multiple, speaker, talker, context, speech, voice, recognition, vector, sequence, feature, word, n-1		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116825095 A (ALIBABA (CHINA) CO., LTD.) 29 September 2023 (2023-09-29) description, paragraphs 46-175, and figures 1-11	1-10
X	CN 113889089 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 04 January 2022 (2022-01-04) description, paragraphs 30-122, and figures 1-6	1-10
A	CN 111009237 A (BEIJING DAJIA INTERNET INFORMATION TECHNOLOGY CO., LTD.) 14 April 2020 (2020-04-14) entire document	1-10
A	CN 109215662 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 15 January 2019 (2019-01-15) entire document	1-10
A	CN 111899727 A (AI SPEECH LTD.) 06 November 2020 (2020-11-06) entire document	1-10
A	US 2020043467 A1 (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 06 February 2020 (2020-02-06) 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 21 August 2024		Date of mailing of the international search report 22 August 2024
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/ CN) China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/097609

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	116825095	A	29 September 2023	None			
CN	113889089	A	04 January 2022	CN	113889089	B	19 July 2024
CN	111009237	A	14 April 2020	CN	111009237	B	01 July 2022
CN	109215662	A	15 January 2019	CN	109215662	B	20 June 2023
CN	111899727	A	06 November 2020	CN	111899727	B	06 May 2022
US	2020043467	A1	06 February 2020	US	10699700	B2	30 June 2020
				WO	2020024646	A1	06 February 2020

A. 主题的分类

G10L 15/20(2006.01)i; G10L 15/183(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)

IPC: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))

CNABS, CNTXT, CNKI: 多, 说话人, 话者, 上下文, 语音, 识别, 向量, 序列, 特征, 字, 词, 前一帧 VEN, ENTXT,USTXT, WOTXT, ELSEVIER: multiple, speaker, talker, context, speech, voice, recognition, vector, sequence, feature, word, n-1

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
PX	CN 116825095 A (阿里巴巴(中国)有限公司) 2023年9月29日 (2023 - 09 - 29) 说明书第46-175段,图1-11	1-10
X	CN 113889089 A (北京百度网讯科技有限公司) 2022年1月4日 (2022 - 01 - 04) 说明书第30-122段,图1-6	1-10
A	CN 111009237 A (北京达佳互联信息技术有限公司) 2020年4月14日 (2020 - 04 - 14) 全文	1-10
A	CN 109215662 A (平安科技(深圳)有限公司) 2019年1月15日 (2019 - 01 - 15) 全文	1-10
A	CN 111899727 A (苏州思必驰信息科技有限公司) 2020年11月6日 (2020 - 11 - 06) 全文	1-10
A	US 2020043467 A1 (TENCENT TECH SHENZHEN CO LTD) 2020年2月6日 (2020 - 02 - 06) 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

2024年8月21日

国际检索报告邮寄日期

2024年8月22日

ISA/CN的名称和邮寄地址

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

授权官员

王婷玉

电话号码 (+86) 020-28958521

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/097609

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	116825095	A	2023年9月29日	无	
CN	113889089	A	2022年1月4日	CN	113889089 B 2024年7月19日
CN	111009237	A	2020年4月14日	CN	111009237 B 2022年7月1日
CN	109215662	A	2019年1月15日	CN	109215662 B 2023年6月20日
CN	111899727	A	2020年11月6日	CN	111899727 B 2022年5月6日
US	2020043467	A1	2020年2月6日	US	10699700 B2 2020年6月30日
				WO	2020024646 A1 2020年2月6日