

[19] 中华人民共和国国家知识产权局

[ 51 ] Int. Cl<sup>7</sup>

G10L 15/00

G10L 15/20 G10L 15/14

G10L 13/00



# [12] 发 明 专 利 说 明 书

[21] ZL 专利号 01140615.1

[45] 授权公告日 2004 年 7 月 28 日

[11] 授权公告号 CN 1159703C

[22] 申请日 2001.9.18 [21] 申请号 01140615.1

[30] 优先权

[32] 2000.9.18 [33] JP [31] 282349/2000

[71] 专利权人 日本先锋公司

地址 日本东京都

[72] 发明人 濑尾寻 驹村光弥 外山聪一

审查员 朱 骥

[74] 专利代理机构 北京市柳沈律师事务所

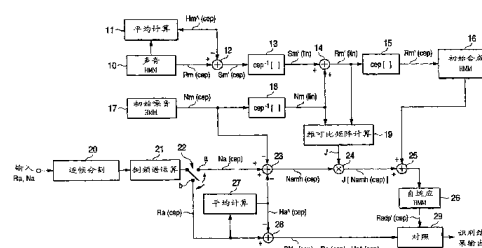
代理人 黄小临 王志森

权利要求书 3 页 说明书 21 页 附图 4 页

[54] 发明名称 声音识别系统

[57] 摘要

从含有可积性失真和可加性噪声的初始噪声 HMM 的声音 HMM10 中生成初始合成 HMM16, 同时由雅可比矩阵计算单元 19 计算雅可比矩阵 J。将可积性失真的估计值  $H_a(\text{cep})$ 、可加性噪声  $N_a(\text{cep})$  和初始噪声 HMM17 的可加性噪声  $N_m(\text{cep})$  合成所得的噪声变化量  $N_{amh}(\text{cep})$  与雅可比矩阵相乘, 其中将相乘结果和初始合成 HMM16 进行合成, 并生成自适应 HMM26。从而可以事先生成与从真正讲话声音中生成的观察值序列  $R_{Na}(\text{cep})$  相匹配的自适应 HMM26。当通过将观察值序列  $R_{Na}(\text{cep})$  与自适应 HMM26 对照进行声音识别时, 抵消了可积性失真和可加性噪声的影响, 其中可以获得与利用纯声音实现声音识别的情况等效的效果, 并可以实现抗干扰的声音识别系统。



1. 一种声音识别系统, 包括:
  - 声音模型, 从事先采集的、包含可积性失真的声音中生成;
  - 5 噪声模型, 从事先采集的可加性噪声中生成;
  - 第一估计单元, 估计包含在声音模型中的可积性失真的第一估计值;
  - 合成声音模型, 根据附加了噪声的声音生成, 所述声音是通过合成声音模型、可积性失真的第一估计值和噪声模型的可加性噪声进行处理而得到的;
  - 计算单元, 根据噪声模型的可加性噪声和附加了噪声的声音计算雅可比
  - 10 矩阵;
  - 第二估计单元, 估计包含在讲话声音中的可积性失真的第二估计值;
  - 第一运算单元, 通过将在非讲话期间生成的可加性噪声、噪声模型的可加性噪声、和第二估计单元估计的可积性失真的第二估计值合成在一起的处理, 获得噪声变化量, 所述第一运算单元同时将噪声变化量与计算单元计算的雅可比矩阵相乘;
  - 15 自适应模型, 通过将第一运算单元获得的相乘结果与合成声音模型合成在一起的处理生成; 和
  - 第二运算单元, 通过将讲话声音与第二估计单元估计的可积性失真的第二估计值合成在一起的处理, 生成观察值序列;
  - 20 其中, 声音识别是通过将观察值序列与自适应模型相对照实现的。
2. 根据权利要求 1 所述的声音识别系统, 其中, 第一估计单元通过从倒频谱域中非讲话期间生成的可加性噪声中减去噪声模型的可加性噪声和由第二估计单元估计的可积性失真的第二估计值, 实现合成处理。
3. 一种声音识别系统, 包括:
  - 25 声音模型, 从事先采集的、包含可积性失真的声音中生成;
  - 噪声模型, 从事先采集的可加性噪声中生成;
  - 第一估计单元, 估计包含在声音模型中的可积性失真的第一估计值;
  - 第二估计单元, 根据对讲话声音的识别结果, 估计声音模型中包含随讲话声音变化的误差的可积性失真的第二估计值;
  - 30 第一运算单元, 将噪声模型的可加性噪声与第一估计单元估计的可积性失真值合成在一起;

合成声音模型，其根据通过将第一运算单元的合成结果与声音模型合成在一起的处理所得的附加了噪声的声音而生成；

计算单元，从第一运算单元所得的合成结果和附加了噪声的声音中计算雅可比矩阵；

- 5        第三估计单元，估计真正讲话声音中包含随真正讲话声音变化的误差的可积性失真的第三估计值；

- 第二运算单元，通过将在非讲话期间生成的可加性噪声、第一运算单元所得的合成结果、第二估计单元估计的第二估计值、和第三估计单元估计的第三估计值合成在一起的处理，获取噪声变化量，同时第二运算单元将噪声变化量与计算单元计算的雅可比矩阵相乘；
- 10

      自适应模型，通过将第二运算单元获得的相乘结果与合成声音模型合成在一起的处理生成；和

      第三运算单元，通过将真正讲话声音与第二估计单元估计的第二估计值和第三估计单元估计的第三估计值合成在一起的处理，生成观察值序列；

- 15        其中，声音识别是通过将观察值序列与自适应模型相对照实现的。

      4. 根据权利要求3所述的语音识别系统，其中，第二运算单元通过从倒频谱域中非讲话期间生成的可加性噪声中加上第二估计单元估计的第二估计值，减去第一运算单元产生的合成结果以及第三估计单元估计的第三估计值，实现合成处理。

- 20        5. 一种语音识别方法，包括下列步骤：

      从事先采集的和包含可积性失真的声音中生成声音模型；

      从事先采集的可加性噪声中生成噪声模型；

      估计包含在声音模型中的可积性失真的第一估计值；

- 将声音模型、可积性失真的第一估计值和噪声模型的可加性噪声合成在一起，以从附加了噪声的声音中生成合成声音模型；
- 25

      根据噪声模型的可加性噪声和附加了噪声的声音，计算雅可比矩阵；

      估计包含在讲话声音中的可积性失真的第二估计值；

      将在非讲话期间生成的可加性噪声、噪声模型的可加性噪声、和第二估计值合成在一起，获得噪声变化量，

- 30        将噪声变化量与雅可比矩阵相乘；

      将在相乘步骤中获得的相乘结果与合成声音模型合成在一起，生成自适

应模型；和

将讲话声音与第二估计值合成在一起，生成观察值序列；

将观察值序列与自适应模型相对照，以实现声音识别。

- 5 6. 根据权利要求5所述的聲音识别方法，其中，合成步骤从倒频谱域中非讲话期间产生的可加性噪声中减去噪声模型的可加性噪声和由第二估计单元估计的可积性失真的第二估计值。

7. 一种声音识别方法，包括下列步骤：

从事先采集的和包含可积性失真的声音中生成声音模型；

从事先采集的可加性噪声中生成噪声模型；

- 10 估计包含在声音模型中的可积性失真的第一估计值；

根据对讲话声音的识别结果，估计声音模型中包含随讲话声音变化的误差的可积性失真的第二估计值；

第一合成步骤，将噪声模型的可加性噪声与可积性失真的第一估计值合成在一起；

- 15 第二合成步骤，将第一合成步骤中所得的合成结果与声音模型合成在一起，以从附加了噪声的声音中生成合成声音模型；

从第一合成步骤中所得的合成结果和附加了噪声的声音中计算雅可比矩阵；

- 20 估计真正讲话声音中包含随真正讲话声音变化的误差的可积性失真的第三估计值；

第三合成步骤，将在非讲话期间生成的可加性噪声、第一合成步骤所得的合成结果、第二估计值、和第三估计值合成在一起，获取噪声变化量，

将噪声变化量与雅可比矩阵相乘；

- 25 将在相乘步骤中所得的相乘结果与合成声音模型合成在一起，生成自适应模型；和

将真正讲话声音、第二估计值和第三估计值合成在一起，生成观察值序列；

将观察值序列与自适应模型相对照，实现声音识别。

- 30 8. 根据权利要求7所述的聲音识别方法，其中，第三合成步骤通过从倒频谱域中非讲话期间生成的可加性噪声中加上第二估计值，减去第一合成结果以及第三估计值。

## 声音识别系统

5

## 技术领域

本发明涉及传输系统等中抗噪声和失真干扰的声音识别系统。

## 背景技术

10

在传统技术中，在例如包含在汽车中的导航仪器等那样的电子设备中，能够进行人机通信的声音识别系统已经引起人们注意。如图 3 所示，根据信息处理算法构成的声音识别系统是众所周知的。

声音识别系统首先利用隐式马尔可夫模型 (Hidden Markov Model, HMM) 生成由字或子字 (音素、音节等) 组成的声学模型 (声音 HMM)，当发出要识别的声音  $R_a$  时生成相对于讲话声音  $R_a$  而言是倒频谱的时间序列的观察值序列  $R_a(\text{cep})$ ，将观察值序列  $R_a(\text{cep})$  与声音 HMM 相对照，选择给出最大似然的 HMM，并输出它作为识别结果。

更详细地说，声音识别系统配置了遵照 HMM 方法生成上述声音 HMM 的声音 HMM 生成部分 5，该声音 HMM 生成部分 5 包括声音数据库 1、逐帧分割单元 2、倒频谱运算单元 3、和练习 (training) 单元 4。

逐帧分割单元 2 将已经经过实验采集的和存储在声音数据库 1 中的、测试对象的大量声音数据划分成每一个由 10 至 20 msec 左右组成的帧，倒频谱运算单元 3 运算各个用倒频谱表示的逐帧数据，从而获得倒频谱的时间序列  $R_m(\text{cep})$ 。

