[19]中华人民共和国国家知识产权局

[51]Int.Cl[6]
G10L 3/00
G10L 5/06 G10L 7/08
G10L 9/06 G10L 9/18
G10L 5/04

[12]发明专利申请公开说明书

[21]申请号 97195936.6

[43]公开日 1999年7月21日

[11]公开号 CN 1223739A

[22]申请日 97.6.27 [21]申请号 97195936.6
[30]优先权
[32]96.6.28 [33]US [31]08/673,435
[86]国际申请 PCT/US97/11683 97.6.27
[87]国际公布 WO98/00834 英 98.1.8
[85]进入国家阶段日期 98.12.28
[71]申请人 微软公司
地址 美国华盛顿
[72]发明人 晓－文·洪 学东·D·黄 眉－宇·黄
励·蒋 云－成·鞠
米林德·V·马哈贾
米切尔·J·若扎克

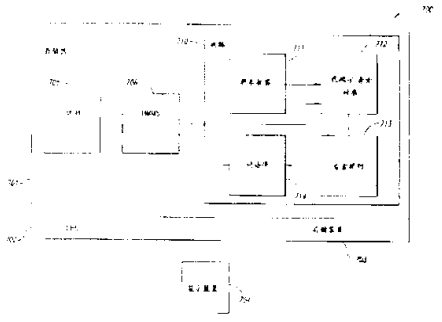
[74]专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 酆 迅

权利要求书 12 页 说明书 11 页 附图页数 11 页

[54]发明名称 用于语音识别的动态调节的训练方法和系统

[57]摘要

一种用来动态选择用来训练语音识别系统的词的方法和系统。语音识别系统使 用隐藏的马尔克夫模型模拟每个音素,并且把每个词表示为音素的一个序列。训练系统根据相应代码字将作为音素的部分读出的概率排列用于每帧的每个音 素。训练系统收集其相应词是已知的读出发音。训练系统然后把每个发音的代 码字与它认为是其部分的音素对准。训练系统然后使用对准帧的对准代码字来 计算对于每个音素的平均等级。最后,训练系统选择包含具有低等级的音素的 词用于训练。



ISSN1008-4274

权 利 要 求 书

1.一种用来动态选择用来训练语音识别系统的词的计算机系统中的方法，该语音识别系统用来识别多个词，该语音识别系统具有组成每个词的音素的指示，该语音识别系统具有用于每个音素的模型，每个模型用来产生代码字的每个可能序列对应于模拟的音素的概率，该方法包括:

对于每个代码字，根据代码字将作为音素的部分读出的概率排列音素;

收集其相应词是已知的多个读出发音;

对于每个收集的发音，

把收集的发音转换成代码字序列；及

根据音素模型，把代码字序列中的每个代码字与收集发音对应的已知词的音素对准;

对于每个音素，

对于在每一个收集发音中与音素对准的所有代码字，累计该音素的等级；及

通过把累计的等级除以在收集发音中与该音素对准的代码字的总数，计算该音素的平均等级;

辨别具有低平均等级的音素；及

选择包含辨别音素的词，作为用来训练语音识别系统的词。

2.根据权利要求 1 所述的方法，其中模型是隐藏马尔克夫模型。

3.根据权利要求 1 所述的方法，其中对准使用一种基于维特比的对准算法。

4.根据权利要求 1 所述的方法，包括把选择的词呈现给讲话者以便训练。

5.根据权利要求 4 所述的方法，其中在词将由讲话者读出的概率下，最好选择选出的词。

6.根据权利要求 1 所述的方法，其中音素的辨别包括辨别多于一个具有低平均等级的音素，并且其中选择过程选择包含多于一个辨别音素的词。

7.根据权利要求 6 所述的方法，包括把选择的词呈现给讲话者以便训练。

8.根据权利要求 7 所述的方法，其中每个词具有指示该词将由讲话者读出的概率的语言模型概率，及其中按基于选择词的语言模型概率的顺序，把选择的词呈现给讲话者。

9.一种用来动态选择用来训练语音识别系统的词的计算机系统中的方法，该语音识别系统用来识别多个词，该语音识别系统具有组成每个词的语音单元的指示，该语音识别系统具有用于每个语音单元的模型，每个模型用来产生特征向量的每个可能序列对应于模拟的语音单元的概率，该方法包括:

收集其相应词是已知的多个读出发音;

对于每个收集的发音，

把收集的发音转换成特征向量序列；及

根据已知词的语音单元的模型，把特征向量序列中的每个特征向量与收集发音所对应的已知词的语音单元对准;

从与每个语音单元对准的特征向量中，辨别哪些语音单元模拟得最不准确；及

选择包含辨别音素的词，作为用来训练语音识别系统的词。

10.根据权利要求 9 所述的方法，其中哪些语音单元模拟得最不准确的辨别包括：根据对准语音单元和特征向量计算帧准确度度量；及通过组合基于该语音单元的帧准确度度量，计算对于每个独特语音单元的组合准确度度量。

11.根据权利要求 10 所述的方法，其中帧准确度度量是特征向量包含在特征向量与其对准的语音单元内的概率，与特征向量包含在任何语音单元内的最大概率的比值。

12.根据权利要求 10 所述的方法，其中帧准确度度量是特征向量包含在特征向量与其对准的语音单元内的概率，与特征向量包含在任何语音单元内的最大概率的差值。

13.根据权利要求 10 所述的方法，其中帧准确度度量是每个语音单元基于这些与语音单元对准的向量帧将作为语音单元的部分读出的概率的等

级。

14.根据权利要求 13 所述的方法，其中与语音单元对准的这些特征向量将作为语音单元的部分读出的概率是声学模型概率。

15.根据权利要求 13 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的平均值。

16.根据权利要求 13 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的最大值。

17.根据权利要求 13 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的最小值。

18.根据权利要求 13 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的加权平均值。

19.根据权利要求 9 所述的方法，其中哪些语音单元模拟得最不准确的辨别是基于在识别期间误识别的词的分析。

20.根据权利要求 9 所述的方法，其中语音单元是一个音素。

21.根据权利要求 9 所述的方法，其中语音单元是一个句素。

22.根据权利要求 9 所述的方法，其中语音单元是一个上下文依赖的音素状态。

23.根据权利要求 9 所述的方法，其中语音单元是一个词。

24.根据权利要求 9 所述的方法，其中特征向量是量化向量。

25.根据权利要求 9 所述的方法，包括根据选择词来训练语音识别系统。

26.根据权利要求 25 所述的方法，其中训练包括产生指示对于每个语音单元用于每个特征向量的概率的、反映在训练期间模型与所读出的词之间差别的误差分布。

27.根据权利要求 26 所述的方法，其中误差分布的概率在语音识别期间分解成模型的概率。

28.一种用来将词的发音教给讲话者的计算机系统中的方法，每个词用语音方法由语音单元表示，每个语音单元具有用来产生特征向量的各序列对应于模拟的语音单元的概率的模型，该方法包括：

收集来自讲话者的其相应词是已知的多个读出发音；

对于每个收集的发音，

把收集的发音转换成特征向量序列；及

根据已知词的语音单元的模型，把特征向量序列中的每个特征向量与收集发音对应的已知词的语音单元对准；

从与每个语音单元对准的特征向量中，辨别讲话者读得不准确的语音单元；及

选择包含辨别音素的词，作为用来教讲话者的词。

29.根据权利要求 28 所述的方法，包括把选择的词呈现给讲话者。

30.根据权利要求 29 所述的方法，包括接收对应于每个读出词的语音发音和估计接收语音发音的准确度。

31.根据权利要求 28 所述的方法，其中哪些语音单元模拟得最不准确的辨别包括：根据对准语音单元和特征向量计算帧准确度度量；及通过组合基于该语音单元的帧准确度度量，计算对于每个独特语音单元的组合准确度度量。

32.根据权利要求 31 所述的方法，其中帧准确度度量是特征向量包含在特征向量与其对准的语音单元内的概率，与特征向量包含在任何语音单元内的最大概率的比值。

33.根据权利要求 31 所述的方法，其中帧准确度度量是特征向量包含在特征向量与其对准的语音单元内的概率，与特征向量包含在任何语音单元内的最大概率的差值。

34.根据权利要求 31 所述的方法，其中帧准确度度量是基于这些与语音单元对准的向量帧将作为语音单元的部分读出的概率的每个语音单元等级。

35.根据权利要求 34 所述的方法，其中与语音单元对准的这些特征向量将作为语音单元的部分读出的概率是声学模型概率。

36.根据权利要求 34 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的平均值。

37.根据权利要求 34 所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的最大值。

38.根据权利要求 34 所述的方法，其中组合准确度量度是与该语音单元

对准的每个特征向量包含在该语音单元内的概率的最小值。

39.根据权利要求34所述的方法，其中组合准确度量度是与该语音单元对准的每个特征向量包含在该语音单元内的概率的加权平均值。

40.根据权利要求28所述的方法，其中语音单元是一个音素。

41.根据权利要求28所述的方法，其中语音单元是一个句素。

42.根据权利要求28所述的方法，其中特征向量是量化向量。

43.一种用来选择用来训练语音识别系统的词的计算机系统中的方法，该语音识别系统用来识别多个词，读出的每个词带有语音单元，该方法包括:

