[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 13/04 (2006.01)
G10L 21/06 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710139735.2

[43] 公开日 2009 年 2 月 4 日

[11] 公开号 CN 101359473A

[22] 申请日 2007.7.30
[21] 申请号 200710139735.2
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 施 琴 秦 勇 刘 义 双志伟

[74] 专利代理机构 北京市中咨律师事务所
代理人 于 静 杨晓光

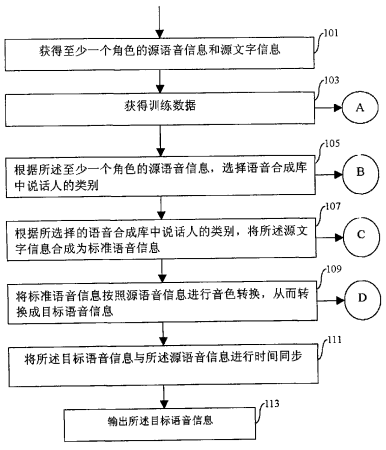
权利要求书 3 页 说明书 19 页 附图 11 页

[54] 发明名称

自动进行语音转换的方法和装置

[57] 摘要

本发明提出了一种能够显著改进音色转换的质量，并保证转换的声音相似度的方法和装置。 本发明在语音合成库中设置有若干标准说话人，根据不同的角色，本发明选用不同的标准说话人的声音进行语音合成，所述被选中的标准说话人的声音与原始角色之间已经存在一定程度的相似性。 然后本发明将这种与原始声音具有一定程度相似性的标准语音进一步进行音色转换，以精确模仿原始说话人的声音，从而使得转换后的声音在保证相似度的同时，更加接近原始的语音特征。



1、一种用于自动进行语音转换的方法，所述方法包括:

获得源语音信息和源文字信息;

根据源语音信息，选择语音合成库中的标准说话人;

根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息；以及

将所述标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息。

2、一种如权利要求0所述的方法，进一步包括获得训练数据的步骤，所述获得训练数据的步骤包括:

对齐所述源文字信息和源语音信息。

3、一种如权利要求0所述的方法，其中所述获得训练数据的步骤还包括:

对所述源语音信息的角色进行聚类。

4、一种如权利要求0所述的方法，进一步包括将所述目标语音信息与所述源语音信息进行时间同步的步骤。

5、一种如权利要求0所述的方法，其中所述选择语音合成库中的标准说话人的步骤进一步包括:

根据语音合成库中的标准说话人的标准语音信息与源语音信息之间的基频差异和频谱差异，选择声学特征差异最小的语音合成库中的标准说话人。

6、一种如权利要求0所述的方法，其中所述将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的步骤进一步包括:

根据语音合成库中的标准语音信息与源语音信息之间的基频差异和频谱差异，对所述标准语音信息进行音色转换，将其转换成目标语音信息。

7、一种如权利要求0或0所述的方法，其中所述基频差异包括基频的均值差异和方差差异。

8、一种如权利要求0所述的方法，其中将所述目标语音信息与所述源语音信息进行时间同步的步骤包括根据源语音信息进行同步。

9、一种如权利要求0所述的方法，其中将所述目标语音信息与所述源语音信息进行时间同步的步骤包括根据源语音信息所对应的画面信息进行同步。

10、一种用于自动进行语音转换的系统，所述系统包括:

获得源语音信息和源文字信息的单元;

根据所述源语音信息，选择语音合成库中的标准说话人的单元;

根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元；以及

将所述标准语音信息按照源语音进行音色转换，从而得到目标语音信息的单元。

11、一种如权利要求0所述的系统，进一步包括获得训练数据的单元，所述获得训练数据的单元包括:

对齐所述源文字信息和源语音信息的单元。

12、一种如权利要求0所述的系统，其中所述获得训练数据的单元还包括:

对所述源语音信息的角色进行聚类的单元。

13、一种如权利要求0所述的系统，进一步包括将所述目标语音信息与所述源语音信息进行时间同步的单元。

14、一种如权利要求0所述的系统，其中所述选择语音合成库中的标准说话人的单元进一步包括:

根据语音合成库中的标准说话人的标准语音信息与源语音信息之间的基频差异和频谱差异，选择声学特征差异最小的语音合成库中的标准说话人的单元。

15、一种如权利要求0所述的系统，其中所述将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的单元进一步包括:

根据语音合成库中的标准语音信息与源语音信息之间的基频差异和频

谱差异，对所述标准语音信息进行音色转换，将其转换成目标语音信息的单元。

16、一种如权利要求0或0所述的系统，其中所述基频差异包括基频的均值差异和方差差异。

17、一种如权利要求0所述的系统，其中将所述目标语音信息与所述源语音信息进行时间同步的单元包括根据源语音信息进行同步的单元。

18、一种如权利要求0所述的系统，其中将所述目标语音信息与所述源语音信息进行时间同步的单元包括根据源语音信息所对应的画面信息进行同步的单元。

19、一种媒体播放装置，所述媒体播放装置至少用于播放语音信息，所述装置包括:

获得源语音信息和源文字信息的单元;

根据所述源语音信息，选择语音合成库中的标准说话人的单元;

根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元; 以及

将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的单元。

20、一种媒体写装置，所述装置包括:

获得源语音信息和源文字信息的单元;

根据所述源语音信息，选择语音合成库中的标准说话人的单元;