接着，练习单元 4 进行处理，以练习作为声音的特征 (特征矢量) 的倒频谱的时间序列  $R_m(\text{cep})$ ，并使其反映到声学模型 (声音 HMM) 的参数上，从而事先生成由字或子字组成的声音 HMM 6。

当真正进行讲话时，逐帧分割单元 7 象逐帧分割单元 2 那样，将讲话声音的数据  $R_a$  划分成供输入用的帧，倒频谱运算单元 8 一个接着一个地运算各个逐帧讲话声音数据，从而生成作为倒频谱时间序列的观察值序列  $R_a(\text{cep})$ 。

并且,对照单元 9 将观察值序列  $Ra(cep)$  与用字或子字表示的声音 HMM 相对照,输出与观察值序列  $Ra(cep)$  相比具有最大似然的声音 HMM,作为声音识别的结果。

然而,在图 3 所示的声音识别系统中,当采集声音数据  $R_m$  以生成声音 HMM 6 时,会采集已经受到麦克风和电子传输系统中受可积性失真影响的声音数据  $R_m$ ,而其中的问题在于,难以满意地生成准确的声音 HMM 6。

另外,当发出要识别的讲话声音  $R_a$  时,诸如室内噪声、背景噪声等的可加性噪声,和诸如从嘴巴到麦克风的传输特性、麦克风和电子传输系统中的传输特性等的可积性失真,都对观察值序列  $Ra(cep)$  产生负面影响,其中,这样的问题在于声音识别率降低了。

为了解决这些和其它问题,构筑几乎不受可加性噪声和可积性失真影响的声音识别系统,即,抗干扰声音识别系统是基本课题。

为了对付上述题目,本发明已经作了尝试,通过将 HMM 合成法应用于可加性噪声和将倒频谱平均归一化法(cepstrum means normalization method, CMN)应用于可积性失真,实现抗干扰声音识别系统。

图 4 是显示声音识别系统结构的方块图。声音识别系统配置了声音 HMM 10、初始噪声 HMM 17、初始合成 HMM 16 和自适应 HMM 26,其中,当发出要识别的声音时,由对照单元 29 以字或子字为单位将已经通过发出的声音获得的、作为倒频谱时间序列的观察值序列  $RNa(cep)$  与自适应 HMM 26 相互对照,并输出相对于观察值序列  $RNa(cep)$  具有最大似然的自适应 HMM,作为声音识别的结果。

并且,由于如果应用 HMM 合成法的话,就会增加计算量,因此,为了通过减少运算量而实现高速处理,应用基于泰勒展开的模型自适应法。也就是说,通过提供雅可比矩阵计算单元 19 来试图减少运算量,雅可比矩阵计算单元 19 计算所谓“雅可比矩阵  $J$ ”的泰勒展开一阶导数矩阵。

上述声音 HMM 10 是通过 HMM 方法,利用采集的并不包括任何可加性噪声的讲话声音  $R_m$  事先生成的声学模型。也就是说,声音 HMM 10 是通过根据与图 3 所示的声音 HMM 生成部分 5 的方法相似的 HMM 方法进行处理先生成的。

此外,通过实验采集消声室中测试对象的讲话声音  $R_m$ ,生成不受可加性噪声影响的声音 HMM 10。但是,由于无法消除麦克风和电子传输系统等中可积性失真带来的影响,声音 HMM 10 成为可积性失真带来的影响仍然存在的声

学模型。

因此，在假定经实验采集的讲话声音  $R_m$  由纯声音  $S_m$ （不包含任何可加性噪声和可积性失真的声音）和可积性失真  $H_m$  组成的情况下，如果用线性谱域（lin）表示讲话声音  $R_m$ ，那么，它可以用纯声音  $S_m$  与可积性失真  $H_m$  相乘所得的线性谱积来表示，即， $R_m(\text{lin})=S_m(\text{lin})H_m(\text{lin})$ 。此外，如果用倒频域（cep）表示讲话声音  $R_m$ ，那么，它可以用纯声音  $S_m$  与可积性失真  $H_m$  相加所得的倒频谱和来表示，即， $R_m(\text{cep})=S_m(\text{cep})+H_m(\text{cep})$ 。

并且，如果用线性谱域（lin）表示声音 HMM 10，那么，它可以用  $R_m(\text{lin})=S_m(\text{lin})H_m(\text{lin})$  来表示。而如果用倒频域（cep）表示声音 HMM 10，那么，它可以用  $R_m(\text{cep})=S_m(\text{cep})+H_m(\text{cep})$  来表示。

上述初始噪声 HMM 17 是将非讲话期间的声音（对应于可加性噪声）收集起来作为初始噪声数据  $N_m$ ，并利用初始噪声数据  $N_m$  训练的声学模型，初始噪声 HMM 17 是通过与图 3 所示的声音 HMM 生成部分 5 中的处理相似的处理先生成的。因此，如果用线性谱域（lin）表示初始噪声 HMM 17，它就变成  $N_m(\text{lin})$ ，而如果用倒频域（cep）表示初始噪声 HMM 17，它就变成  $N_m(\text{cep})$ 。

初始合成 HMM 16 是通过下列处理生成的。

将倒频域（cep）中的声音（声学模型） $R_m(\text{cep})=S_m(\text{cep})+H_m(\text{cep})$  从声音 HMM 10 提供到平均计算单元 11 和减法器 12，平均计算单元 11 通过对用于训练声学模型的声音数据库中的特征矢量求平均，和利用 CMN 法对声音 HMM 的平均矢量求平均，获得可积性失真的估计值  $H_m^{\wedge}(\text{cep})$ ，并将其提供给减法器 12。因此，在减法器 12 中进行  $R_m(\text{cep})-H_m^{\wedge}(\text{cep})$  的运算，减法器 12 输出消除了可积性失真的估计值  $H_m^{\wedge}(\text{cep})$  的声音  $S_m'(\text{cep})$ 。

这里，通过进行估计值  $H_m^{\wedge}(\text{cep})$  几乎等于可积性失真  $H_m(\text{cep})$  的近似，假设已经获得了排除了任何可积性失真干扰的声音  $S_m'(\text{cep})$ 。

接着，倒频谱逆变换单元 13 将倒频域中的声音  $S_m'(\text{cep})$  变换成线性谱域中的声音  $S_m'(\text{lin})$ ，并将其提供给加法器 14，同时，倒频谱逆变换单元 13 将从初始噪声 HMM 17 中输出的、倒频域中的初始噪声  $N_m(\text{cep})$  变换成线性谱域中的初始声音  $N_m(\text{lin})$ ，并将其提供给加法器 14，因而，加法器 14 通过将声音  $S_m'(\text{lin})$  与线性谱域中的初始噪声  $N_m(\text{lin})$  相加，生成叠加了可加性噪声的声音  $R_m'(\text{lin})=S_m'(\text{lin})+N_m(\text{lin})$ ，并将其提供给倒频谱变换单元 15。

并且，倒频谱变换单元 15 将附加了可加性噪声的声音  $Rm'(lin)$  变换成倒频谱域中的附加了可加性噪声的声音  $Rm'(cep)$ ，并生成初始合成 HMM 16。

于是，使初始合成 HMM 16 变成具有附加了可加性噪声的声音  $Rm'(cep)$  的特征的声学模型。该声学模型表示如下：

5

$$\begin{aligned} Rm'(cep) &= cep[cep^{-1}[Sm(cep) + Hm(cep) - Hm^{\wedge}(cep)] + Nm(lin)] \\ &= cep[Sm'(lin) + Nm(lin)] \\ &\doteq cep[Sm(lin) + Nm(lin)] \quad \dots \dots (1) \end{aligned}$$

10 此外，在上述表达式中， $cep[]$  表示在倒频谱变换单元 15 中实现的倒频谱变换，和  $cep^{-1}[]$  表示由倒频谱逆变换单元 13 和 18 实现的倒频谱逆变换。

接着，对雅可比矩阵计算单元 19 加以描述。如上所述，提供雅可比矩阵计算单元 19 是为了减少计算量。在假设实际使用环境中的可加性噪声  $Na(cep)$  与初始噪声 HMM 17 中的初始噪声  $Nm(cep)$  之间的变化量  $\Delta Nm(cep) = Na(cep) - Nm(cep)$  很微小的情况下，通过泰勒展开获得合成模型与噪声谱的变化量  $\Delta Nm(cep)$  相对应的变化量，其中，依照获得的变化量来补偿初始合成模型 16。并且，使通过补偿获得的声学模型变成自适应 HMM 26。

更详细地说，线性谱表示如下：

20 
$$Rm(lin) = Sm(lin) + Nm(lin) \quad \dots \dots (2)$$

此处， $Sm(lin)$  是不包括可积性失真和可加性噪声的纯声音  $Sm$  的线性谱； $Rm(lin)$  是不包括可积性失真，但包括可加性噪声的声音  $Rm$  的线性谱；和

25  $Nm(lin)$  是可加性噪声  $Nm$  的线性谱。

此外，如果用倒频谱域中的术语表示包括可加性噪声的声音  $Rm$ ，那么，下列表达式成立：

$$Rm(cep) = IDCT[\log(\exp(DCT[Sm(cep)]) + \exp(DCT[Nm(cep)]))] \quad \dots \dots (3)$$

30

其中， $IDCT[]$  是离散余弦逆变换， $DCT[]$  是离散余弦变换， $\log()$  是

对数变换, 和  $\exp()$  是指数变换。

假设纯声音  $S_m$  不变化而在实际讲话环境中可加性噪声在  $N_m$  到  $N_a$  范围内变化, 初始合成模型  $\Delta R_m(cep)$  中的变化量, 即作为包括  $N_a$  的声音的  $R_{mc}(cep)$  和作为包括倒频谱域中  $N_m$  的声音的  $R_m(cep)$  之间的差值, 能够近似地从表达式 (3) 的泰勒展开的一阶导数项得到, 如下面的表达式 (4) 所示。

$$\Delta R_m(cep) = \frac{\partial R_m(cep)}{\partial N_m(cep)} \Delta N_m(cep) = J(\Delta N_m(cep)) \quad \dots \dots (4)$$