接收对其确定相应词的多个读出发音;

对于接收读出发音的确定词的每一个的每个语音单元，确定语音识别系统在识别确定词内的语音单元时的上下文依赖准确度;

对于每个语音单元，根据上下文依赖准确度确定语音识别系统在识别语音单元时的上下文无关准确度；及

选择包含确定具有最低上下文无关准确度的语音单元的词，用来训练语音识别系统。

44.根据权利要求36所述的方法，其中语音单元是一个音素。

45.根据权利要求36所述的方法，其中语音单元是一个句素。

46.根据权利要求36所述的方法，其中语音单元是一个上下文依赖的音素状态。

47.根据权利要求36所述的方法，其中语音单元是一个词。

48.根据权利要求43所述的方法，其中根据用来训练语音识别系统的上下文依赖准确度来选择词。

49.根据权利要求43所述的方法，其中语音识别系统具有用于每个语音单元、指示量化向量序列对应于语音单元的概率的模型，并且其中上下文依赖准确度的确定包括:

对于每个接收的发音，

把读出的发音转换成量化向量序列;

根据用于确定词的语音单元的模型，把序列中的每个量化向量与所确定词的语音单元对准；及

辨别每个对准量化向量作为与该向量与之对准的语音单元的部分而读出的概率，其中辨别的概率用来确定上下文依赖准确度。

50.根据权利要求 43 所述的方法，包括把选择的词呈现给讲话者以便训练。

51.根据权利要求 50 所述的方法，其中每个词具有指示该词将要读出的概率的语言模型概率，及其中按基于词的语言模型概率的顺序，把选择的词呈现给讲话者。

52.根据权利要求 43 所述的方法，其中词的选择包括选择具有多于一个具有所确定上下文依赖准确度较低的语音单元的词。

53.根据权利要求 52 所述的方法，包括把选择的词呈现给讲话者以便训练。

54.根据权利要求 53 所述的方法，其中每个词具有指示该词将要读出的概率的语言模型概率，及其中按基于选择词的语言模型概率的顺序，把选择的词呈现给讲话者。

55.根据权利要求 43 所述的方法，其中多个读出发音的接收出现在识别过程期间，并且其中识别过程定期要求讲话者在选择的词上训练。

56.根据权利要求 55 所述的方法，其中重复地进行训练和识别。

57.根据权利要求 43 所述的方法，其中多个读出发音的接收出现在识别过程期间，并且其中当识别过程辨别到这时识别过程不正确地识别一定部分的读出发音时，自动提示在选择的词上进行训练。

58.根据权利要求 43 所述的方法，其中多个读出发音的接收出现在识别过程期间，并且其中在识别过程期间误识别接收的读出发音。

59.一种包含用来使计算机系统教讲话者词的发音的指令的计算机可读介质，每个读出的词带有语音单元，该介质的特征在于：

从讲话者接收其相应词是已知的多个读出发音；

由读出发音辨别哪些语音单元由讲话者读得不准确；及

选择包含辨别语音单元的词，用来教讲话者。

60.根据权利要求 59 所述的计算机可读介质，包括把选择的词呈现给讲话者。

61.根据权利要求 60 所述的计算机可读介质，包括接收对应于呈现给讲

话者每个读出词的读出发音、和估计所接收读出发音的准确度。

62.根据权利要求 59 所述的计算机可读介质，其中语音单元是一个音素。

63.根据权利要求 59 所述的计算机可读介质，其中语音单元是一个句素。

64.一种包含用来使计算机系统选择用来训练语音识别系统的词的指令的计算机可读介质，该语音识别系统用来识别多个词，该语音识别系统具有组成每个词的语音单元的指示，该语音识别系统具有用于每个语音单元的模型，每个模型用来指示特征向量的每个可能序列对应于模拟的语音单元的概率，该介质的特征在于:

接收确定相应词的多个读出发音;

对于每个收集的发音，

把收集的发音转换成特征向量序列；及

根据确定词的语音单元模型，把特征向量序列中的每个特征向量与和收集发音对应的 确定词的语音单元对准;

由与每个语音单元对准的特征向量，辨别哪些模拟得最不准确的语音单元；及

选择包含辨别音素的词，作为用来训练语音识别系统的词。

65.根据权利要求 64 所述的计算机可读介质，其中哪些语音单元模拟得最不准确的辨别包括：根据与语音单元对准的那些特征向量将作为语音单元的部分读出的概率，计算每个语音单元的等级。

66.根据权利要求 65 所述的计算机可读介质，其中与语音单元对准的那些特征向量将作为语音单元的部分读出的概率是声学模型概率。

67.根据权利要求 64 所述的计算机可读介质，包括根据选择的词训练语音识别系统。

68.根据权利要求 67 所述的计算机可读介质，其中训练包括产生指示对于每个语音单元用于每个特征向量的概率的、反映在训练期间模型与所读出的词之间差别的误差分布。

69.根据权利要求 68 所述的计算机可读介质，其中误差分布的概率在语音识别期间分解成模型的概率。

70.一种用来动态选择用来训练语音识别系统的词的计算机系统中的方法，每个词包括语音单元，该方法包括:

收集其相应词是已知的多个读出发音;

从读出发音中，辨别哪些语音单元由语音识别系统模拟得最不准确;及

选择包含辨别语音发音的词，用于语音识别系统的训练。

71.根据权利要求 70 所述的方法，其中语音单元是一个音素。

72.根据权利要求 70 所述的方法，其中语音单元是一个句素。

73.根据权利要求 70 所述的方法，其中语音单元是一个上下文依赖的音素状态。

74.根据权利要求 70 所述的方法，其中语音单元是一个词。

75.根据权利要求 70 所述的方法，其中语音识别系统具有用于每个语音单元的指示量化向量序列对应于语音单元的概率的模型，并且其中辨别包括:

对于每个收集的发音，

把收集的发音转换成量化向量序列;

根据用于确定词的语音单元的模型，把序列中的每个量化向量与已知词的语音单元对准;及

辨别每个对准量化向量作为与该向量对准的音素的部分而读出的概率。

76.根据权利要求 70 所述的方法，包括把选择的词呈现给讲话者以便训练。

77.根据权利要求 76 所述的方法，其中每个词具有指示该词将要读出的概率的语言模型概率，及其中按基于词的语言模型概率的顺序，把选择的词呈现给讲话者。

78.根据权利要求 70 所述的方法，其中词的选择包括选择具有多于一个不准确模拟的语音单元的词。

79.根据权利要求 78 所述的方法，包括把选择的词呈现给讲话者以便训练。

80.根据权利要求 79 所述的方法，其中每个词具有指示该词将要读出的

概率的语言模型概率，及其中按基于选择词的语言模型概率的顺序，把选择的词呈现给讲话者。

81.根据权利要求 70 所述的方法，其中多个读出发音的接收在识别过程期间出现，并且其中识别过程定期地要求讲话者在选择词上训练。

82.根据权利要求 81 所述的方法，其中重复地进行训练和识别。

83.根据权利要求 70 所述的方法，其中多个读出发音的接收在识别过程期间出现，并且其中当识别过程辨别到这时识别过程不正确地识别一定部分的读出发音时，在选择的词上进行训练。

84.根据权利要求 70 所述的方法，其中多个读出发音的接收在识别过程期间出现，并且其中在识别过程期间接收的读出发音被误识别。

85.根据权利要求 70 所述的方法，其中哪些语音单元模拟得不准确的辨别包括:

把读出发音的特征向量对准语音单元;

根据对准的语音单元和特征向量计算对于每个特征向量的帧准确度度量; 及

通过组合基于该语音单元的帧准确度度量，计算对于每个独特语音单元的组合准确度度量。

86.根据权利要求 85 所述的方法，其中帧准确度度量是该帧的读出发音包含在读出发音与其对准的语音单元内的概率，与该帧的读出发音包含在任何语音单元内的最大概率的比值。

87.根据权利要求 85 所述的方法，其中帧准确度度量是该帧的读出发音包含在读出发音与其对准的语音单元内的概率，与该帧的读出发音包含在任何语音单元内的最大概率的差值。

88.根据权利要求 85 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的平均值。

89.根据权利要求 85 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的最大值的平均值。

90.根据权利要求 85 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的最小值。

91.根据权利要求 85 所述的方法，其中帧准确度度量是与该语音单元对

准的每个帧的读出发音包含在该语音单元内的概率的加权平均值之和。

92.根据权利要求 70 所述的方法，其中哪些语音单元模拟得不准确的辨别包括计数在识别期间在一个误识别词中识别一个语音单元的次数。

93.根据权利要求 70 所述的方法，其中哪些语音单元模拟得不准确的辨别包括计数在识别期间在一个正确词中未识别一个语音单元的次数。

94.根据权利要求 70 所述的方法，其中哪些语音单元模拟得不准确的辨别是基于正确与不正确音素模型的概率值之间的差。

95.根据权利要求 70 所述的方法，其中语音单元是一个上下文依赖音素。

96.根据权利要求 70 所述的方法，其中语音单元是一个词。

97.根据权利要求 70 所述的方法，其中哪些语音单元模拟得不准确的辨别包括：把读出语音的帧的序列与语音单元对准，和根据帧的对准序列包含在该语音单元中的概率，计算语音单元的等级。