根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元;

将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的单元; 以及

将所述目标语音信息写入至少一个存储装置的单元。

21、一种计算机程序产品，该计算机程序产品包括存储在计算机可读存储介质中的程序代码，所述程序代码用于完成权利要求0－0中任何一个权利要求的方法的操作。

自动进行语音转换的方法和装置

技术领域

本发明涉及语音转换的领域，并且本发明特别涉及将文字信息进行语音合成和音色转换的方法和装置。

背景技术

当人们观看一段影音文件（如外文电影）时，语言不通经常构成一个显著的阅读障碍。现有的影片发行商们可以在相对较短的时间内将外文字幕（如英文）翻译成本地文字字幕（如中文），并且同步发行带有本地文字字幕的电影供观众欣赏。然而阅读字幕仍然会影响大部分观众的观看感受，因为观众的视线需要在字幕和画面之间不断的快速切换，尤其对于儿童、老人、视力有障碍或阅读有障碍的人群，阅读字幕所带来的负面影响尤为突出。为了照顾其它地区的观众市场，影音文件的发行商们可以聘请配音演员对影音文件赋予中文配音。然而这一过程往往需要较长的时间，并且需要花费大量的人力成本。

语音合成技术（TTS Text to Speech）可以将文字信息转换成语音信息。美国专利 US5970459 提供了一种利用 TTS 技术将电影字幕转换成本地语音的方法。该方法分析原始语音数据和原始说话人的嘴型（shape of lip），先将文字信息利用 TTS 技术转换成语音信息，然后按照嘴型的运动对这些语音信息进行同步，从而形成电影的配音效果。然而该方案并没有使用音色转换技术，无法使合成的声音与电影原声音色相近，最终的配音效果与原声的声音特征相差很远。

音色转换技术可以把原始说话人的声音转换成目标说话人的声音。现有技术中经常利用频率弯曲的方法将原始说话人的声音频谱转换成目标说

话人的声音频谱，从而按照目标说话人的声音特征，包括声音的语速、语调，制造出相应的声音数据。频率弯曲（frequency wrapping）技术是一种用于补偿不同说话者之间的声音频谱的差异的方法，它广泛应用于语音识别和语音转换领域。按照频率弯曲技术，给定一个声音的一频谱截面，该方法通过施加一频率弯曲函数来生成一新的频谱截面，使一个说话人的声音听起来象另一个说话人的声音。

在现有技术中已提出了许多用于发现性能良好的频率弯曲函数的自动训练方法。一种方法是最大似然线性回归。该方法的描述可参见: L.F.Uebel, 和 P.C. Woodland 的 " An investigation into vocal tract length normalization, " EUROSPEECH ' 99, Budapest, Hungary, 1999, 第 2527-2530 页。然而，这种方法需要大量的训练数据，这限制了它在很多应用场合中的使用。

另一种方法是利用共振峰映射技术进行声音的转换，该方法的描述可参见: Zhiwei Shuang, Raimo Bakis, Yong Qin 的 "Voice Conversion Based on Mapping Formants" in Workshop on Speech to Speech Translation, Barcellona, June 2006。具体而言，该方法根据源说话人和目标说话人之间的共振峰（formant）的关系来获得频率弯曲函数。共振峰是指在发音时由于声道本身的共振而在声音频谱中形成的声音强度较大的若干频率区域。共振峰与声道的形状有关，因此每一个人的共振峰通常是不同的。而不同说话人的共振峰可用于表示不同说话人之间的声学差异。并且该方法还利用基频调整技术使得仅仅利用少量的训练数据就能够进行声音的频率弯曲。然而该方法所未能解决的问题是如果原始说话人与目标说话人之间的声音相差很远，由于频率弯曲所带来的音质损伤就会急剧增加从而损坏输出声音的质量。

实际上在衡量音色转换的优略时，存在两种指标，其一是被转换的声音的质量、其二是被转换的声音与目标说话人的相似程度。现有技术中二者常常处于相互牵制的状态，很难同时满足。也就是说即便将现有的音色转换技术应用于美国专利US5970459中的配音方法时也难以形成很好的配

音效果。

发明内容

为了解决现有技术的上述问题，本发明提出了一种能够显著改进音色转换的质量，并保证转换的声音相似度的方法和装置。本发明在语音合成库中设置有若干标准说话人，根据不同的角色，本发明选用不同的标准说话人的声音进行语音合成，所述被选中的标准说话人的声音与原始角色之间已经存在一定程度的相似性。然后本发明将这种与原始声音具有一定程度相似性的标准语音进一步进行音色转换，以精确模仿原始说话人的声音，从而使得转换后的声音在保证相似度的同时，更加接近原始的语音特征。

具体而言，本发明提供了一种用于自动进行语音转换的方法，所述方法包括：获得源语音信息和源文字信息；根据源语音信息，选择语音合成库中的标准说话人；根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息；以及将所述标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息。

本发明还提供了一种用于自动进行语音转换的系统，所述系统包括：获得源语音信息和源文字信息的单元；根据所述源语音信息，选择语音合成库中的标准说话人的单元；根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元；以及将所述标准语音信息按照源语音进行音色转换，从而得到目标语音信息的单元。

本发明还提供了一种媒体播放装置，所述媒体播放装置至少用于播放语音信息，所述装置包括：获得源语音信息和源文字信息的单元；根据所述源语音信息，选择语音合成库中的标准说话人的单元；根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元；以及将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的单元。