其中  $\partial R_m(cep)/\partial N_m(cep)$  是雅可比矩阵, 并且  $\Delta N_m(cep) = N_a(cep) - N_m(cep)$  是在实际讲话环境中的可加性噪声与在倒频谱域中的初始噪声中的可加性噪声之间的差值。

表达式 (4) 还可以如下面的表达式 (5) 所示来表达。

$$\begin{aligned} R_{mc}(cep) &= R_m(cep) + \frac{\partial R_m(cep)}{\partial N_m(cep)} (N_a(cep) - N_m(cep)) \\ &= IDCT[\log(\exp(DCT[S_m(cep)]) + \exp(DCT[N_a(cep)]))] \quad \dots \dots (5) \end{aligned}$$

雅可比矩阵的第  $I$  行和第  $J$  列的元素通过下面的表达式 (6) 来计算。

$$[J]_{ij} = \sum_{k=1}^P \frac{R_m'(cep)_k}{N_m(cep)_k} F_{ik}^{-1} F_{kj} \quad \dots \dots (6)$$

其中  $F_{kj}$  是余弦变换矩阵的第  $k$  行第  $j$  列的元素, 而  $F_{ik}^{-1}$  是余弦逆变换矩阵的第  $i$  行第  $k$  列的元素。

因此, 雅可比矩阵计算单元 19 能够通过利用从加法器 14 中接收的在线性谱域中的叠加了可加性噪声的声音  $R_m(lin)$  和从倒频谱逆变换单元 18 接收的在线性谱域中的初始噪声  $N_m(lin)$ , 依照表达式 (6) 事先计算出雅可比矩阵。

根据在实际讲话环境中生成的可加性噪声可以自适应地补偿初始合成 HMM 16。初始合成模型的变化量能够通过将可加性噪声之间的变化量  $\Delta N_m(cep)$

与雅可比矩阵  $J$  相乘而获得。因此，可以通过将合成模型中的变化量附加到初始合成模型中生成自适应模型。

接着，对生成自适应 HMM 26 的处理加以描述。

随着用户打开配备在声音识别系统中的讲话开始开关（未示出），麦克风（未示出）采集讲话声音，并且逐帧分割单元 20 以指定的时间间隔为单位分割讲话声音的数据  $R_a$ 。进一步，倒频谱运算单元 21 将数据  $R_a$  变成倒频谱域（cep）中的讲话声音数据  $R_a(\text{cep})$ 。

首先，随着用户打开上述讲话开始开关，转换单元 22 在实际开始讲话之前的非讲话期间内切换到触点“a”一侧。因此，在用户试图讲话的环境中背景噪声（可加性噪声） $N_a$  的倒频谱  $N_a(\text{cep})$  通过转换单元 22 提供给减法器 23。

减法器 23 从背景噪声  $N_a$  的倒频谱  $N_a(\text{cep})$  中减去初始噪声  $N_m$  的倒频谱  $N_m(\text{cep})$ ，并将相减的结果  $N_a(\text{cep}) - N_m(\text{cep})$  提供给乘法器 24，其中乘法器 24 将上述相减结果  $N_a(\text{cep}) - N_m(\text{cep})$  与雅可比矩阵  $J$  相乘，并将相乘结果  $J[N_a(\text{cep}) - N_m(\text{cep})]$  提供给加法器 25。加法器 25 以字或子字为单位将相乘结果  $J[N_a(\text{cep}) - N_m(\text{cep})]$  加入初始合成 HMM 16 的声学模型  $R_m^{\wedge}(\text{cep})$  中，从而，生成已经得到实际讲话环境中的背景噪声  $N_a$  自适应补偿的自适应 HMM 26。也就是说，如果用倒频谱域（cep）中的术语表示自适应 HMM 26，那么，下列表达式成立：

20

$$\begin{aligned} Radp(\text{cep}) &= Rm'(\text{cep}) + J[N_a(\text{cep}) - N_m(\text{cep})] \\ &\doteq \text{cep}[Sm(\text{lin}) + Na(\text{lin})] \quad \dots \dots (7) \end{aligned}$$

此外，在表达式（7）中， $\text{cep}[\ ]$  表示倒频谱变换。

因此，当生成自适应 HMM 26 生成时，转换单元 22 切换到触点“b”一侧，并输入要识别的讲话声音  $R_a$ ，作为倒频谱域中的讲话声音  $R_a(\text{cep})$ 。这里，如果假设讲话声音  $R_a(\text{cep})$  包含纯声音  $S_a$ 、可积性失真  $H_a$ 、和可加性噪声  $N_a$  的线性谱的  $S_a(\text{lin})$ 、 $H_a(\text{lin})$  和  $N_a(\text{lin})$ ，那么，倒频谱域中的讲话声音  $R_a(\text{cep})$  表示成：

30

$$Ra(\text{cep}) = \text{cep}[Sa(\text{lin})Ha(\text{lin}) + Na(\text{lin})]$$

因此, 平均计算单元 27 通过 CMN 法获得可积性失真  $Ha(cep)$  的估计值  $Ha^{\wedge}(cep)$ , 减法器 28 从讲话声音  $Ra(cep)$  中减去估计值  $Ha^{\wedge}(cep)$ , 其中, 将相减的结果  $Ra(cep) - Ha^{\wedge}(cep)$  提供给对照单元 29, 作为观察值序列  $RNa(cep)$ 。

并且, 对照单元 29 以字或子字为单位将观察值序列  $RNa(cep)$  与自适应 HMM 26 相对照, 并输出相对于观察值序列  $RNa(cep)$  具有最大似然的自适应 HMM, 作为识别的结果。也就是说, 观察值序列  $RNa(cep)$  可以用下列表达式来表示。

$$\begin{aligned}
 RNa(cep) &= Ra(cep) - Ha^{\wedge}(cep) \\
 &= cep \left[ \frac{Sa(lin)Ha(lin)}{Ha^{\wedge}(lin)} + \frac{Na(lin)}{Ha^{\wedge}(lin)} \right] \\
 &\cong cep \left[ Sa(lin) + \frac{Na(lin)}{Ha^{\wedge}(lin)} \right] \quad \dots\dots (8)
 \end{aligned}$$

通过将上述表达式 (8) 所表示的、观察值序列的特征矢量  $RNa(cep)$  与上述表达式 (7) 所示的自适应 HMM 26  $Radp(cep)$  的那些特征矢量相对照, 实现声音识别。

但是, 在如图 4 所示的、本发明人提出的声音识别系统中, 通过将自适应 HMM 26 与讲话声音的观察值序列  $Radp(cep)$  相对照, 实现声音识别。然而, 问题在于还没有建立起作为对应观察值序列  $RNa(cep)$  的充分模型的自适应 HMM 26。

也就是说, 在将上述表达式 (7) 与上述表达式 (8) 相比较的情况下, 自适应 HMM 26 的特征在于, 将可加性噪声的线性谱  $Na(lin)$  加入纯声音的线性谱  $Sm(lin)$  当中, 并将其转换成倒频谱。但是, 观察值序列  $RNa(cep)$  的特征在于, 将可加性噪声的线性谱  $Na(lin)$  与可积性失真的线性谱  $Ha^{\wedge}(lin)$  之比  $Na(lin) / Ha^{\wedge}(lin)$  加入纯声音的线性谱  $Sm(lin)$  当中, 并将其变换倒倒频谱域中。

因此, 自适应 HMM 26 不是那种可以完全消除可积性失真带来的影响的模型。这样, 在对照单元 29 将自适应 HMM 26 与观察值序列  $RNa(cep)$  相对照的

情况下,会出现自适应 HMM 26 不能完全模仿观察值序列  $RNa(cep)$  的情况。最后,存在着无法提高声音识别率的问题。

## 发明内容

5

本发明的目的是克服这些和其它问题,为此,提供一种抗可加性噪声和可积性失真干扰的声音识别系统。

10 为了实现上述目的,根据本发明的声音识别系统包括:声音模型,从事先采集的和包含可积性失真的声音中生成;噪声模型,从事先采集的可加性噪声中生成;第一估计单元,估计包含在声音模型中的可积性失真;合成声音模型,根据附加了噪声的声音生成,所述附加了噪声的声音是通过将声音模型、第一估计单元估计的可积性失真的第一估计值和噪声模型的可加性噪声合成在一起的处理获得的;计算单元,根据噪声模型的可加性噪声和附加了噪声的声音计算雅可比矩阵;第二估计单元,估计包含在讲话声音中的可积性失真;第一运算单元,通过将在非讲话期间生成的可加性噪声、噪声模型的可加性噪声和第二估计单元估计的可积性失真的第二估计值合成在一起的处理,获得噪声变化量,同时将噪声变化量与计算单元计算的雅可比矩阵相乘;自适应模型,通过将第一运算单元获得的相乘结果与合成声音模型合成在一起的处理生成;和第二运算单元,通过将讲话声音、第二估计单元估计的可积性失真的第二估计值合成在一起的处理,生成观察值序列;其中,声音识别是通过将观察值序列与自适应模型相对照实现的。

25 根据如上所述构成的声音识别系统,通过将含有可积性失真的声音模型、由第一估计单元估计的可积性失真的第一估计值和噪声模型的可加性噪声合成在一起的处理,生成附加了噪声的声音,并从附加了噪声的声音中生成合成声音模型。

此外,通过将在非讲话期间生成的可加性噪声、噪声模型的可加性噪声和第二估计单元估计的噪声(可积性失真)的第二估计值合成在一起的处理,生成噪声变化量。

30 此外,从噪声模型的可加性噪声和附加了噪声的声音(即,通过将含有可积性失真的声音模型、由第一估计单元估计的可积性失真的第一估计值和噪声模型的可加性噪声合成在一起的处理获得的声音)中计算出雅可比矩阵。