98.一种用来动态选择用来训练语音识别系统的词的计算机系统，每个词包括语音单元，该计算机系统包括：

一个样本收集元件，收集其相应词是已知的多个读出发音，并且把读出发音转换成代码字；

一个对准元件，把代码字与每个词的语音单元对准；

一个语音单元排列元件，由读出的发音辨别哪些语音单元由语音识别系统模拟得不准确；及

一个词选择元件，选择包含辨别发音的词，用来训练语音识别系统。

99.根据权利要求 98 所述的计算机系统，其中语音识别系统具有对于每个语音单元的指示量化向量序列对应于语音单元的概率的模型；其中对准元件根据用于确定词的语音单元的模型，把每个代码字与确定词的语音单元对准；及其中语音单元排列元件辨别每个对准代码字作为代码字与其对准的语音单元的部分读出的概率。

100.根据权利要求 98 所述的计算机系统，其中每个词具有指示该词将要读出的概率的语言模型概率，及包括一个按基于词的语言模型概率的顺序、把选择的词呈现给讲话者的呈现元件。

101.根据权利要求 98 所述的计算机系统，其中词选择元件选择具有多

个模拟得不准确的语音单元的词。

102.一种用来估计识别系统在识别词时的准确度的计算机识别系统中的方法，每个词包括语音单元，该方法包括:

收集其相应词是已知的多个读出发音；及

通过把读出发音的帧与语音单元对准、和根据该帧的读出语音包含在读出语音与之对准的语音单元中的概率来计算用于每帧的帧准确度度量，来辨别每个语音单元的准确度。

103.根据权利要求 102 所述的方法，其中帧准确度度量是该帧的读出发音包含在读出发音与其对准的语音单元内的概率，与该帧的读出发音包含在任何语音单元内的最大概率的比值。

104.根据权利要求 102 所述的方法，其中帧准确度度量是该帧的读出发音包含在读出发音与其对准的语音单元内的概率，与该帧的读出发音包含在任何语音单元内的最大概率的差值。

105.根据权利要求 102 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的平均值。

106.根据权利要求 102 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的最大值。

107.根据权利要求 102 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的最小值。

108.根据权利要求 102 所述的方法，其中帧准确度度量是与该语音单元对准的每个帧的读出发音包含在该语音单元内的概率的加权平均值。

109.根据权利要求 102 所述的方法，其中计算包括计数在识别期间在一个误识别词中未识别一个语音单元的次数。

110.根据权利要求 102 所述的方法，其中计算包括计数在识别期间在一个正确词中未识别一个语音单元的次数。

111.根据权利要求 102 所述的方法，其中语音单元是一个上下文依赖音素。

112.根据权利要求 102 所述的方法，其中语音单元是一个词。

113.根据权利要求 102 所述的方法，其中语音单元是一个音素的状态。

114.根据权利要求 102 所述的方法，其中语音单元是一个句素。

115.一种用来估计识别系统在识别词时的准确度的计算机识别系统中的方法，每个词包括语音单元，该方法包括:

收集其相应词是已知的多个读出发音；及

通过把读出发音的帧与语音单元对准、和计数在识别期间在一个正确词中未识别一个语音单元的次数，来辨别每个语音单元的准确度。

116.一种用来估计识别系统在识别词时的准确度的计算机识别系统中的方法，每个词包括语音单元，该方法包括:

收集其相应词是已知的多个读出发音；及

通过把读出发音的帧与语音单元对准、和计数在识别期间在一个误识别词中未识别一个语音单元的次数，来辨别每个语音单元的准确度。

说　明　书

用于语音识别的动态调节的训练方法和系统

本发明涉及计算机语音识别，更具体地说，涉及训练一种计算机语音识别系统。

借助于计算机系统迅速和准确地识别人类语音早就是计算机系统开发者长期追求的目标。由这样一种计算机语音识别(CSR)系统产生的益处是显著的。例如，与其把文件用键盘打入计算机系统中，倒不如人简单地读出文件的词，并且CSR系统识别词且存储每个词的字母，就象已经用键盘打出词一样。由于人们一般说得比打字快，所以能提高效率。而且，人们不必学习如何打字。计算机还能用于许多这样的用途中，在因为人的手忙于打字之外的任务而无法使用的场合中。

典型的CSR系统借助于把读出的发音与词汇中每个词的模型相比较来识别词。把其模型与发音匹配最好的词识别为讲出的词。CSR系统将每个词看作一个组成该词的音素序列的模型。为了识别发音，CSR系统辨别一个词序列，该序列的音素与发音很好地匹配。然而，这些音素可能与组成词的音素对应得不准确。因而，CSR系统一般使用概率分析，以确定哪个词最接近地对应于辨别的音素。

当识别发音时，CSR系统把代表发音的模拟信号转换成用来进一步处理的更适用形式。CSR系统首先把模拟信号转换成数字形式。CSR系统然后对该数字形式采用信号处理技术，如快速傅里叶变换(FFT)、线性预测编码(LPC)、或滤波器组，以抽取发音的适当参数表示。通常使用的表示是一个带有代表在各间隔(称为“帧”)处的频带和/或能带的FFT或LPC系数的“特征向量”。诸间隔依据计算机系统的计算容量和识别过程的希望准确度可短可长。典型间隔可能在10毫秒的范围内。就是说，CSR系统对于每10毫秒的发音产生一个特征向量。每帧一般25毫秒长。因此，每10毫秒产生一个25毫秒长的帧。在连续的帧之间有重叠。

为了便于特征向量的处理，把每个特征向量量化成有限数量(例如256

个)的“量化向量”之一。就是说，CSR 系统定义多个选择为代表特征向量典型或平均范围的量化向量。CSR 系统然后把每个特征向量与每一个量化向量相比较，并且选择最接近类似于特征向量的量化向量，以表示特征向量。每个量化向量唯一地由一个数辨别(例如在 1 与 256 之间)，这个数称作“代码字”。当一个特征向量表示为量化向量时，有信息丢失，因为多个不同的特征向量映到相同的量化向量上。为了保证这种信息丢失不会严重影响识别，CSR 系统可以定义几千或几百万个量化向量。存储这样大数量的量化向量的定义所需的存储量会相当大。因而，为了减小所需的存储量，CSR 系统将特征向量分段，并且把每段量化成小数量(例如 256 个)量化向量之一。因而，每个特征向量由用于每个段的量化向量(由一个代码字辨别)表示。为了解释简单起见，描述没有将特征向量分段并因而每个特征向量(或帧)仅有一个代码字的 CSR 系统。

如以上讨论的那样，读出的发音经常与词的模型对应得不准确。找到准确对应性的困难归因于语音中的巨大变化，语音不能由词模型完全和准确地捕捉。这些变化例如是由讲话者的音调、人讲话的速度和音高、讲话者的当前健康状况(例如感冒)、讲话者的年龄和性别等等产生的。使用概率技术的 CSR 系统在准确识别语音方面，比寻找准确对应的技术更成功。

通常用于语音识别的这样一种概率技术是隐藏马尔可夫(Markov)模型。CSR 系统可以把隐藏马尔可夫模型(“HMM”)用于词汇中的每个词。用于词的 HMM 包括由其能导出代码字的任何序列对应于该词的概率的概率信息。因而，为了识别发音，CSR 系统把发音转换成一个代码字序列，并且然后把 HMM 用于每个词，以确定词对应于发音的概率。CSR 系统把发音识别为具有最高概率的词。

HMM 由一个状态图表示。状态图传统上用来确定系统接收一个输入序列后处于的状态。状态图包括状态和源状态与目标状态之间的过渡段。每个过渡段与一个输入有关，该输入指示当系统接收到该输入且处于源状态时，系统将过渡到目标状态。这样一种状态图，例如能由识别代码字每个序列的系统使用，这些代码字组成词汇中的词。当系统处理每个代码字时，系统根据当前状态和正在处理的代码字确定下一个状态。在这个例子中，状态图可能具有对应于每个词的一定最终状态。然而，如果表示一个

词的多种发音，那么每个词可能具有多个最终状态。如果在处理代码字之后，系统处于对应于一个词的最终状态，那么能把代码字的序列识别为最终状态的词。

然而，一个 HMM 具有与对于每个代码字从一个状态到另一个状态的每个过渡段有关的概率。例如，如果 HMM 处于状态 2，那么某一代码字导致从当前状态过渡到下一状态的概率可能是.1，而相同代码字导致从当前状态过渡到下一不同状态的概率可能是.2。类似地，另一个不同代码字导致从当前状态过渡到下一状态的概率可能是.1。由于 HMM 具有与其状态图有关的概率，所以对于给定的代码字序列的最终状态的确定仅能用概率表示。因而，为了确定对于一个代码字序列的每个可能最终状态的概率，需要辨别用于 HMM 状态图的状态的每个可能序列，并且需要计算有关的概率。状态的每个这种序列称为状态路径。