本发明还提供了一种媒体写装置，所述装置包括：获得源语音信息和源文字信息的单元；根据所述源语音信息，选择语音合成库中的标准说话

人的单元；根据所选择的语音合成库中的标准说话人，将所述源文字信息合成为标准语音信息的单元；将标准语音信息按照源语音信息进行音色转换，从而得到目标语音信息的单元；以及将所述目标语音信息写入至少一个存储装置的单元。

通过本发明的方法和装置，可以按照原始说话人的声音，把影音文件中的字幕自动转换成声音信息。保证转换后的声音与原声的相似度的同时，进一步提高声音的转换质量，使得转换后的声音更加逼真。

上述描述大致列举了本发明的优越之处，通过结合附图与本发明最佳实施例的详细说明，本发明的这些以及其它优点将更加明显。

附图说明

本说明中所参考的附图只用于示例本发明的典型实施例，不应该认为是对本发明范围的限制。

图 1 为进行语音转换的流程图。

图 2 为获得训练数据的流程图。

图 3 为选择语音合成库中说话人类别的流程图。

图 4 为计算标准说话人与源说话人基频差异的流程图。

图 5 为源说话人与标准说话人基频差异均值比较示意图。

图 6 为源说话人与标准说话人基频差异方差比较示意图。

图 7 为计算标准说话人与源说话人频谱差异的流程图。

图 8 为源说话人与标准说话人频谱差异比较示意图。

图 9 为将所述源文字信息合成为标准语音信息的流程图。

图 10 为将标准语音信息按照源语音信息进行音色转换的流程图。

图 11 为自动语音转换系统的结构框图。

图 12 为带有自动语音转换系统的影音文件配音装置的结构框图。

图 13 为带有自动语音转换系统的影音文件播放器的结构框图。

具体实施方式

下列讨论中，提供大量具体的细节以帮助彻底了解本发明。然而，很显然对于本领域技术人员来说，即使没有这些具体细节，并不影响对本发明的理解。并且应该认识到，使用如下的任何具体术语仅仅是为了方便描述，因此，本发明不应当局限于只用在这样的术语所标识和/或暗示的任何特定应用中。

除非另有说明，本发明所述的功能可用硬件或软件或它们的结合来运行。然而，在一优选实施列中，除非另有说明，这些功能是由处理器，如计算机或电子数据处理器，按照编码，如计算机程序编码，的集成电路来执行的。一般来说，为了实现本发明的实施例而执行的方法可以是操作系统或特定应用程序的一部分、程序、模块、对象或指令序列。本发明的软件通常包括将由本地计算机呈现成机器可读格式的众多指令，因此是可执行指令。此外，程序包括相对于程序来说驻留在本地或在存储器中找到的变量和数据结构。另外，下文描述的各种程序可以根据在本发明的特定实施例中实现它们的应用方法来识别。当携带指向本发明的功能的计算机可读指令时，这样的信号承载媒体代表本发明的实施例。

本发明以配有中文字幕的英文电影文件为例进行说明，但是本领域的普通技术人员理解，本发明并不局限于这一应用场景。图 1 为进行语音转换的流程图。步骤 101 用于获得至少一个角色的源语音信息和源文字信息。比如，所述源语音信息可以是英文电影原声：

Tom: I'm afraid I can't go to the meeting tomorrow.

Chris: Well, I'm going in any event.

源文字信息可以是电影片断中与该句话所对应的中文字幕：

汤姆：我恐怕不能参加明天的会议了。

克莉丝：好吧，但无论如何我会去的。

步骤 103 用于获得训练数据，所述训练数据包括语音信息和文字信息，其中语音信息用于进行后续的标准说话人的选择和音色转换，文字信息用于进行语音合成。理论上，如果所提供的语音信息和文字信息能够严格对

应，并且语音信息已经进行了很好的分割，可以略去这一步骤。但是在当前的电影文件大多无法提供准备好的训练数据，因此，本发明需要在进行声音转换之前对训练数据进行预处理。这一步骤将在下文中进行更详细的说明。

步骤 105 用于根据所述至少一个角色的源语音信息，选择语音合成库中说话人的类别。所述语音合成（TTS）是指将文字信息转换成语音信息的过程，语音合成库中存储若干标准说话人的声音。传统上，语音合成库中可以只存储一个说话人的声音，如存储某个电视台播音员的一段或者若干段录音。所存储的声音以一句话为一个单位，根据需求的不同所存储的单句数量不等，经验表明最少需要存储几百句话，通常存储的单句数量在 5000 句左右。本领域的普通技术人理解，所存储的单句数量越多，能够用于合成的语音信息就越丰富。语音合成库中所存储的单句会被分割成若干小的单元，比如一个字、一个音节（syllables）、一个音素（phonemes），甚至 10 毫秒的语音段。语音合成库中的标准说话人的录音可以和待转换的文字没有任何关系，比如语音合成库中所记录的是一段新闻播报员所播报的事实新闻，而待合成的文字信息是一段电影片断。只要所述文字所包含的“字”、“音节”或“音素”的发音能够在语音合成库的标准说话人的声音单元中找到就可以完成语音合成的过程。