通过将噪声变化量与雅可比矩阵相乘的结果与合成声音模型合成在一起的处理，生成自适应模型。也就是说，自适应模型成为含有合成声音模型以及噪声变化量和雅可比矩阵的信息的模型。

当发出声音时，第二运算单元通过将由第二估计单元估计的可积性失真的第二估计值与包括可积性失真和可加性噪声的讲话声音合成在一起的处理，生成观察值序列。

并且，通过将观察值序列与自适应模型相对照，实现声音识别。

这里，与观察值序列相对照的自适应模型包括与包含在观察值序列中的噪声成分相匹配的噪声成分。因此，如果将上述观察值序列与自适应模型相对照，则可以抵消由噪声成分带来的影响，和能产生与将声音模型的纯声音与发出的纯声音相对照所得的效果几乎相同的效果。因此，可以实现抗可积性失真和可加性噪声干扰的合适声音识别系统。

此外，为了实现上述目的，根据本发明的声音识别系统包括：声音模型，从事先采集的包含可积性失真的声音中生成；噪声模型，从事先采集的可加性噪声中生成；第一估计单元，估计包含在声音模型中的可积性失真；第二估计单元，根据识别对讲话声音的结果，估计包含在声音模型中的包含根据讲话声音变化的误差的可积性失真；第一运算单元，将噪声模型的可加性噪声与第一估计单元估计的可积性失真的第一估计值合成在一起；合成声音模型，其根据通过将第一运算单元的合成结果与声音模型合成在一起的处理所得的附加了噪声的声音而生成；计算单元，从第一运算单元所得的合成结果和附加了噪声的声音中计算雅可比矩阵；第三估计单元，估计包含在讲话声音中的包含根据讲话声音变化的误差的可积性失真；第二运算单元，通过将非讲话期间生成的可加性噪声、第一运算单元所得的合成结果、第二估计单元估计的包含根据讲话声音变化的误差的可积性失真的第二估计值和第三估计单元估计的包含根据讲话声音变化的误差的可积性失真的第三估计值合成在一起的处理，获取噪声变化量，同时通过计算单元将噪声变化量与雅可比矩阵相乘；自适应模型，通过将第二运算单元获得的相乘结果与合成声音模型合成在一起的处理生成；和第三运算单元，通过将讲话声音与第三估计单元估计的包含根据讲话声音变化的误差的可积性失真的第三估计值合成在一起的处理，生成观察值序列；其中，声音识别是通过将观察值序列与自适应模型相对照实现的。

根据如上所述构成的声音识别系统，通过将噪声变化量与雅可比矩阵相乘的结果与合成声音模型合成在一起的处理，生成自适应模型。也就是说，自适应模型成为包括合成声音模型、噪声变化量和雅可比矩阵的信息的模型。

并且，当发出声音时，第三估计单元通过将由第二估计单元估计的可积性失真的第三估计值、由第三估计单元估计的可积性失真的第三估计值和包括可积性失真和可加性噪声的讲话声音合成在一起的处理，生成观察值序列，通过将观察值序列与自适应模型相对照，实现声音识别。

此外，为了实现上述目的，根据本发明的声音识别方法，包括下列步骤：从事先采集的和包含可积性失真的声音中生成声音模型；从事先采集的可加性噪声中生成噪声模型；估计包含在声音模型中的可积性失真的第一估计值；根据对讲话声音的识别结果，估计声音模型中包含随讲话声音变化的误差的可积性失真的第二估计值；第一合成步骤，将噪声模型的可加性噪声与可积性失真的第一估计值合成在一起；第二合成步骤，将第一合成步骤中所得的合成结果与声音模型合成在一起，以从附加了噪声的声音中生成合成声音模型；从第一合成步骤中所得的合成结果和附加了噪声的声音中计算雅可比矩阵；估计真正讲话声音中包含随真正讲话声音变化的误差的可积性失真的第三估计值；第三合成步骤，将在非讲话期间生成的可加性噪声、第一合成步骤所得的合成结果、第二估计值、和第三估计值合成在一起，获取噪声变化量，将噪声变化量与雅可比矩阵相乘；将在相乘步骤中所得的相乘结果与合成声音模型合成在一起，生成自适应模型；和将真正讲话声音、第二估计值和第三估计值合成在一起，生成观察值序列；将观察值序列与自适应模型相对照，实现声音识别

这里，与观察值序列相对照的自适应模型包括与包含在观察值序列中的噪声成分相匹配的噪声成分。因此，如果将上述观察值序列与自适应模型相对照，则可以抵消由噪声成分带来的影响，和能产生与将声音模型的纯声音与发出的纯声音相对照所得的效果几乎相同的效果。因此，可以实现抗可积性失真和可加性噪声干扰的合适声音识别系统。

#### 附图说明

图 1 是显示根据第一实施例的声音识别系统的结构的方块图；

图 2 是显示根据第二实施例的声音识别系统的结构的方块图;

图 3 是显示传统技术中的声音识别系统的结构的方块图; 和

图 4 是显示当开发本发明的声音识别系统时, 本发明人设计和研究的声  
音识别系统的结构的方块图。

5

### 具体实施方式

下面参照附图, 对本发明的实施例加以描述。此外, 图 1 是显示根据本  
发明的声音识别系统的第一实施例的结构的方块图, 和图 2 是显示第二实施  
10 例的结构的方块图。

首先, 在描述第一和第二实施例之前, 对参考符号加以描述。

假设用于生成如后所述的声音 HMM 10、时间域中的声音是  $R_m$ , 用于生成  
初始噪声 HMM 17 的在时间域中的可加性噪声是  $N_m$ , 和在实际发出声音中的  
时间域中的讲话声音是  $R_a$ 。

15 另外, 上述声音  $R_m$  是事先采集的, 并由可积性失真  $H_m$  和不包括噪声的

纯声音  $S_m$  组成。此外，讲话声音是  $R_a$  由可积性失真  $H_a$ 、可加性噪声  $N_a$  和不包括噪声的纯声音  $S_a$  组成。另外，当不发出声音时的，诸如室内噪声之类的背景噪声被称为  $N_a$ 。

并且，线性谱域中的信息用 (lin) 表示，倒频谱域中的信息用 (cep) 表示。

- 5 例如，倒频谱域中初始噪声 HMM 17 的声学模型的标记被称为  $N_m(\text{cep})$ ，线性谱域中初始噪声 HMM 17 的声学模型的标记被称为  $N_m(\text{lin})$ 。

### 第一实施例

现在参照图 1，对第一实施例加以描述。另外，在图 1 中，与图 4 中相应部分相同或相似的那些部分给予相同的标号。

- 10 声音识别系统配置了声音 HMM 10、初始噪声 HMM 17、初始合成 HMM 16、用于当产生时显著减少运算或计算量的雅可比矩阵计算单元 19 和自适应 HMM 26，其中，当发出要识别的声音时，由对照单元 29 以字或子字为单位将从发出的声音获得的、作为倒频谱时间序列的观察值序列  $R_{Na}(\text{cep})$  与自适应 HMM 26 相对照，并输出相对于观察值序列  $R_{Na}(\text{cep})$  具有最大似然的自适应 HMM，  
15 作为声音识别的结果。

- 声音 HMM 10 是通过 HMM 方法，利用经过实验采集的和不包括可加性噪声的声音  $R_m$  事先生成的声学模型。但是，虽然通过实验采集消声室中测试对象的声音  $R_m$ ，生成不受可加性噪声带来的影响的声音 HMM 10，但由于无法消除麦克风和电子传输系统等中可积性失真带来的影响，因此，声音 HMM 10 是可  
20 积性失真带来的影响仍然存在的那一种。

- 因此，如果用线性谱域 (lin) 表示声音 HMM 10，那么，它可以用纯声音  $S_m$  和可积性失真  $H_m$  的线性谱  $S_m(\text{lin})$  和  $H_m(\text{lin})$  的乘积  $R_m(\text{lin}) = S_m(\text{lin})H_m(\text{lin})$  来表示，并且如果用倒频谱域 (cep) 表示声音 HMM 10，那么，它可以用纯声音  $S_m$  和可积性失真  $H_m$  的倒频谱  $S_m(\text{cep})$  和  $H_m(\text{cep})$   
25 的和  $R_m(\text{cep}) = S_m(\text{cep}) + H_m(\text{cep})$  来表示。

初始噪声 HMM 17 是通过上述实验将非讲话期间的声音（相当于可加性噪声）收集起来作为初始噪声数据  $N_m$ ，并利用初始噪声数据  $N_m$  事先练习的声学模型。因此，如果用线性谱域 (lin) 表示初始噪声 HMM 17，它就变成  $N_m(\text{lin})$ ，而如果用倒频谱域 (cep) 表示初始噪声 HMM 17，它就变成  $N_m(\text{cep})$ 。

- 30 初始合成 HMM 16 是通过将线性谱域中声音 HMM 10 的声学模型和初始噪声 HMM 17 的声学模型相加，并将相加结果变换成倒频谱域中的声学模型事先

生成的。

也就是说，在发出实际要识别的声音之前，将声音（声学模型） $R_m(cep) = S_m(cep) + H_m(cep)$ 从声音 HMM 10 提供到平均计算单元 11 和减法器 12。平均计算单元 11 利用 CMN 方法获得可积性失真  $H_m(cep)$  的估计值  $\hat{H}_m(cep)$ ，接着，减法器 12 进行  $R_m(cep) - \hat{H}_m(cep)$  的运算，从而生成从中消除了可积性失真的估计值  $\hat{H}_m(cep)$  的声音  $S_m'$ 。另外，倒频谱逆变换单元 13 将倒频谱域中的声音  $S_m'(cep)$  变换成线性谱域中的声音  $S_m'(lin)$ ，同时，倒频谱逆变换单元 18 将来自初始噪声 HMM 17 的、倒频谱域中的初始噪声  $N_m(cep)$ （初始噪声的声学模型）变换成线性谱域中的初始声音  $N_m(lin)$ ，和加法器 14 将线性谱域中的声音  $S_m'(lin)$  与初始噪声  $N_m(lin)$  相加，从而生成附加了可加性噪声的声音  $R_m'(lin) = S_m'(lin) + N_m(lin)$ 。并且，倒频谱变换单元 15 将附加了可加性噪声的声音  $R_m'(lin)$  变换成倒频谱域中的附加了可加性噪声的声音  $R_m'(cep)$ ，其中生成初始合成 HMM 16。