为了简化识别，与其使用带有表示对于每个可能词用于代码字的每个可能序列的概率的大状态图的 HMM，倒不如 CSR 系统用 HMM 表示每个可能的语音单元，并且把每个词表示为语音单元序列。传统上，语音单元就是音素。然而，已经使用了其他语音单元，如句素（Senones）。(见 Hwang 等，“用句素预测未知的三音素(Predicting Unseen Triphones with Senones)”，Proc.ICASSP'93，1993 年，卷II，第 311-314 页。)对于用于每个语音单元的 HMM，CSR 系统通过连接用于组成词的音素的 HMM 和估计生成 的 HMM，来估计表示某一词的音素序列的概率。

每个 HMM 包含对于每个状态中每个代码字将导致彼此状态过渡的概率。与每个状态过渡有关的概率由用于该状态的代码字依赖输出概率、和用于状态的代码字无关过渡概率表示。用于状态的代码字依赖输出概率反映在代码字序列导致 HMM 处于该状态之后，音素将包含该代码字作为下一个代码字的可能性。状态的代码字无关过渡概率指示 HMM 将从该状态过渡到每个下一状态的概率。因而，当输入代码字时 HMM 将从当前状态过渡到下一状态的概率，是从当前状态到下一状态的过渡概率与用于接收代码字的输出概率的乘积。

图 1 表明用于音素的样本 HMM。HMM 包含三个状态和离开每个状态的两个过渡段。一般地，CSR 系统使用相同的状态图表示每个音素，但

带有音素依赖输出和过渡概率。根据这种 HMM，过渡仅出现在过渡至相同状态或模拟语音左至右本性的下一个状态。每个状态带有包含输出和过渡概率的相关的输出概率表和过渡概率表。如图 1 中所示，当 HMM 处于状态 2 时，用于代码字 5 的输出概率是.1，而当 HMM 处于状态 2 时，到状态 3 的过渡概率是.8。因而，当接收代码字 5 时 HMM 从状态 2 过渡到状态 3 的概率是.08(即.1×.8)。

为了确定代码字序列表示音素的概率，CSR 系统可以产生概率网格。用于音素的 HMM 的概率网格表示用于对代码字序列的每个可能状态路径的概率计算。概率网格包含用于对序列中每个代码字 HMM 可能处于其中的每个可能状态的节点。每个节点包含至今处理过的代码字将使 HMM 处于与该节点有关的状态中的累计概率。具体代码字的节点中的概率之和指示至今处理过的代码字表示音素字首部分的可能性。

图 2 是表明概率网格的图。概率网格表示当处理代码字序列“7、5、2、1、2”时，用于图 1 中所示 HMM 的每个可能状态的概率计算。横轴对应于代码字，而纵轴对应于 HMM 的状态。网格的每个节点包含每个源状态的概率乘以输出和过渡概率时的最大概率，而不是概率之和。例如，节点 201 包含 8.6E-6 的概率，该概率是 3.6E-4×.01×.9 和 1.4E-3×.03×.2 的最大值。有许多引导到任何节点的不同状态路径(即状态序列)。例如，节点 201 可以通过状态路径“1、2、3、3、”“1、2、2、3、”和“1、1、2、3、”到达。每个状态路径具有当处理代码字序列时 HMM 跟随该状态路径的概率。每个节点中的概率是引导到节点的每个状态路径的概率中的最大值。这些最大概率用于如下讨论的维特比（Viterbi）对准。

图 3 表明用于词的概率网格。纵轴对应于用于组成词的音素的 HMM 的状态的连接。节点 301 表示词的最终状态，并且包含引导到该节点的所有状态路径的最大概率。图 3 中的加粗线表示其终点在节点 301 处的最大概率的状态路径。在一定的用途中(例如训练 CSR 系统)，辨别具有引导到具体节点的最大概率的状态路径是有益的。一种用来辨别这样一种状态路径的熟知算法是维特比算法。在维特比算法已经确定了到最终状态的最大概率状态路径之后，有可能在网格中从最终节点回追且确定最大概率状态路径上的前一个节点，一路回到开始状态。例如，其终点在图 2 节点 203

处的最大概率的状态路径是“1、2、2、2、2、3”。当概率网格表示组成词的音素时，那么每个状态能依据音素和音素中的状态辨别。

CSR 系统的准确度部分地取决于用于每个音素的 HMM 的输出和过渡概率准确度。典型的 CSR 系统“训练” CSR 系统，从而输出和过渡概率准确地反映普通讲话者的语音。在训练期间，CSR 系统从各讲话者大量各种各样的词收集代码字序列。这样选择词，从而很多次读出每个音素。根据这些代码字序列，CSR 系统计算用于每个 HMM 的输出和过渡概率。各种用来计算这些概率的迭代计算法是熟知的，并且在 Huang 等的“用于语音识别的隐藏马尔克夫模型”(Edinburgh University Press，1990 年)中，进行了描述。

然而，伴随这种训练技术的一个问题是，这样的普通 HMM 可能不准确地模拟其语音模式与普通模式不同的人们的语音。一般地说，每个人都具有不同于普通模式的一定语音模式。因此，CSR 系统允许讲话者训练 HMM，以适应讲话者的语音模式。在这样的训练中，CSR 系统通过使用由系统的实际用户读出的训练发音来细化 HMM 参数，如输出和过渡概率及由代码字表示的量化向量。通过使用用户提供的数据以及由大量讲话者无关数据产生的信息和参数两者，导出适应的参数。因而，概率反映讲话者依赖特征。在 Huang 和 Lee 的“关于讲话者无关的、讲话者依赖的、和讲话者适应的语音识别(On Speaker-Independent, Speaker-Dependent,and Speaker-Adaptive Speech Recognition)”，Proc.ICASSP'91，1991 年，第 877-880 页中描述了一种这样的训练技术。

一般通过向讲话者呈现大量各种预选的词，来训练 CSR 系统。选择这些词，以保证能收集语音对应于每个音素的代表性样本。就这种代表性样本而言，CSR 系统能保证，不准确反映讲话者的音素发音的任何 HMM 能被适当地修改。当进行另外的训练时，例如因为讲话者不满意识别的准确度，CSR 系统就把另外的预选词呈现给讲话者。

尽管预选词的使用能提供适当的训练，但讲话者可能对于必须读出大量的词感到灰心。的确，由于词预选成包括每个音素，所以要求讲话者高效地读出其音素以可接受的准确度被模拟的词。因此，使训练系统能动态地选择用于训练的、将趋于优化训练准确度的、和减小要求讲话者读出的

词数量的词，将是有益的。

本发明涉及一种用来动态选择用来训练语音识别系统的词的方法和系统。每个词由语音识别系统模拟为包含音素的单元。训练系统收集其相应词是已知的读出发音。训练系统根据读出的发音辨别哪些语音单元由语音识别系统模拟得不准确。训练系统然后选择包含用于语音识别系统训练的所辨别语音单元的词。

在本发明的一个方面，语音识别系统把每个词模拟为一个音素序列，并且具有用于每个音素的 HMM 。训练系统通过把每个发音的每个代码字与已知词的音素对准，来辨别哪些音素模拟得不准确，收集的发音根据音素模型对应于这些已知的词。训练系统然后通过估计每个代码字对准的音素并且把代码字与其他音素相比较，来计算准确模拟音素的准确度指示。

图 1 表明用于音素的样本 HMM 。

图 2 是表明概率网格的图。

图 3 表明用于一个词的概率网格。

图 4 表示每个代码字与音素的对准。

图 5A 表示每个音素包含每个代码字的概率。

图 5B 表示用于每个代码字的每个音素的等级。

图 5C 表示用于每帧的代码字的每个音素的等级。

图 6 表示与音素对准的代码字的普通等级的样本计算。

图 7 是在其上运行一种最佳训练系统的计算机系统的方块图。

图 8 是训练系统的流程图。

图 9 是根据 HMM 的准确度用来排列音素的程序的流程图。

本发明提供了一种用来动态选择用来训练计算机语音识别(CSR)系统的词的方法和系统。在一个实施例中，训练系统辨别哪些语音单元，如音素，由 CSR 系统模拟得最不准确。训练系统然后辨别包含一个或多个这些最不准确模拟音素的词。训练系统提示讲话者读出这些辨别的词。训练系统然后对应于读出的词修改音素的模型。通过选择包含模拟得最不准确的音素的词，训练系统能集中训练其模型偏离讲话者的实际语音模式最大的模型。而且，不要求讲话者读出已经准确模拟的词。

训练系统通过估计由讲话者读出的、其对应词是已知的各种发音，来

确定哪些音素模拟得最不准确。训练系统把发音转换成代码词，然后在语音识别期间通过一个称为把代码字与音素对准的过程，来确定能把每个代码字考虑成哪个音素的部分。一旦对准完成，训练系统就在识别代码字是音素的部分时，为每个代码字确定对准音素的模型的准确度。例如，如果一个代码字与一个音素对准，并且模型预测与其他音素相比该代码字在该音素内的概率非常低，则该模型在识别该代码字为该音素的部分时的准确度较低。在确定用于每个代码字的音素模型的准确度之后，训练系统计算模型在识别对准代码字是音素的部分时的总准确度。总准确度能通过平均用于与该音素对准的每个代码字的准确度来计算。这些具有最低总准确度的音素模拟得最不准确。