本发明在此采用多于一个的标准说话人，目的是为了使标准说话人的声音和电影原声更加接近，从而在后续的音色转换过程中减少音质损伤。选择语音合成库中说话人的类别就是选择一个音色最接近的标准说话人作为 TTS 的标准说话人声音。本领域的普通技术人员了解，根据一些基本的声学特征，如语调（intonation）、音调（tone），可以把不同的声音进行归类，比如女高音、女低音、男高音、男低音、童音等。这些归类有助于对源语音信息有一个粗略的定义，而这一定义过程可以显著的提升音色转换过程的效果和质量。分类越细，最终的转换效果可能越好，但是分类越细带来的计算成本和存储成本也较高。本发明以 4 个标准说话人的声音（女 1、女 2、男 1、男 2）为例进行说明，但是本发明并不限于这样的分类方

法。更详细的内容将在后文中进行说明。

在步骤 107 中，根据所选择的语音合成库中说话人的类别，也就是所选择的标准说话人，将所述源文字信息合成为标准语音信息。比如通过步骤 105 的选择，选中了男 1（男高音）作为汤姆那句话的标准说话人，所述源文字信息“*我恐怕不能参加明天的会议了*”，将会被用男 1 的声音表达出来。详细的步骤将在后文中描述。

在步骤 109 中，本发明将标准语音信息按照源语音信息进行音色转换，从而转换成目标语音信息。在上一步骤中，用男 1 的标准语音来表达汤姆的台词，虽然男 1 的声音在某种程度上与电影原声中汤姆的声音类似，比如都是男声的声音，而且音调都比较高，但是二者的相似度是十分粗略的。这样的配音效果会大大损害观众对电影配音的观感，因此必须进行音色转换的步骤以使男 1 的声音能够听起来像汤姆在电影原声中的声音特征。经过这样的转换过程，所产生的与汤姆的原声十分接近的中文发音就被称为目标语音。更详细的步骤将在下文中进行说明。

在步骤 111 中，所述目标语音信息与所述源语音信息进行时间同步。因为同一句话的中文和英文表达时长不同，比如英文的 *“I’m afraid I can’t go to the meeting tomorrow”* 可能就比中文的“*我恐怕不能参加明天的会议了*”略短一些，前者用时 2.60 秒，后者用时 2.90 秒。这样所引起的常见问题是，画面中的人物已经结束说话，而合成的声音还在继续。当然，也有可能画面中的人物还没有结束说话，而目标语音已经停止。因此，我们需要对目标语音和源语音信息或者画面进行同步。由于源语音信息和图像信息通常是同步的，因此可以有两种方法进行这一同步过程，其一是使目标语音信息与源语音信息同步，其二是使目标语音信息与图像信息进行同步。下面分别进行说明。

在第一种同步方法中，可以利用源语音信息的开始和结束时间进行同步。开始和结束的时间可以利用简单的静音检测获得，也可以利用将文字信息和语音信息对齐的方式获得（比如，已知源语音信息 *“I’m afraid I can’t go to the meeting tomorrow”* 所处的时间位置为 01:20:00,000 到

01:20:02,600，则源文字信息“*我恐怕不能参加明天的会议了*”所对应的中文目标语音的时间位置也应调整为 01:20:00,000 到 01:20:02,600）。在获得源语音信息的开始和结束时间后，将目标语音信息的起始时间设定为与源语音信息的起始时间一致（比如 01:20:00,000），同时，将对目标语音信息的时长进行调整（比如由 2.90 秒调整为 2.60 秒）以保证目标语音的结束时间和源语音的结束时间一致。注意，这种时长的调整一般而言可以是均匀进行的（比如上文中将一个时长 2.90 秒的句子均匀压缩为 2.60 秒），从而保证对每一个音的压缩都是一致的，这样可以保证一个句子经过压缩或者延长之后听起来声音仍然是自然平滑的。当然对于一些很长的句子有明显停顿的地方，也可以将其分成若干段进行同步。

在第二种同步方法中，根据画面的信息对目标语音进行同步。本领域的普通技术人员理解，人物的脸部信息，特别是唇部信息，能够基本准确的表达声音同步信息。对于一些简单的场景，比如固定背景的单一说话人情形，唇部信息可以较好的识别。可以利用所识别的唇部信息，判断语音的起始和结束时间，从而按照与上文类似的方法整合目标语音的时长，设置目标语音的时间位置。

需要指出的是，在一种实施例中，上述同步步骤可以在音色转换之后单独进行，而在另一种实施例中，上述同步步骤可以与音色转换一起进行。后一实施例可能能够带来更优的效果，因为每一次对声音信号的处理都可能造成对声音质量的损害，这是由于对声音的分析和重建所带来的固有缺陷，将两步骤同时完成可以减少对声音数据的处理次数，从而进一步提高声音数据的质量。

最后，在步骤 113 中，经过同步的目标语音数据与画面或文字信息一同输出。从而产生自动配音的效果。

下面参考图 2 对获得训练数据的过程进行说明。在步骤 201 中首先对声音信息进行预处理，过滤背景声音。语音数据，特别是电影中的语音数据可能包含很强的背景噪音或者音乐声音，这些数据用于训练可能会损害训练结果，因此需要排除这些背景声音，而只保留纯粹的语音数据。如果