因此，使初始合成 HMM 16 变成了具有如上述表达式（1）所表示的附加了可加性噪声的声音  $R_m'(cep)$  的特征的声学模型。

雅可比矩阵计算单元 19 输入由倒频谱逆变换单元 18 生成的初始噪声  $N_m(lin)$  和由加法器 14 生成的附加了可加性噪声的声音  $R_m'(lin)$ ，并将这些线性谱域中的初始噪声  $N_m(lin)$  和附加了可加性噪声的声音  $R_m'(lin)$  变换成倒频谱域中的那些相应量。通过这些量引入上述表达式（6）中，每个用于子字的合成 HMM 事先生成作为泰勒展开的一阶导数矩阵的雅可比矩阵  $J = \partial R_m'(cep) / \partial N_m(cep)$ 。

自适应 HMM 26 是通过响应于用户打开配备在声音识别系统中的讲话开始开关（未示出），在真正开始讲话之前经麦克风（未示出）采集的并与可加性噪声相对应的背景噪声，自适应地补偿初始合成 HMM 16 而生成的声学模型，其中自适应 HMM 26 是在开始真正的讲话之前事先生成的。

另外，声音识别系统配置了逐帧分割单元 20、倒频谱运算单元 21、转换单元 22、加法器和减法器 23、乘法器 24、加法器 25、平均计算单元 27、和减法器 28。

当通过打开上述讲话开始开关使麦克风进入声音收集状态时，逐帧分割单元 20 接收从麦克风输出的并经过模拟-数字转换的背景噪声  $N_a$ （对应于可加性噪声）和讲话声音  $R_a$  的输入，并在逐帧地分割它们之后，按照指定的

间隔输出。

倒频谱运算单元 21 将逐帧背景噪声  $Na$  和讲话声音  $Ra$  变换成倒频谱, 生成和输出背景噪声  $Na$  和讲话声音  $Ra$  的倒频谱  $Na(cep)$  和  $Ra(cep)$ 。

转换单元 22 在打开上述讲话开始开关真正开始讲话之前的非讲话期间  
5 内切换到触点“a”一侧, 从而, 将背景噪声的倒频谱  $Na(cep)$  提供给加法器和减法器 23 一侧。并且, 转换单元 22 在真正开始讲话期间切换到触点“b”一侧, 从而将讲话声音的倒频谱  $Ra(cep)$  提供给平均计算单元 27 和加法器 28 一侧。

平均计算单元 27 通过 CMN 方法运算几种讲话声音的倒频谱  $Ra(cep)$  以便  
10 获得平均值, 并获得包含在倒频谱  $Ra(cep)$  中的可积性失真  $Ha$  的估计值  $Ha^{\wedge}(cep)$ 。

此外, 平均计算单元 27 配置了诸如半导体存储器之类的存储估计值  $Ha^{\wedge}(cep)$  的存储单元 (未示出), 和每当操作讲话开始开关时, 存储单元输出在最后一次发出声音时已经获得的估计值  $Ha^{\wedge}(cep)$ , 并且利用此时的讲话声  
15 音将估计值  $Ha^{\wedge}(cep)$  更新和存储新的估计值  $Ha^{\wedge}(cep)$ 。

加法器和减法器 23 通过在倒频谱域中进行运算, 从背景噪声  $Na(cep)$  中减去可积性失真  $Ha$  的估计值  $Ha^{\wedge}(cep)$  和初始噪声  $Nm(cep)$ , 从而获得由下列表达式表示的可加性噪声倒频谱变化量  $Namh(cep)$ , 将其提供给乘法器 24。

$$20 \quad Namh(cep) = Na(cep) - Ha^{\wedge}(cep) - Nm(cep) \quad \dots \dots (9)$$

乘法器 24 将可加性噪声倒频谱变化量  $Namh(cep)$  与雅可比矩阵  $J$  相乘, 将相乘结果  $J[Namh(cep)] = J[Na(cep) - Ha^{\wedge}(cep) - Nm(cep)]$  提供给加法器 25。

加法器 25 以字或子字为单位将上述相乘结果  $J[Namh(cep)]$  与初始合成  
25 HMM 16 的声学模型  $Rm'(cep)$  相加, 从而生成已经通过在实际讲话环境中背景噪声  $Na$  自适应补偿的自适应 HMM 16。也就是说, 如果用倒频谱域(cep)中的术语表示自适应 HMM 16, 那么, 下列表达式成立:

$$30 \quad \begin{aligned} Radp'(cep) &= Rm'(cep) + J[Namh(cep)] \\ &= Rm'(cep) + J[Na(cep) - Ha^{\wedge}(cep) - Nm(cep)] \end{aligned}$$

$$= cep \left[ Sm(lin) + \frac{Na(lin)}{Ha^{\wedge}(lin)} \right] \dots\dots (10)$$

此外，上面表达式(10)中的 cep[ ]表示倒频谱变换。

接着，对这样的声音识别系统的操作加以描述。

- 5        首先，如上所述，在开始真正的声音识别之前，事先准备好声音 HMM 10、初始噪声 HMM 17 和初始合成 HMM 16，并且，由雅可比矩阵计算单元 19 准备雅可比矩阵 J。

10        接着，随着用户打开讲话开始开关，麦克风（未示出）进入声音收集状态，转换单元 22 在开始真正讲话之前的非讲话期间内切换到触点“a”一侧，其中输入背景噪声 Na。因此，背景噪声 Na 的倒频谱 Na(cep)通过逐帧分割单元 20、倒频谱运算单元 21 和转换单元 22 提供给加法器和减法器 23。进一步，将来自平均计算单元 27 的、在最后一次讲话中已经获得的、可积性失真的估计值 Ha<sup>^</sup>(cep)给予加法器和减法器 23，并且从初始噪声 HMM 17 提供初始噪声 Nm(cep)。

- 15        而且，加法器和减法器 23 获取上述表达式(9)所表示的可加性噪声倒频谱变化量 Namh(cep)，乘法器 24 将这个可加性噪声倒频谱变化量 Namh(cep)与雅可比矩阵 J 相乘。接着，加法器 25 以字或子字为单位将相乘结果 J[Namh(cep)]与初始合成 HMM 16 相加，从而生成自适应 HMM 26。

20        然后，当自适应 HMM 26 的生成已完成时，转换单元 22 切换到触点“b”一侧。随着用户真正发出声音，逐帧分割单元 20 和倒频谱运算单元 21 使讲话声音 Ra 变成倒频谱域中的讲话声音 Ra(cep)，并通过转换单元 22 将其提供给平均计算单元 27。

25        从而，平均计算单元 27 通过 CMN 方法，重新生成包含在讲话声音 Ra(cep)中的可积性失真 Ha 在倒频谱域中的估计值 Ha<sup>^</sup>(cep)，并存储它。同时，将重新生成的估计值 Ha<sup>^</sup>(cep)提供给加法器和减法器 28。

30        而且加法器和减法器 28 通过从讲话声音 Ra(cep)中减去估计值 Ha<sup>^</sup>(cep)，生成消除了可积性失真的观察值序列 RNa(cep)=Ra(cep) - Ha<sup>^</sup>(cep)，对照单元 29 以字或子字为单元将观察值序列 RNa(cep)与自适应 HMM 26 相对照，其中输出获得了最大似然的自适应 HMM，作为声音识别的结果。

这里，正如上述表达式(8)所示的，观察值序列  $RNa(cep)$  的特征在于，将可加性噪声的线性谱  $Na(lin)$  与可积性失真的线性谱  $Ha^{(lin)}$  之比  $Na(lin)/Ha^{(lin)}$  加入纯声音的线性谱  $Sa(lin)$  当中，并将其转换成倒频谱。另一方面，正如上述表达式(10)所示的，自适应 HMM 26 的特征在于，将可加性噪声的线性谱  $Na(lin)$  与可积性失真的线性谱  $Ha^{(lin)}$  之比  $Na(lin)/Ha^{(lin)}$  加入纯声音的线性谱  $Sm(lin)$  当中，并将其转换成倒频谱。

也就是说，由于使自适应 HMM 26 的噪声谱和观察值序列  $RNa(cep)$  的噪声谱两者都变成了  $Na(lin)/Ha^{(lin)}$ ，因此，自适应 HMM 26 变成与观察值序列  $RNa(cep)$  相匹配的声学模型。

此外，由于当对照单元 29 将观察值序列  $RNa(cep)$  与自适应 HMM 26 相对照时，两个噪声谱都是  $Na(lin)/Ha^{(lin)}$ ，因此，自适应 HMM 26 和观察值序列  $RNa(cep)$  的噪声（可加性噪声和可积性失真）所带来的影响基本上相互抵消，可以将纯声音  $Sm(cep)$  与纯讲话声音  $Sa(cep)$  相对照，其中声音识别率可以得到进一步的提高。

因此，根据本实施例的声音识别系统，可以使自适应 HMM 26 变成与当真正讲话出现时获得的观察值序列  $RNa(cep)$  相匹配的声学模型，并将其构造成能够对付可加性噪声和可积性失真。因此，这种声音识别系统是抗可加性噪声和可积性失真干扰的，并且可以提高声音识别率。此外，由于已经实现了足以应用雅可比矩阵的配置，因此，可以显著地缩短声音识别所需的时间。