训练系统选择用来训练的、包括模拟得最不准确的音素的词。训练系统可以使用几种不同的选择技术。训练系统可以辨别一定数量的模拟得最不准确的音素。训练系统然后可以选择任何包含至少一个辨别音素的词。可替换地，训练系统最好选择包含多于一个辨别音素的词，以减小讲话者需要读出以在辨别音素上训练的词的数量。而且，训练系统最好选择常说的词，以有助于保证讲话者不读出讲话者可能不熟悉的生僻词。

训练系统通过首先产生用于代码字和已知词的概率网格，把代码字序列与词的音素对准。训练系统然后辨别引导到最可几状态的最可几状态路径。这样一种状态路径的辨别最好使用基于维特比的算法。训练系统然后使用状态路径辨别哪些代码字能识别为哪些音素(与之对准的)的部分。

训练系统或通过具体提示训练的讲话者，或者通过保存由 CSR 系统误识别的发音以及正确的词，能收集在确定音素模型准确度时所用的语音。具体提示一般发生在训练对话期间。训练系统通过提示讲话者读出各种预选的词而开始，并然后相应修改模型。训练系统然后选择包含识别得最不准确的音素的词，并且提示讲话者读出这些词，且相应修改模型。训练系统能反复进行这种修改。如果收集的语音是误识别的发音，那么训练系统最初会选择包含那些确定为模拟得最不准确的误识别发音的音素的词。能通过把发音与误识别词的音素对准与正确词的音素对准相比较来确定哪个音素模型模拟得最不准确。在导出音素模型准确度的量度时能使用的因素包括：在误识别词中音素不正确识别的次数、和正确与不正确音素模型的

概率值的差。使用误识别发音的优点在于，训练是基于讲话者在通常说话期间实际使用的词中的音素。

一种最佳的CSR系统还自动地确定何时应该进行训练。通常，讲话者认为难以进行训练对话。因此，他们不可能着手训练对话，除非识别系统的准确度存在重大问题。而且，训练是如此困难，以致于讲话者会经常修改其语音模式以匹配模型。为了使训练过程更加讲话者友好，CSR系统能定期地或者当确定足够多的句素被不准确地模拟而认为训练是必要时，自动地着手一个短期训练对话。例如，能以每天为基础自动地着手训练对话，或者当注意到20个句素与模型匹配得不够准确时着手进行。

本发明的技术还能在例如当讲话者学习一门新语言时用来把词的适当发音教授给讲话者。指导系统与其认为音素模拟得的不准确，倒不如认为音素被讲话者错误地读出。因而，一旦辨别到发音最不准确的音素，指导系统就特别强调教给讲话者如何发音带有这些音素的词。而且，指导系统根据使用模拟音素计算的读出音素的准确度，将讲话者的发音分级。

在一个最佳实施例中，训练系统根据给定音素包含一个给定代码字的声学模型概率确定，该音素包含该代码字的可能性。对于每个代码词，训练系统根据音素的概率排列每个音素。就是说，具有最大概率的音素分配到最高等级(即等级1)。然后当计算模型的准确度时，使用这些等级。特别是，训练系统使用与它对准的所有代码字来计算音素的平均等级。训练系统然后选择包含具有用于训练的低平均等级的那些音素的词。

图4-6表明各音素的准确度计算。这个例子表明基于一个词的读出的计算。然而，在实际中，这样一种计算可能基于多个词。输入词包括音素10、12、和2。训练系统把相应发音划分成具有如下代码字的15帧：5、10、255、2、3、50、32、256、6、4、6、10、2、3、和5。训练系统然后把代码字与音素对准。图4表示每个代码字与音素的对准。表401带有一根对应于音素的横轴、和一根对应于帧的纵轴。表的项指示每个音素与其对准的代码字。代码字5、10、255、和2与音素10对准；代码字3、50、32、256、6、和4与音素12对准；及代码字6、10、2、3、和5与音素2对准。

图5A表示声学模型代码字/音素概率表。该表带有一根对应于代码字

的纵轴、和一根对应于音素的横轴。表中的每项包含相应音素包含该代码词的概率。例如，音素 10 包含代码字 6 的概率是.01，而音素 3 包含代码字 5 的概率是.04。每行列的概率之和是 1。

图 5B 表示代码字/音素等级表。这个表包含对于每个代码字的、该代码字相对于每个音素的概率的等级。例如，代码字 6 对于音素 10 具有等级为 33，这意味着代码字 6 在 32 个其他音素中的可能性比在音素 10 中的大，而且代码字 6 在音素 10 中的可能性比在 7 个其他音素中的大 (假定总共 40 个音素)。因而，表的每等级包含具有从 1 至 40 的数的项。

音素用于每个代码字的等级能以几种方式产生。例如，对于每帧，CSR 系统能辨别音素能产生用于该帧的代码字的声学模型概率。对于该帧，具有最大概率的音素分配到等级 1，具有第二大概率的音素分配到等级 2，以此类推。能根据来自代码字/音素概率表的信息动态地计算等级。图 5C 表示用于每帧的代码字的每个音素的等级。用于一帧的这些等级能通过按减小顺序动态地将对于该帧用于所有音素的概率分类来产生。另外，依据可存储的量，能一次产生该等级，如代码字/概率等级表中所示。

图 6 表示使用与音素对准的帧对这些音素的平均等级的样本计算。表 601 带有一根对应于音素的横轴、和一根对应于输入发音的代码字的纵轴。表的每项包含用于对准代码字的相应音素的等级。例如，代码字 5、10、255、和 2 与音素 10 对准，并且对于这些代码字音素 10 分别具有等级 19、31、15 和 1。表的底部包含等级的和、对准代码字的计数、及平均等级。例如，对于音素 10 等级的和是 66，对准代码字的计数是 4，及对于音素 10 平均等级因此是 16。如表所示，对于音素 12 平均等级是 13，而对于音素 2 平均等级是 19。由于音素 12 具有最高的平均等级，所以 CSR 系统认为该音素比其他两个音素模拟得更准确。反之，由于音素 2 具有最低的平均等级，所以 CSR 系统认为该音素比其他两个音素模拟得更不准确，并且最好选择该音素用于训练。

图 7 是在其上运行一种最佳训练系统的计算机系统的方块图。计算机系统 700 包含一个存储器 701、一个中央处理单元 702、存储装置 703、及显示装置 704。训练系统可以永久地存储在计算机可读存储介质上，如磁盘上，并且装入用于执行的计算机系统的存储器中。一种最佳的 CSR 系

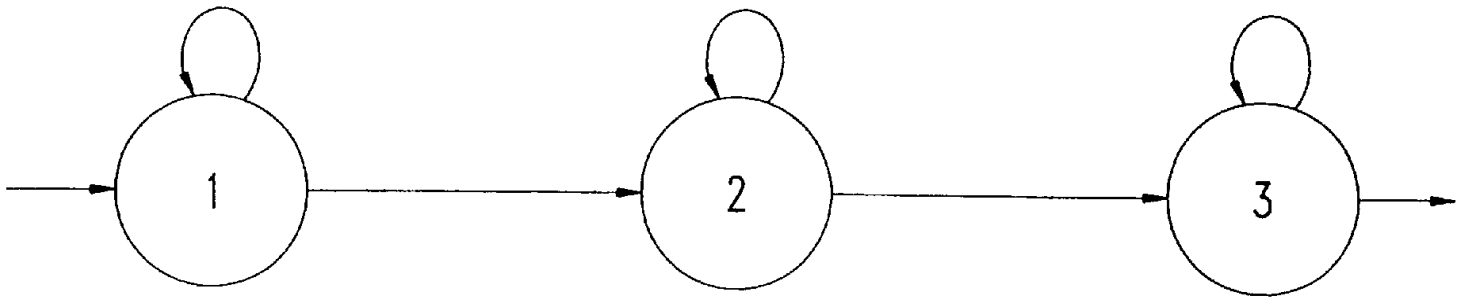
统包括一个识别元件 705、一个 HMM 元件 706、和一个训练元件 710。HMM 元件包含一个用于每个音素的隐藏马尔克夫模型、和每个词对其音素的映象。训练元件包含一个样本收集元件 711、一个代码字/音素对准元件 712、一个音素排列元件 713、及一个词选择元件 714。样本收集元件或通过具体提示用户、或收集误识别的发音来收集发音的各种样本和其相应的词。样本收集元件把发音转换成代码字。代码字/音素对准元件接收代码字和其相应的词，并且使用 HMM 把每个代码字与词的音素对准。音素排列元件使用代码字与音素的对准，以使用与音素对准的代码字计算这些音素的平均等级。词选择元件然后使用平均等级从可用的词汇(未表示)中选择词。