电影数据按照 Mpeg 协议进行存储，则可以自动的区分不同的声音信道，如图 11 中的背景声音信道 1105 和前景声音信道 1107。但是当源影音数据中没有对前景声音和背景声音进行区分，或者即便进行了区分，前景声音中仍然混入了一些非语音的或者无字幕对应的语音声音（如，一群小孩的混乱的叫嚷生）时，就可以进行上述过滤步骤。这一过滤过程可以通过语音识别技术中的隐马尔可夫模型(HMM)进行，该模型较好地描述了语音现象的特性，基于 HMM 的语音识别算法也取得了比较好的识别效果。

在步骤 203 中，对字幕进行预处理，过滤那些无语音信息对应的文字信息。由于字幕中可能包含一些非语音的解释性信息，这一部分信息无需进行语音合成，因此也需要进行预先过滤。如:

00:00:52,000 --> 00:01:02,000

www.1000fr.com present

一种简单的过滤方法就是设定好一系列特殊关键词进行过滤。比如对于上面这种形式的数据，我们可以设定关键字<font 和 ，对这两个关键字之间的信息进行过滤。在影音文件中这样的解释性文字信息大多是有规律的，因此设定一个关键字过滤集合基本上可以满足绝大部分过滤需求。当然，在过滤大量不可预测的解释性文字信息时，也可以使用其它的方法，比如利用 TTS 技术寻找是否存在与文字信息对应的语音信息，如果没有找到与“www.1000fr.com present”所对应的语音信息，则认为这一段内容应当被过滤掉。此外，在一些比较简单的例子中，原有的影音文件可能并不包含这些解释性文字信息，这样就不需要进行上述过滤步骤。此外，还需要特别说明的是，上述步骤 201 与 203 并没有特别的先后限制，二者的顺序是可以互换的。

在步骤 205，需要将文字信息和语音信息进行对齐，即将一段文字信息与一段源语音信息的起始和终止时间对应。对齐之后才能准确提取相应的源语音信息作为对某一句文字信息的语音训练数据，进行标准说话人选择，音色转换，以及定位最终的目标语音信息的时间位置的步骤。在一种实施例中，如果字幕信息本身就包含某一段文本对应的音频流（即源语音

信息）的时间起点和终点（现有的影音文件大多是这样一种情况），可以利用这一时间信息将文字和源语音信息进行对齐，这样可以大大提高对应的精度。在另一种实施例中，如果该段文字中没有精确标定相应的时间信息，仍然可以通过语音识别技术将相应的源语音转换为文字然后寻找匹配的字幕信息，并在该字幕信息上标定源语音的起始和终止时间点。本领域的普通技术人员理解，任何其它的有助于实现语音和文字对齐的算法都在本发明的保护范围之内。

有时，字幕信息有可能出现标定错误，这是由于原始的影音文件制造者造成的文字和源语音的不匹配，一种简单的纠正方法就是当检查到文字信息和语音信息不匹配时，过滤不匹配的文字和语音信息（步骤 207）。注意这种匹配度检查所关注的是英文的源语音和英文的源字幕，因为用同一语言进行检查会大大降低计算的成本和难度，只要将源语音转换为文字然后与英文源字幕进行匹配计算，或者将源英文字幕转换为语音然后与英文源语音进行匹配计算就可以实现。当然，对于一段很简单的，字幕和语音能够很好对应的影音文件，可以省略上述匹配步骤。

下面在步骤 209、211、213 中，进行不同说话人的数据分割。在步骤 209 中判断源文字信息中的说话人角色是否已被标定。如果字幕信息中已经标定了说话人的信息，则可以根据这一信息容易的将不同说话人所对应的文字信息和语音信息进行分割。如:

Tom: I'm afraid I can't go to the meeting tomorrow。

这里直接用 Tom 标识出说话人的角色，这样就可以直接将对应的语音和文字信息作为说话人 Tom 的训练数据，从而按照不同的角色对说话人的语音和文字信息进行分割（步骤 211）。相反，如果字幕信息中没有标定说话人的信息，则还需要对说话人的语音信息和文字信息进行额外的分割（步骤 213），即对说话人进行自动分类。具体而言，可以利用说话人的频谱和韵律特征对所有的源语音信息进行自动分类，从而形成若干个类。这样就得到每一个类的训练数据。之后可以赋予每一个类一个特定的说话人标识，如“角色 A”。需要指出的是，自动的分类的结果可能使不同的

说话人被归为一类，因为他们的声音特征极为相似，也可能使同一说话人的不同语音被分为多类，因为该说话人在不同语境下的声音特征表现出明显差异（比如在愤怒时和高兴时的语音特征相差很远）。但是这样的归类并不会过分影响最终的配音效果，因为后续的音色转换过程仍然会使输出的目标语音接近源语音的发音。

在步骤 215 中，经过处理的文字信息和源语音信息可以作为训练数据待用。

图 3 为选择语音合成库中说话人类别的流程图。如上所述，选择标准说话人的目的是为了使语音合成步骤中所使用的标准说话人声音与源声音尽量接近，从而减少后续音色转换步骤所带来的音质损伤。正是因为标准说话人选择的过程直接决定了后续音色转换的优劣，因此具体的标准说话人的选择方法与音色转换的方法相关。为了寻找与源声音声学特征差异最小的标准说话人声音，大致可以利用以下两个因素对声音特征的差异进行度量：一是声音的基频差异（也称韵律上的差异），通常用 F_0 表示，二是声音的频谱差异，通常用 $F_1...F_n$ 表示。在一个自然的复合音里，有一个振幅最大、频率最低的分音，一般被称为“基音“，他的振动频率被成为“基频”。一般来说，对音高的感知主要决定于基频。由于基频反应的是声带振动的频率，其与具体的说话内容无关，因此也称为超音段特征，而频谱 $F_1...F_n$ 反应的是声道的形状，其与具体的说话内容有关，因此也被称为音段特征。这两种频率共同定义了一段声音的声学特征。本发明分别根据这两种特征选择声音差异最小的标准说话人。