## 20 第二实施例

下面参照图 2，对根据第二实施例的声音识别系统加以描述。此外，在图 2 中，给予与图 1 中相应部分相同或相似的那些部分以相同的标号。

在图 2 中，对与第一实施例不同的那些点作如下描述。根据第二实施例的声音识别系统配置了两个平均计算单元 11a 和 11b、加法器 12a、加法器和减法器 23a、和加法器 30。

另外，没有提供图 1 所示的减法器 12。因此，声音 HMM 10 在倒频谱域中的声音  $Rm(cep)$  被直接提供给倒频谱逆变换单元 13。

第一平均计算单元 11a 通过获取声音 HMM 10 的平均矢量的总平均值，获得大量声音 HMM 的平均值，并获得可积性失真的估计值  $Hm1^{(cep)}$ 。

通过对在最后一次讲话中进行  $Ha^{(lin)}$  估计（如后所述）用的讲话声音的范围（相当于过去数次讲话）内的与识别结果相对应的子字声音 HMM 的平均失

量求平均，第二平均计算单元 11b 生成包括真正可积性失真  $Hm(cep)$  和根据讲话声音变化的可积性失真的误差  $Hme(cep)$  的可积性失真的新估计值  $Hm2^{\wedge}(cep) = Hm(cep) + Hme(cep)$ 。

5 加法器 12a 将来自初始噪声 HMM 17 的初始噪声  $Nm(cep)$  与来自第一平均计算单元 11a 的估计值  $Hm1^{\wedge}(cep)$  相加，并将相加结果  $Nmh(cep) = Nm(cep) + Hm1^{\wedge}(cep)$  提供给加法器和减法器 23a 和倒频谱逆变换单元 18。

除了来自上述加法器 12a 的相加结果  $Nmh(cep)$  之外，还将来自第二平均计算单元 11b 的估计值  $Hm2^{\wedge}(cep)$  给予加法器和减法器 23a。并且，还将来自平均计算单元 27 的估计值  $Ha^{\wedge}(cep)$ 、和非讲话期间的背景噪声  $Na(cep)$  给予  
10 加法器和减法器 23a，从而，加法器和减法器 23a 生成如下列表达式 (11) 所表示的、用于生成自适应 HMM 26 的可加性噪声倒频谱变化量  $Namh'(cep)$ 。

$$\begin{aligned} Namh'(cep) &= Na(cep) + Hm2^{\wedge}(cep) - Ha^{\wedge}(cep) - Nmh(cep) \\ &= Na(cep) + Hm2^{\wedge}(cep) - Ha^{\wedge}(cep) - Nm(cep) - Hm1^{\wedge}(cep) \quad \dots \dots (11) \end{aligned}$$

15

如上所示，本实施例的初始合成 HMM 16 是通过由加法器 12a 已经生成的相加结果  $Nmh(cep)$ 、和还没有消除来自声音 HMM 10 的可积性失真  $Hm$  的声音  $Rm(cep)$  事先生成的。也就是说，倒频谱逆变换单元 13 将还没有消除可积性失真  $Hm$  的声音  $Rm(cep)$  变换成线性谱域中的声音  $Rm(lin)$ ，和倒频谱逆变换  
20 单元 18 将上述相加结果  $Nmh(cep)$  变换成线性谱域中的相加结果  $Nmh(lin)$ 。并且，加法器 14 将  $Rm(lin)$  与  $Nmh(lin)$  彼此相加。倒频谱变换单元 15 将通过加法器 14 生成的附加了可加性噪声的声音  $Rmh(lin) = Rm(lin) + Nmh(lin)$  变换成倒频谱域中的附加了可加性噪声的声音  $Rmh(cep)$ ，从而生成初始合成 HMM 16。

25 因此，初始合成 HMM 16 由下列表达式 (12) 表示。

$$\begin{aligned} Rmh(cep) &= Rm(cep) + Nm(cep) + Hm1^{\wedge}(cep) \\ &= Sm(cep) + Hm(cep) + Nm(cep) + Hm1^{\wedge}(cep) \\ &= cep[Rmh(lin)] \\ 30 \quad &= cep[Sm(lin)Hm(lin) + Nm(lin)Hm1^{\wedge}(lin)] \quad \dots \dots (12) \end{aligned}$$

此外，表达式 (12) 中的  $cep[ ]$  表示倒频谱变换。

在根据本实施例的平均计算单元 27 中，如果讲话声音  $Ra$  的倒频谱  $Ra(cep)$  是在讲话期间通过转换单元 22 的触点 “b” 提供的，那么，随讲话声音变化的可积性失真的误差  $Hae(cep)$  包括在附加到  $Ra(cep)$  中的真正可积性失真  $Ha(cep)$  中，可积性失真的估计值由  $Ha^{\wedge}(cep) = Ha(cep) + Hae(cep)$  表示。

因此，图 2 所示的平均计算单元 27 与图 1 所示的第一实施例的平均计算单元 27 不同之处在于，假设上面的  $Ha^{\wedge}(cep)$  等于  $Ha(cep)$ ，而假设本实施例的  $Ha^{\wedge}(cep)$  等于  $Ha(cep) + Hqe(cep)$ ，其中  $Hae$  是随讲话的内容而改变的可积性失真的估计值的误差。如果识别结果是正确的，则使讲话的内容与识别结果相同。当估计  $Hm2^{\wedge}$  时，通过利用识别结果，用于估计  $Ha^{\wedge}$  的讲话内容被使用。因此，建立了  $Hae = Hme$ ，并且从而随讲话声音而变化的可积性失真中的误差能够假设等同于  $Hm2^{\wedge}$  和  $Ha^{\wedge}$ 。

根据本发明的雅可比矩阵计算单元 19 接收由倒频谱逆变换单元 18 生成的噪声  $Nmh(lin)$  和加法器 14 生成的附加了可加性噪声的声音  $Rmh(lin)$ ，并将其引入到上述表达式 (6) 中，其中泰勒展开的一阶导数  $J = \partial Rm(cep) / \partial Nm(cep)$  是作为雅可比矩阵  $J$  事先产生的。

并且，乘法器 24 将雅可比矩阵  $J$  与来自加法器和减法器 23a 的可加性噪声的倒频谱变化量  $Namh'(cep)$  相乘，并将相乘结果  $J[Namh'(cep)]$  提供给加法器 25。进一步，加法器 25 将相乘结果  $J[Namh'(cep)]$  与初始合成 HMM 16 相加，从而生成自适应 HMM 26。

此外，根据上述表达式 (11)，使相乘结果  $J[Namh'(cep)]$  变成  $J[Na(cep) + Hm2^{\wedge}(cep) - Ha^{\wedge}(cep) - Nm(cep) - Hm1^{\wedge}(cep)]$ ，于是，自适应 HMM 26 成为：

$$\begin{aligned} Radp'(cep) &= Rmh(cep) + J[Na(cep) + Hm2^{\wedge}(cep) - Ha^{\wedge}(cep) - Nm(cep) - Hm1^{\wedge}(cep)] \\ &= cep \left[ Sm(lin)Hm(lin) + \frac{Na(lin)Hm2^{\wedge}(lin)}{Ha^{\wedge}(lin)} \right] \end{aligned} \quad \dots \dots (13)$$

加法器 30 将来自第二平均计算单元 11b 的估计值  $Hm2^{\wedge}(cep)$  与由加法器

和减法器 28 提供的讲话声音  $Ra(cep)$  与估计值  $Ha^\wedge(cep)$  之间的差值  $Ra(cep) - Ha^\wedge(cep)$  相加，生成观察值序列  $RNah(cep) = Ra(cep) - Ha^\wedge(cep) + Hm2^\wedge(cep)$ ，并将其提供给对照单元 29。

因此，如表达式 (14) 所示，观察值序列  $RNah(cep)$  成为：

5

$$\begin{aligned}
 RNah(cep) &= Ra(cep) - Ha^\wedge(cep) + Hm2^\wedge(cep) \\
 &= cep \left[ \frac{Sa(lin)Ha(lin)Hm2^\wedge(lin)}{Ha^\wedge(lin)} + \frac{Na(lin)Hm2^\wedge(lin)}{Ha^\wedge(lin)} \right] \\
 &= cep \left[ \frac{Sa(lin)Ha(lin)Hm(lin)Hae(cep)}{Ha(lin)Hme(cep)} + \frac{Na(lin)Hm2^\wedge(lin)}{Ha^\wedge(lin)} \right] \\
 &= cep \left[ Sa(lin)Hm(lin) + \frac{Na(lin)Hm2^\wedge(lin)}{Ha^\wedge(lin)} \right] \quad \dots \dots (14)
 \end{aligned}$$

10

下面对如上所述构成的声音识别系统的操作加以描述。

首先，如上所述，在开始真正的声音识别之前，事先准备好声音 HMM 10、初始噪声 HMM 17 和初始合成 HMM 16，并且，由雅可比矩阵计算单元 19 准备雅可比矩阵 J。

15

随着用户打开讲话开始开关，麦克风（未示出）进入声音收集状态，转换单元 22 在开始真正讲话之前的非讲话期间内切换到触点“a”一侧，其中输入背景噪声  $Na$ 。因此，背景噪声  $Na$  的倒频谱  $Na(cep)$  通过逐帧分割单元 20、倒频谱运算单元 21 和转换单元 22 提供给加法器和减法器 23a。进一步，将平均计算单元 27 在最后一次讲话中已经获得的可积性失真的估计值  $Ha^\wedge(cep)$  提供给加法器和减法器 23a。同时，将来自加法器 12a 的噪声  $Nmh(cep)$  和来自第二平均计算单元 11b 的估计值  $Hm2^\wedge(cep)$  也提供给它。通过对在最后一次讲话中进行  $Ha^\wedge$  估计（如后所述）用的讲话声音范围（相当于过去数次讲话）内的与识别结果相对应的子字声音 HMM 的平均矢量求平均，可以估计出  $Hm2^\wedge(cep)$ 。