图 8 是训练系统的流程图。在步骤 801，训练系统根据用于每个音素的 HMM 的准确度排列所有音素，如图 9 中所描述的那样。在步骤 802，训练系统辨别模拟得最不准确的音素，即具有较低等级的那些音素。在步骤 803，训练系统根据辨别的音素选择用于训练的词。在步骤 804，训练系统提示讲话者读出选择词的每一个。在步骤 805，训练系统根据用于选择词的发音修改 HMM。

图 9 是根据 HMM 的准确度用来排列音素的程序的流程图。在一个实施例中，这个程序通过提示讲话者读出训练词来收集发音，以在排列时使用。程序然后计算每个音素准确度的指示。在步骤 901，程序从第一个训练词开始选择下一个训练词。训练词可以是预先建立的、或预先定义的训练词的集合，或者可以动态地预先选择。在步骤 902，如果已经选择所有的训练词，那么程序继续到步骤 911，否则程序继续到步骤 903。在步骤 903，程序提示讲话者读出选择的词，并且接收相应的发音。在步骤 904，程序把发音转换成代码字序列。在步骤 905，程序把每个代码字与每个最可能对应的词的音素对准。在步骤 906-910，程序循环选择每个代码字和累计与该代码字对准的音素的等级。在步骤 906，程序从第一个代码字开始选择下一个。在步骤 907，如果已经选择所有的代码字，那么程序循环到步骤 901，以选择下一个训练词，否则程序继续到步骤 908。在步骤 908，程序在对准代码字的范围内辨别音素的等级。在步骤 909，程序累计用于对准音素的辨别等级。在步骤 910，程序增大与该音素对准的代码字的数

量计数，并且循环到 906 以选择下一个代码字。在步骤 911，程序通过把累计等级除以计数来计算每个音素的平均等级，并且返回。

尽管按照最佳实施例已经描述了本发明，但不打算把本发明限于这些实施例。在本发明的实质范围内的改进对于熟悉本专业的技术人员将是显而易见的。例如，尽管按照识别离散的语音发音描述了本发明，但本发明能容易地用于连续的语音识别系统中。此外，本发明的技术能用于不使用隐藏马尔克夫模型的识别系统。而且，使用产生代码字的声学模型概率之外的量度，如通过使用用识别器的音素误识别的计数，也能计算音素的等级。根据求和的不同级而不是帧级能计算语音单元的等级。例如，以语音段级能求和等级，这里语音段包括多个帧或语音的可变长度时段。以粒度的不同级，如音素、在音素中的状态、句素、一种在上下文依赖音素中的状态、或完整的词本身，能计算语音单元的等级，和进行在选择用于训练的词中的语音单元的选择。上下文依赖音素可以取决于多个周围音素或词的上下文。也可以把一个完整的词考虑为在训练时用于模拟和选择的单元。当词汇大小较小时，或者当某些词经常使用并且可能混淆时，如英语字母和数字，使用完整的词作为单元是便利的。CSR 系统能使用等级之外的准确度量度。例如，CSR 系统可以使用音素概率与用于该帧的最好音素概率的差值或比值。而且，使用求平均之外的技术，如计算跨过多个出现的相同语音单元的准确度度量的最大值、最小值、或加权和，能组合跨过不同帧的等级或准确度量度信息。最后，CSR 系统能使用关于音素模型准确度的收集信息(总称为误差分布)，以改进识别过程本身。例如，如果误差分布表示识别该音素模型的机会在其已知出现期间较小，如由误差分布辨别的那样，则在识别期间能增大语音单元的概率。本发明的范围由如下的权利要求书限定。



状态 1 2 3

输出概率表

代码字	输出概率
1	.01
2	.2
⋮	⋮
5	.01
6	.02
7	.03
⋮	⋮
256	.001

代码字	输出概率
1	.1
2	.03
⋮	⋮
5	.1
6	.03
7	.05
⋮	⋮
256	.05

代码字	输出概率
1	.05
2	.01
⋮	⋮
5	.03
6	.01
7	.01
⋮	⋮
256	.02

过渡概率表

状态	过渡概率
本身	.4
下一个	.6

状态	过渡概率
本身	.2
下一个	.8

状态	过渡概率
本身	.9
下一个	.1

图 1

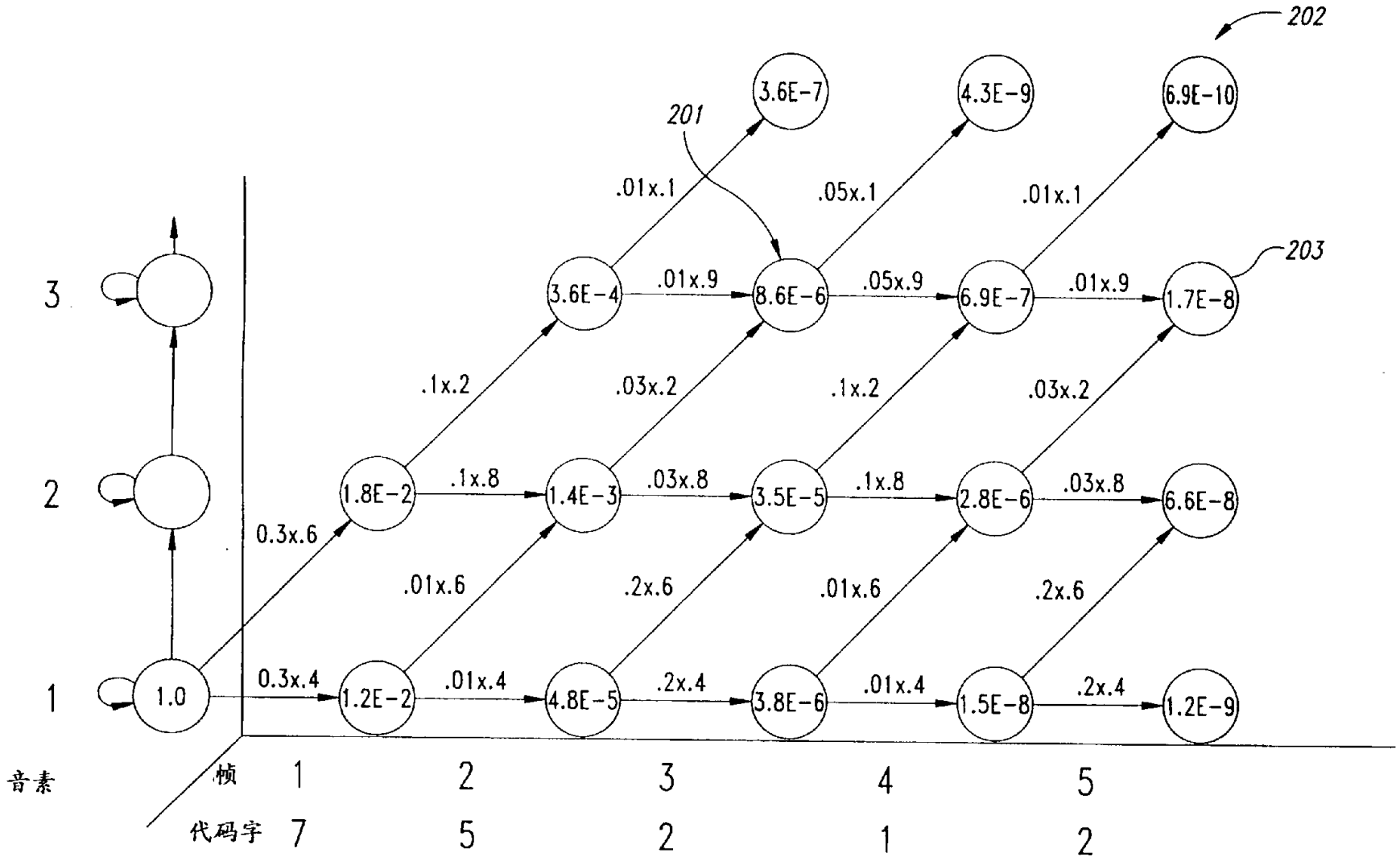

202
201
203
3.6E-7
4.3E-9
6.9E-10
.01x.1
.05x.1
.01x.1
3
3.6E-4
.01x.9
8.6E-6
.05x.9
6.9E-7
.01x.9
1.7E-8
.1x.2
.03x.2
.1x.2
.03x.2
2
1.8E-2
.1x.8
1.4E-3
.03x.8
3.5E-5
.1x.8
2.8E-6
.03x.8
6.6E-8
0.3x.6
.01x.6
.2x.6
.01x.6
.2x.6
1
1.0
0.3x.4
1.2E-2
.01x.4
4.8E-5
.2x.4
3.8E-6
.01x.4
1.5E-8
.2x.4
1.2E-9
音素
帧 1 2 3 4 5
代码字 7 5 2 1 2