在步骤 301 中，计算标准说话人的语音与源说话人的语音之间的基频差异。具体而言，参考图 4，在步骤 401 中准备源说话人（如 Tom）与多个标准说话人（如女 1、女 2、男 1、男 2）的语音训练数据。

在步骤 403 中提取源说话人和目标说话人的相应于多个浊音段的基频 F_0.。然后分别计算对数域基频 $\log(F_0)$ 的均值和/或方差（步骤 405）。并针对每一个标准说话人，计算其基频平均值和/或方差与源说话人的基频均值和/或方差的差异，并且计算这两种差异的加权距离和（步骤 407），

进而判断选择哪个说话人作为标准说话人。

图 5 显示了源说话人与标准说话人基频差异的均值比较。假设源说话人和标准说话人的基频均值如表 1 所示:

	源说话人	女 1	女 2	男 1	男 2
基频均值（HZ）	280	300	260	160	100

表 1

从表 1 中可以容易的看出，源说话人的基频更接近女 1 和女 2，而与男 1 和男 2 的均值相差很远。

然而，如果源说话人的基频均值与两个以上标准说话人的基频均值差异相同（如图 5 所示），或者比较接近，则可以进一步计算源说话人和标准说话人的基频方差。方差是衡量基频的变化范围的指标。在图 6 中把上述三个说话人的基频方差进行了比较，发现源说话人的基频方差（10HZ）与女 1（10HZ）的相同，而与女 2（20HZ）的相异，因此可以选择女 1 作为源说话人在语音合成过程中所使用的标准说话人。

本领域的普通技术人员理解，对于上述基频差异的度量方法并不局限于本说明书中所列举的例子而是可以进行各种各样的变形，只要其能够保证所筛选出来的标准说话人声音在后续的音色转换中所带来的音质损伤最小。在一种实施例中，所述音质损伤的度量可以按照下面给出的公式计算:

$$d(r)=\begin{cases} a_{+}*r^{2}, & r>0 \\ a_{-}*r^{2}, & r<0 \end{cases}$$

其中 d（r）表示音质损伤，$r=\log(F_{0S}/F_{0R})$，F_{0S} 表示源语音的基频均值，F_{0R} 表示标准语音的基频均值。a_{+} 和 a_{-} 分别为两个经验常量。可见，基频均值差异（F_{0S}/F_{0R}）虽然与音色转换的音质损伤存在一定联系，但是并不一定是正比的关系。

返回到图 3 的步骤 303，还要进一步计算标准说话人和源说话人的频谱差异。

下面参考图 7 详细说明计算标准说话人与源说话人的频谱差异的过程。如前文所述，共振峰（format）是指在发音时由于声道本身的共振而在声音频谱中形成的声音强度较大的若干频率区域。说话人的声音特征主要反应在前四个共振峰频率上，即 F_1、F_2、F_3、F_4。一般而言，第一共振峰 F_1 的取值范围在 300－700HZ 范围内，第二共振峰 F_2 的取值范围在 1000－1800HZ 的范围内，第三共振峰 F_3 的取值范围在 2500－3000HZ 的范围内，第四共振峰 F_4 的取值范围在 3800－4200HZ 的范围内。

本发明通过比较源说话人和标准说话人在若干共振峰上的频谱差异从而选择可能引起音质损伤最小的标准说话人。具体而言，在步骤 701，首先提取源说话人的语音训练数据，然后在步骤 703，准备与源说话人对应的标准说话人的语音训练数据。这些训练数据不要求内容完全相同，但需要包含相同或者相似的特征音素。

接下来，在步骤 705 从标准说话人和源说话人的语音训练数据中选择相对应的语音段，以及将所述语音段进行帧对齐。其中所述相对应的语音段在源说话人和标准说话人的训练数据中具有相同或者相似上下文的相同或者相似音素。此处所说的上下文包括但不限于：相邻的语音、在词中的位置、在词组中的位置、在句子中的位置等。如果找到了多对具有相同或相似上下文的音素，则可优选某些特征音素例如[e]。如果所找到的多对具有相同或相似上下文的音素是彼此相同的，则可优选某些上下文。因为，在某些上下文中，音素的共振峰较小可能受相邻音素的影响。例如，选择具有“爆破音”或“摩擦音”或“静音”作为其相邻音素的语音段。如果所找到的多对具有相同或相似上下文的音素中，彼此的上下文和音素均相同，则可随机选择一对语音段。

之后，对所述语音段进行帧对齐：在一种实施例中，将标准说话人的语音段的中间的帧与源说话人的语音段的中间的帧对齐。由于中间的帧被认为变化较小，因此它较小受相邻音素的共振峰的影响。在这一实施例中，该对中间的帧即被选择作为最佳帧（参见步骤 707），从而用于在后续步骤中提取共振峰参数。在另一种实施例中，还可以采用已知的动态时间规