20

并且，加法器和减法器 23a 获取上述表达式 (11) 所表示的可加性噪声的倒频谱变化量  $Namh'(cep)$ ，乘法器 24 将这个可加性噪声的倒频谱变化量  $Namh'(cep)$  与雅可比矩阵 J 相乘。接着，加法器 25 以字或子字为单位将相乘结果  $J[Namh'(cep)]$  与初始合成 HMM 16 相加，生成自适应 HMM 26。

然后,当自适应 HMM 26 的生成已完成时,转换单元 22 切换到触点“b”一侧。随着用户真正发出声音,逐帧分割单元 20 将讲话声音  $R_a$  分割成帧,并且倒频谱运算单元 21 使讲话声音  $R_a$  变成倒频谱域中的讲话声音  $R_a(\text{cep})$ ,其中通过转换单元 22 将其提供给平均计算单元 27 和减法器 28。

5 从而,平均计算单元 27 更新和存储可积性失真  $H_a$  在倒频谱域中的估计值  $\hat{H}_a(\text{cep})$ 。在更新之前,将最后一次讲话中估计出来的  $\hat{H}_a(\text{cep})$  提供给加法器和减法器 28。

加法器和减法器 28 通过从讲话声音  $R_a(\text{cep})$  中减去估计值  $\hat{H}_a(\text{cep})$ ,生成消除了可积性失真的观察值序列  $R_{Nah}(\text{cep}) = R_a(\text{cep}) - \hat{H}_a(\text{cep})$ ,对照单元 29 以字或子字为单元将观察值序列  $R_{Nah}(\text{cep})$  与自适应 HMM 26 相对照,其中输出具有最大似然的自适应 HMM,作为声音识别的结果。

这里,观察值序列  $R_{Nah}(\text{cep})$  由上述表达式 (14) 来表示,和自适应 HMM 26 由上述表达式 (13) 来表示。也就是说,由于观察值序列  $R_{Nah}(\text{cep})$  的噪声谱和自适应 HMM 26 的噪声谱变成了  $N_a(\text{lin})H_{m2}^{\wedge}(\text{lin}) / \hat{H}_a^{\wedge}(\text{lin})$ ,因此,15 自适应 HMM 26 变成与观察值序列  $R_{Na}(\text{cep})$  相匹配的声学模型。

此外,当对照单元 29 将观察值序列  $R_{Na}(\text{cep})$  与自适应 HMM 26 相对照时,由于噪声谱  $N_a(\text{lin})H_{m2}^{\wedge}(\text{lin}) / \hat{H}_a^{\wedge}(\text{lin})$  对于  $R_{Na}(\text{cep})$  与自适应 HMM 26 是相同的,所以,可积性失真  $H_m(\text{cep})$  叠加在纯声音  $S_m(\text{cep})$  上面的声音  $S_m(\text{cep})H_m(\text{cep})$  与可积性失真  $H_a(\text{cep})$  叠加在纯声音  $S_a(\text{cep})$  上面的声音  $S_a(\text{cep})H_a(\text{cep})$  相互对照。因此,可以通过基本抵消  $N_a(\text{lin})H_{m2}^{\wedge}(\text{lin}) / \hat{H}_a^{\wedge}(\text{lin})$  的影响来对它们加以对照,并且,可以通过基本抵消由于可积性失真  $H_m(\text{cep})$  和  $H_a(\text{cep})$  带来的影响来对它们加以对照。也就是说,可以基本抵消可加性噪声和可积性失真两者带来的影响,其中,由于可以将纯声音  $S_m(\text{cep})$  与纯声音  $S_a(\text{cep})$  相互对照,因此,可以进一步提高声音识别率。

25 因此,根据本实施例的声音识别系统,可以使自适应 HMM 26 变成与在真正讲话时获得的观察值序列  $R_{Nah}(\text{cep})$  相匹配的声学模型,同时,由于系统被构造成足以对付可加性噪声和可积性失真,因此,这种声音识别系统是抗可加性噪声和可积性失真干扰的,其中声音识别率可以得到提高。此外,由于实现了足以应用雅可比矩阵的配置,因此,可以显著缩短声音识别所需的时间。

在上述第一和第二实施例的描述中,诸如声音与噪声相加和声音与噪声

相减之类的运算是在线性谱域和倒频谱域中实现的，在本说明书中，术语“合成”指的是加法和减法的运算。因此，例如，即使在图 2 中的减法器 28 中的  $\hat{H}a(cep)$  与  $Ra(cep)$  相减也可以描述为“ $Ra(cep)$  和  $\hat{H}a(cep)$  的合成”或“合成  $Ra(cep)$  和  $\hat{H}a(cep)$ ”。

- 5        如上所述，本发明的声音识别系统被构造成能从声音模型和噪声模型中生成合成声音模型，同时，根据用作参考模型的合成声音模型生成事先包括可积性失真和可加性噪声的信息的自适应模型，其中，当真正开始讲话时，
- 10        通过将从讲话声音中生成的观察值序列和自适应模型相对照来实现声音识别，和把自适应模型构造成能与观察值序列相匹配。因此，可以提供抗可积性失真和可加性噪声干扰的和能够充分进行声音识别的声音识别系统。