图 2

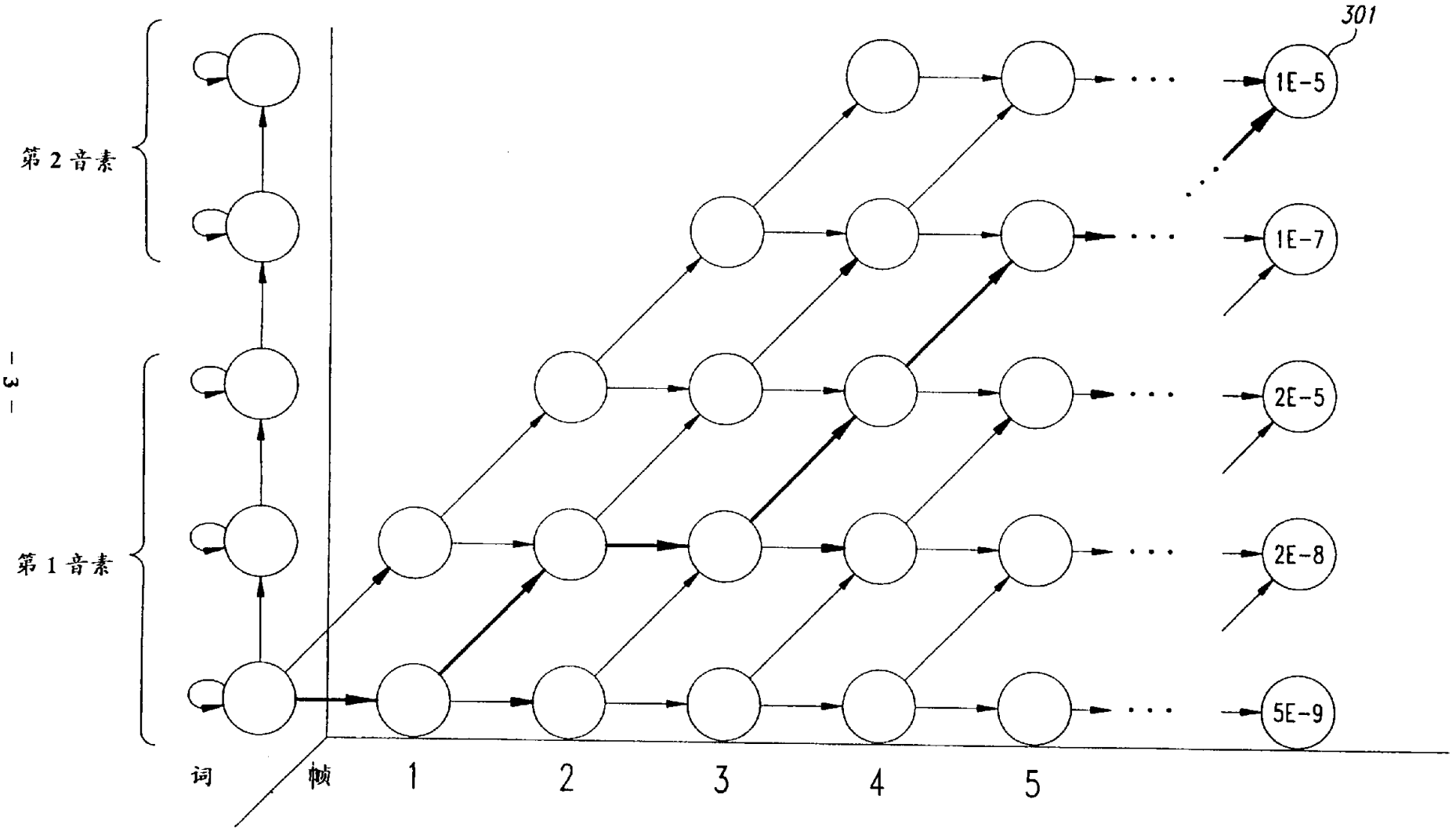


图 3

词 = 10 + 12 + 2

代码字序列 = 5, 10, 255, 2, 3, 50, 32, 256, 6, 4, 6, 10 2, 3, 5

帧/音素对准表

帧	音素 10	音素 12	音素 2
1	5		
2	10		
3	255		
4	2		
5		3	
6		50	
7		32	
8		256	
9		6	
10		4	
11			6
12			10
13			2
14			3
15			5

401

图 4

代码字/音素概率表

音素

代码字	1	2	3	10	12	40
1	.2	.05	.14	.06	.001	.15
2	.02	.001	.03	.25	.06	.05
3	.002	.25	.004	.04	.02	.001
4	.04	.03	.05	.002	.1	.3
5	.02	.2	.04	.03	.025	.021
6	.1	.07	.08	.01	.02	.05
10	.007	.03	.04	.005	.006	.1
32	.04	.03	.4	.1	.15	.02
50	.006	.002	.2	.007	.01	.005
255	.01	.03	.05	.02	.015	.2
256	.2	.09	.06	.001	.07	.05

图 5A

代码字/音素概率表

音素

代码字	1	2	3	…	10	…	12	…	40
1	5	25	13		23		39		6
2	17	40	21		1		5		8
3	36	2	27		6		10		37
4	15	16	11		36		5		1
5	23	3	15		19		20		22
6	5	12	7		33		29		16
⋮									
10	24	18	17		31		27		4
⋮									
32	15	16	1		5		4		23
⋮									
50	26	35	6		25		14		27
⋮									
255	26	10	6		15		24		1
256	2	4	18		37		17		21

图 5B

帧

音素	1	2	3	4	5	
	5	10	255	2	3	
1	23	24	26	17	36	
2	3	18	10	40	2	
⋮						
10	19	31	15	1	6	
⋮						
12	20	27	24	5	10	

图 5C

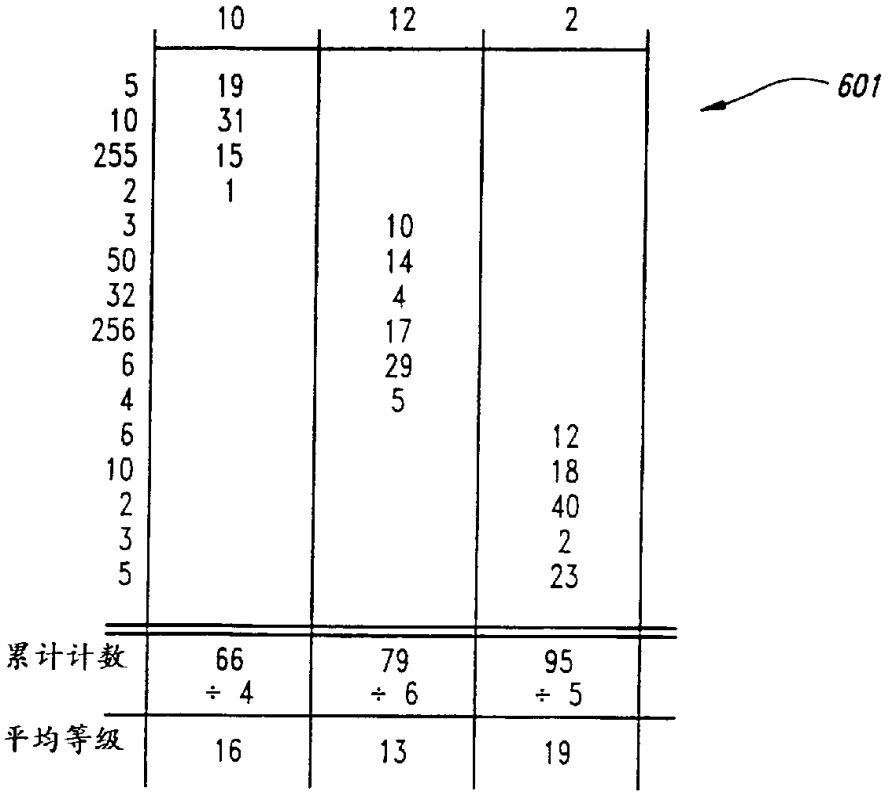

601
10
12
2
5
10
255
2
3
50
32
256
6
4
6
10
2
3
5
19
31
15
1
10
14
4
17
29
5
12
18
40
2
23
累计计数
66
÷ 4
79
÷ 6
95
÷ 5
平均等级
16
13
19