整算法 DTW 进行帧对齐，从而获得多个对齐的帧，并且优选具有最小声学差异的对齐的帧，作为一对对齐的最佳帧（参见步骤 707）。总之，在步骤 707 中所得到的对齐帧具有这样的特点，每个帧都能较好的表达其说话人的声学特征，该对帧之间的声学差异相对较小。

之后，在步骤 709 中，提取所选择的一对帧的匹配的共振峰参数组。可使用任何已知的用于从语音中提取共振峰参数的方法来提取所述相匹配的共振峰参数组。共振峰参数的提取可自动进行，也可以手动进行。一种可能的方式是使用某种语音分析工具例如 PRAAT 来提取共振峰参数。在提取对齐的帧的共振峰参数时，可利用相邻帧的信息来使提取出的共振峰参数更为稳定可靠。在本发明的一种实施例中，将一对匹配的共振峰参数组中的各对匹配的共振峰参数用作关键点来生成一个频率弯曲函数。源说话人的共振峰参数组为[F_{1S}, F_{2S}, F_{3S}, F_{4S}]，标准说话人的共振峰参数组为[F_{1R}, F_{2R}, F_{3R}, F_{4R}]，表 2 中示出了源说话人和标准说话人的共振峰参数示例。虽然本实施例选择第一到第四共振峰作为共振峰参数，因为这些参数已经可以代表某说话人的语音特征，但是本发明并不限于提取更多的、更少的、或其它的共振峰参数。

	第一共振峰（F_1）	第二共振峰（F_2）	第三共振峰（F_3）	第四共振峰（F_4）
标准说话人频率[F_R](HZ)	500	1500	3000	4000
源说话人频率[F_S]（HZ）	600	1700	2700	3900

表 2

在步骤 711 中，根据上述共振峰参数，计算每一个标准说话人与源说话人之间的距离。下面列举两种实施方法来实现该步骤。在第一种实施方法中，直接求解对应共振峰参数之间的加权距离和，并且可以给前三个共振峰频率相同的权重 $W_{高}$，而赋予第四个共振峰频率较低的权重 $W_{低}$，以区分不同共振峰频率对声学特征的不同影响。表 3 表示了按照第一种实施方法所计算的标准说话人与源说话人之间的距离。

	第一共振峰（F_1）	第二共振峰（F_2）	第三共振峰（F_3）	第四共振峰（F_4）
标准说话人频率（F_R）	500	1500	3000	4000
源说话人频率（F_S）	600	1700	2700	3900
共振峰频率差异	100	200	-300	-100
共振峰频率差异权重	$W_{高}=100\%$	$W_{高}=100\%$	$W_{高}=100\%$	$W_{低}=50\%$
两种说话人共振峰频率距离和	（100 + 200 + |-300|）× $W_{高}$ + （|-100|）× $W_{低}$ = 650 这里的差异是绝对值之和			

表 3

在第二种实施方法中，使用匹配的共振峰参数对[F_R, F_S]作为关键点来定义一从源说话人频率轴映射到标准说话人频率轴的分段线性函数。然后计算这条分段线性函数与函数 Y = X 之间的距离。具体而言，可以对两条曲线函数沿 X 轴进行采样得到各自的 Y 值，计算各个采样点 Y 值之间的加权距离和。X 轴的采样可以使用等间隔采样，也可以使用不等间隔采样，如 log 域等间隔采样，或 mel 频谱域等间隔采样。图 8 为第二种实施方法中源说话人与标准说话人频谱差异按照等间隔采样的分段线性函数示意图。由于函数 Y = X 为一条沿 X 轴、Y 轴对称的直线（图中未示出），因此图 8 中所示的分段线性函数与函数 Y = X，在每一个标准说话人共振峰频率[F_{1R}, F_{2R}, F_{3R}, F_{4R}]点上的 Y 值差异反映了源说话人共振峰频率与标准说话人共振峰频率之间的差异。

通过上面的实施方法可以得到一个标准说话人与源说话人之间的距离，即声音频谱差异。通过重复上面步骤就可以计算每一个标准说话人，如[女 1、女 2、男 1、男 2]与源说话人之间的声音频谱差异。

返回图 3 中的步骤 305，根据上述基频差异和频谱差异计算二者的加权距离和，从而选择声音与源说话人最贴近的标准说话人（步骤 307）。本领域的普通技术人员理解，虽然本发明以共同计算基频差异和频谱差异为例进行说明，该方法仅仅构成本发明的一种优选实施例，本发明还可以

实现各种变形：比如可以仅仅根据基频差异选择标准说话人；或者仅仅根据频谱差异选择标准说话人；或者先根据基频差异选出一部分标准说话人，再根据频谱差异从选出的标准说话人中进一步进行筛选；或者先根据频谱差异选出一部分标准说话人，再根据基频差异从选出的标准说话人中进一步进行筛选。总之，标准说话人选择的目的是为了选出与源说话人声音差异相对最小的标准说话人声音，从而在后续的音色转换过程中使用引起音质损伤最小的标准说话人声音进行音色转换，或称声音模拟。