此外，可以提供最好被构造成通过利用雅可比矩阵在运算量或计算量方面显著减少的声音识别系统，从而可以提供高速声音识别系统。

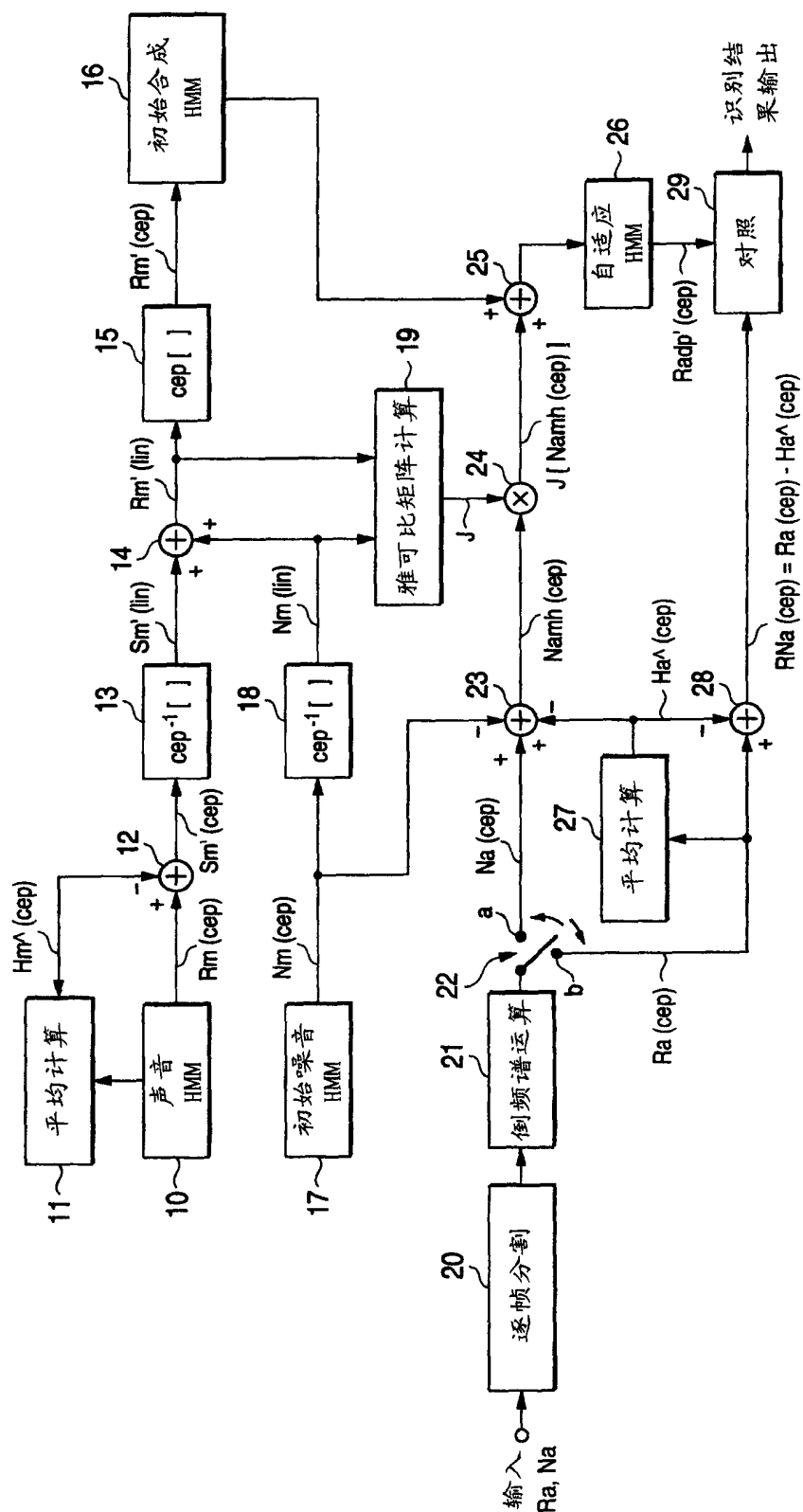


图 1

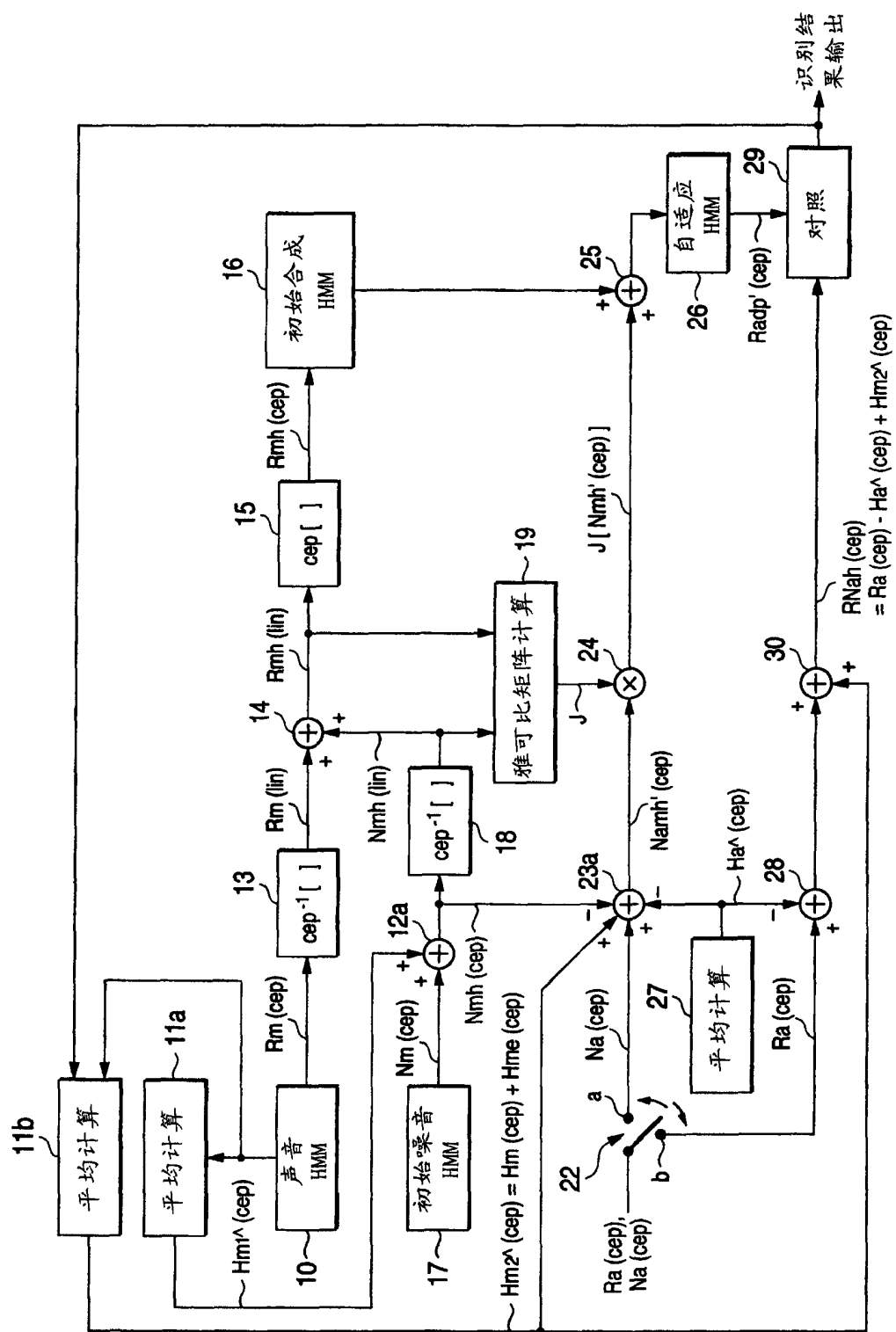


图 2

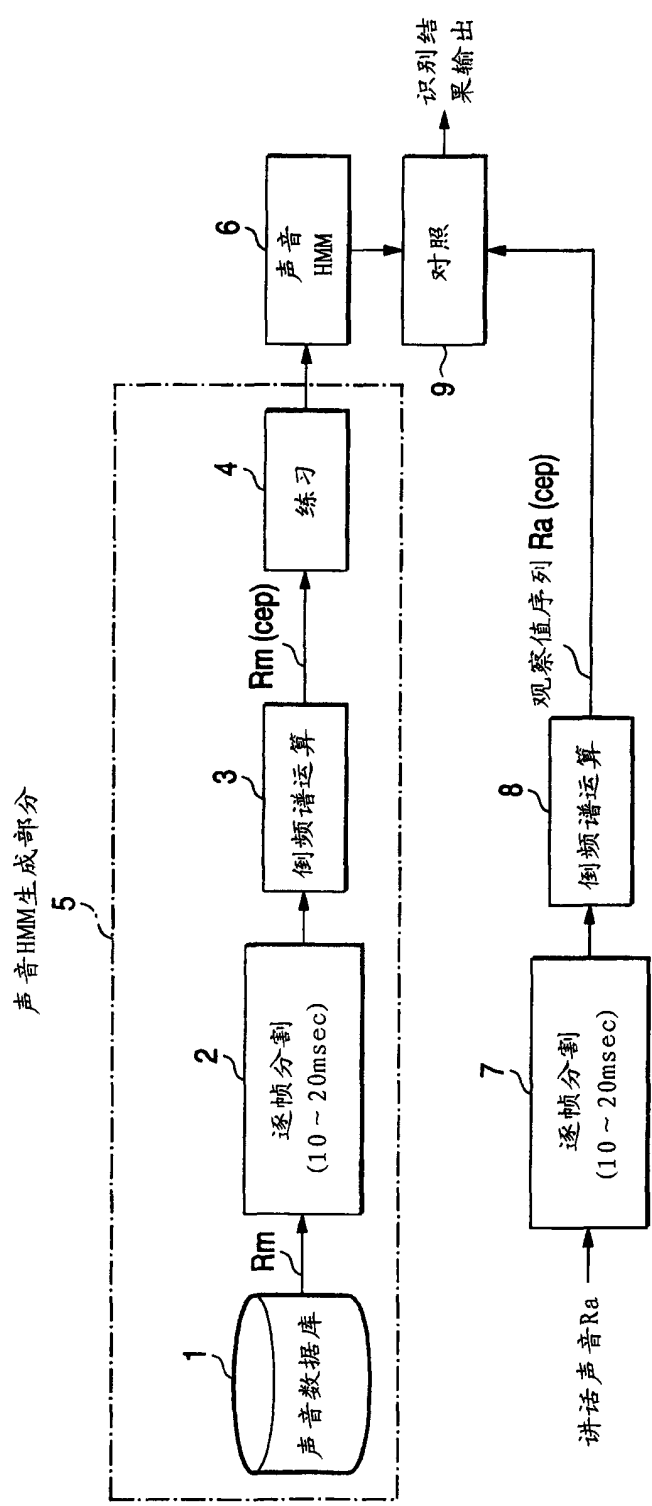
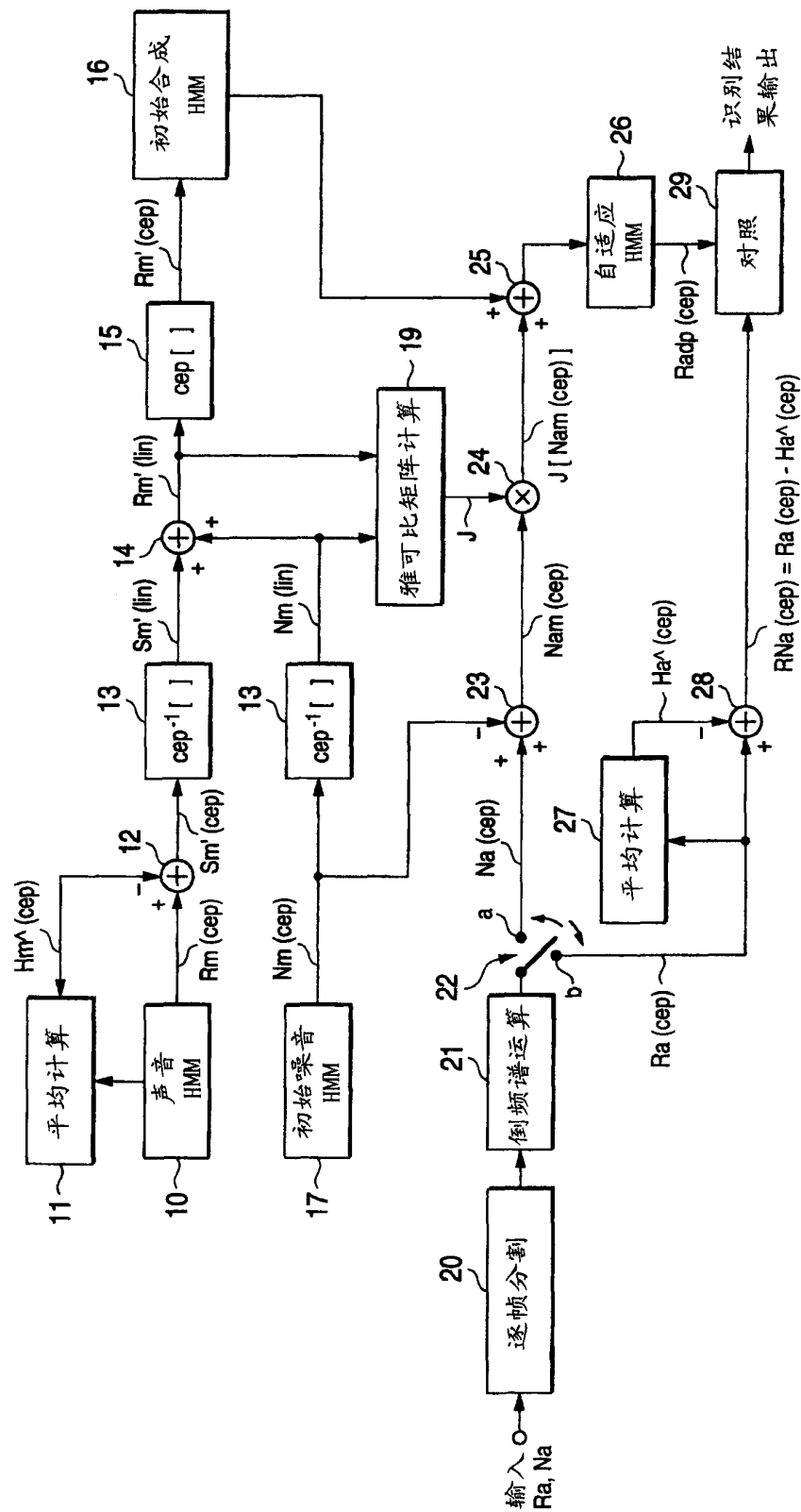


图 3



4  
[X]