图 6

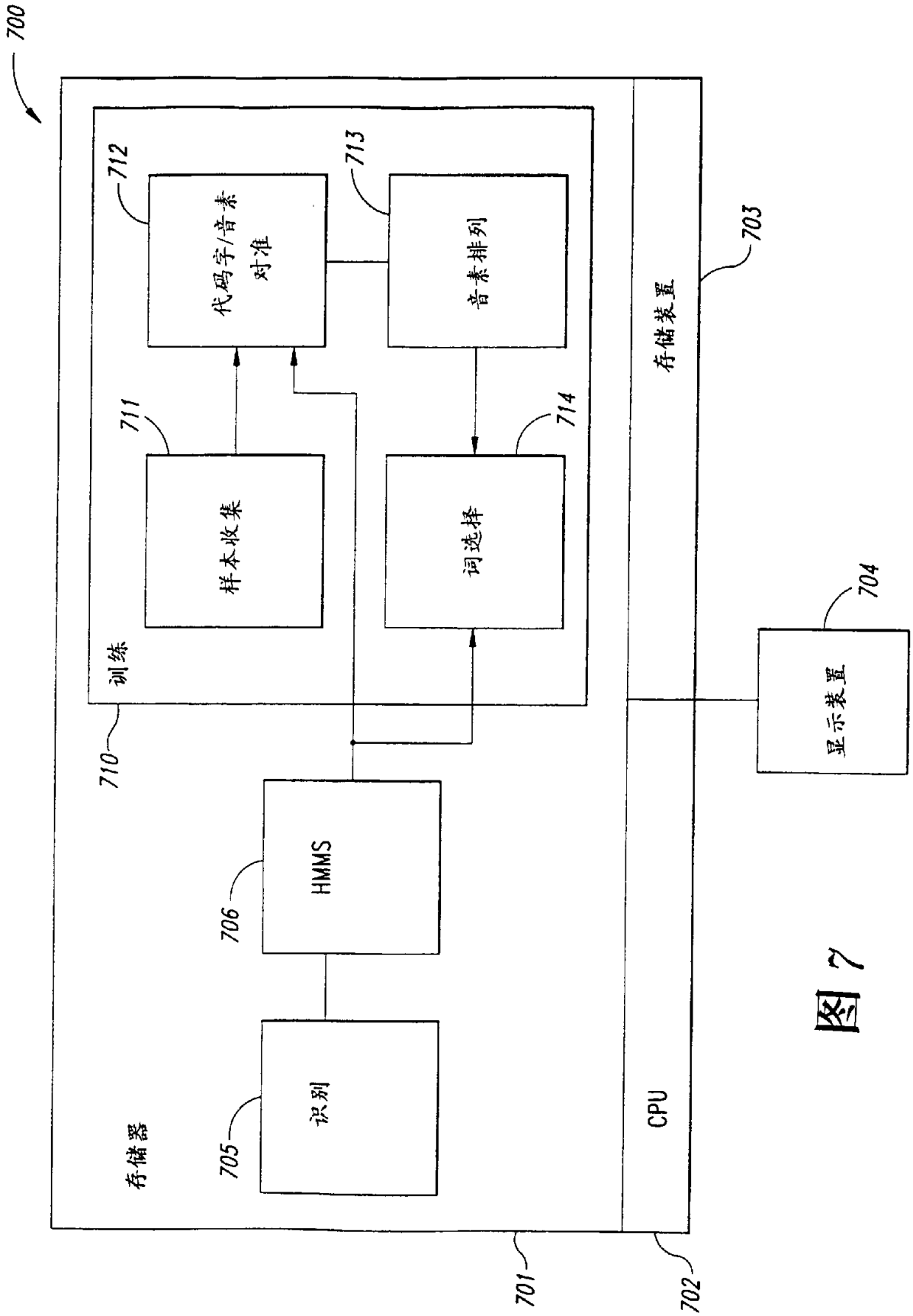

700
712
713
代码字/音素对准
音素排列
711
714
样本收集
词选择
训练
710
存储装置
703
显示装置
704
HMMS
706
识别
705
存储器
CPU
701
702

图 7

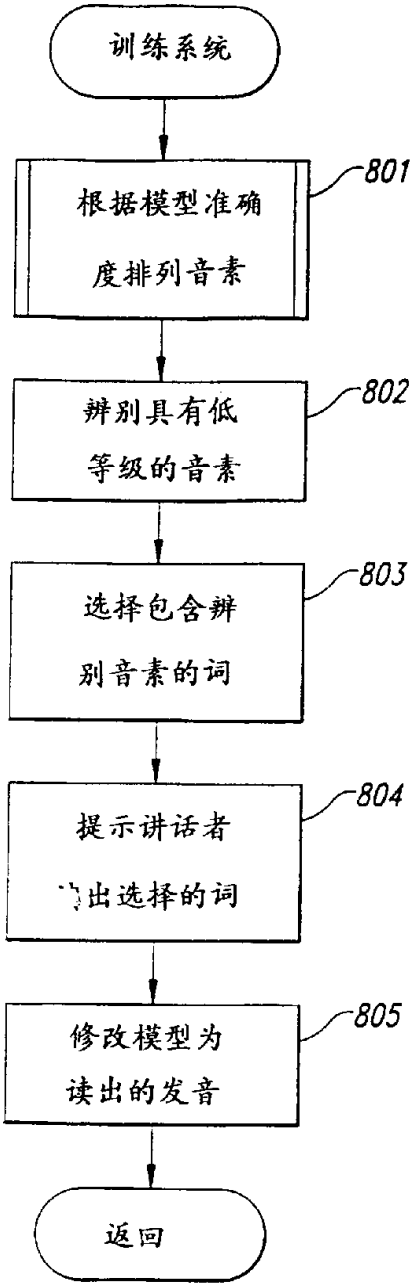

训练系统
根据模型准确度排列音素
801
辨别具有低等级的音素
802
选择包含辨别音素的词
803
提示讲话者读出选择的词
804
修改模型为读出的发音
805
返回

图 8

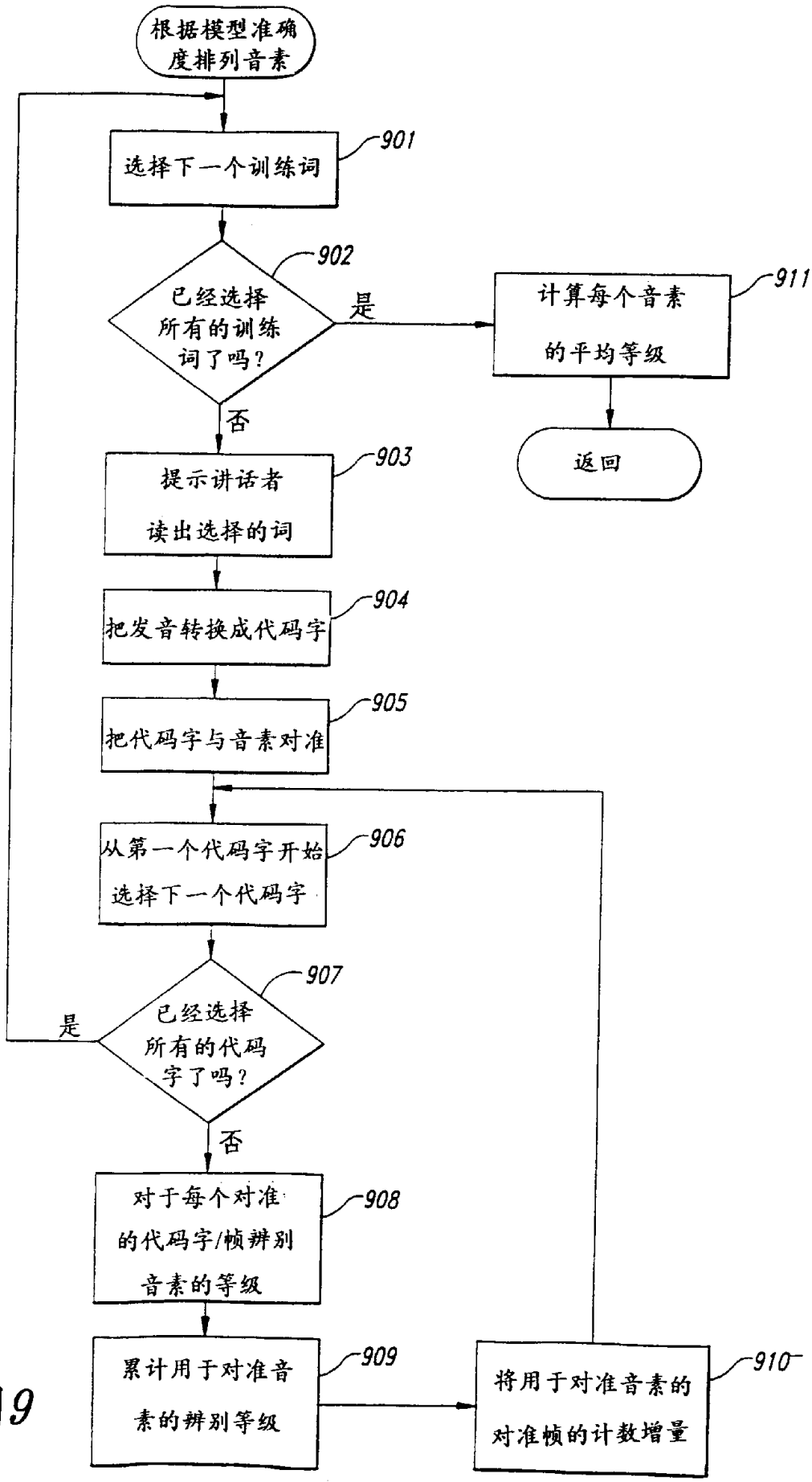

根据模型准确度排列音素
选择下一个训练词
901
已经选择所有的训练词了吗？
902
是
否
计算每个音素的平均等级
911
返回
提示讲话者读出选择的词
903
把发音转换成代码字
904
把代码字与音素对准
905
从第一个代码字开始选择下一个代码字
906
已经选择所有的代码字了吗？
907
是
否
对于每个对准的代码字/帧辨别音素的等级
908
累计用于对准音素的辨别等级
909
将用于对准音素的对准帧的计数增量
910

图9