图 9 示出了将所述源文字信息合成为标准语音信息的流程。首先，在步骤 901 中，选择待合成的一段文字信息，比如电影中的一段字幕“*我恐怕不能参加明天的会议了*”。接着，在步骤 903 中，对上述文字信息进行分词（lexical word segmentation），分词是语言信息处理的前提，其主要目的是把一句话按照自然的说话规律拆分成若干词或字组成（比如“*[我][恐怕][不能][参加][明天的][会议]了*”）。分词的方法有很多种，两种比较基本的分词方法是：基于词典的分词方法和基于频度统计的分词方法。当然本发明并不排除使用任何其它方法进行分词。

接下来，在步骤 905，对经过分词的文字信息进行韵律预测（prosodic structure prediction），可估计合成语音的语调、节奏、重音的位置和时长信息等。然后在步骤 907 从语音合成库调用相应的语音信息，也就是按照韵律预测的结果选择一个标准说话人的语音单元拼接在一起，从而以标准说话人的声音自然流畅的说出上述文字信息。上述语音合成的过程通常被成为拼接合成，尽管本发明以此为例进行说明，实际上本发明并不排除任何其它可以用于进行语音合成的方法，如参数合成等。

图 10 为将标准语音信息按照源语音信息进行音色转换的流程图。由于当前的标准语音信息已经能够准确的按照字幕说出自然流畅的声音，图 10 的方法将会使这一标准语音更加接近源语音的声音。首先，在步骤 1001 中，对所选出的标准语音文件和源语音文件进行语音分析，从而得到标准说话人和源说话人的基频和频谱特征，包括基频频率[F_0]，和共振峰频率[F_1, F_2, F_3, F_4 等]。如果在之前的步骤中已经得到了上述信息，则可以直接

加以利用，而无需重新提取。

然后，在步骤 1003 和 1005 中分别按照源语音文件对标准语音文件进行频谱转换和/或基频调整。通过前文的描述可以得知，利用源说话人和标准说话人的频谱参数可以产生一个频率弯曲函数（参见图 8），将频率弯曲函数应用于标准说话人的语音段中，从而将标准说话人的频谱参数转换为与源说话人的频谱参数一致，就能够得到高相似度的转换语音。如果标准说话人与源说话人声音差别很小，所述频率弯曲函数就更接近一条直线，转换后的语音质量就较高；相反，如果标准说话人与源说话人声音差别很大，所述频率弯曲函数就会更加曲折，转换后的语音质量就会相对下降。在这上述步骤中由于所选择的待转换的标准说话人已经和源说话人声音大致相近，因此音色转换所带来的音质损伤可以被显著减小，从而保证在转换后的语音相似度的同时，提高语音质量。

同理，也可以利用源说话人[F_{0S}]和标准说话人[F_{0R}]的基频参数产生一个基频线性函数,如 $\log F_{0S}=a+b\log F_{0R}$,其中 a 和 b 为常数，这样的基频线性函数较好地反应了源说话人和标准说话人的基频差异，并且利用这样的线性函数可以将标准说话人的基频转换为源说话人的基频。在优选实施例中，基频调整和频谱转换可以不分先后，共同进行，但是本发明并不排除仅仅进行其中的基频调整或频谱转换。

在步骤 1007 中，将按照前述的转换和调整结果，重构标准语音数据，从而产生目标语音数据。

图 11 为自动语音转换系统 1100 的结构框图。在一种实施例中，影音文件 1101 含有不同的轨道，包括音频轨道 1103、字幕轨道 1109 和视频轨道 1111，其中音频轨道 1103 又进一步包括背景(backgroud)声音信道 1105 和前景(foreground)声音信道 1107。背景声音信道 1105 一般存储背景音乐、特殊音效等非说话语音信息，而前景声音信道 1107 一般存储说话人的语音信息。训练数据获得单元 1113 用于获得语音和文字训练数据，并进行相应的对齐处理等。在本实施例中，标准说话人选择单元 1115 利用训练数据获得单元 1113 所获得的语音训练数据从标准语音信息库 1121 中选择合适的

标准说话人。语音合成单元 1119 将文字训练数据按照标准说话人选择单元 1115 所选择的标准说话人声音进行语音合成。音色转换单元 1117 将标准说话人的声音按照源说话人的声音训练数据进行音色转换。同步单元 1123 将音色转换后的目标语音信息与源语音信息或者视频轨道 1111 中的视频信息进行同步。最后，背景声音息、经过自动语音转换后的目标语音信息和视频信息被合成为目标影音文件 1125。

图 12 为带有自动语音转换系统的影音文件配音装置的结构框图。在该图所示的实施例中，载有中文字幕的英文影音文件存储在盘 A 中，影音文件配音装置 1201 包括自动语音转换系统 1203 和目标盘生成器 1205。自动语音转换系统 1203 用于从盘 A 中获得合成的目标影音文件，目标盘生成器 1205 用于将目标影音文件写入目标盘 B。目标盘 B 中载有经过自动中文配音的目标影音文件。

图 13 为带有自动语音转换系统的影音文件播放器的结构框图。在该图所示的实施例中，载有中文字幕的英文影音文件存储在盘 A 中，影音文件播放器 1301，如 DVD 播放器，利用自动语音转换系统 1303 从盘 A 中获得合成的目标影音文件，并且直接传送至电视机或计算机中进行播放。

本领域的普通技术人员理解，虽然本发明以为影音文件进行自动配音为例进行说明，但本发明并不仅限于这一应用场景，任何需要将文字信息转换为特定说话人声音的应用场景都在本发明的保护范围之列，比如在虚拟世界的游戏软件中，游戏者可以通过本发明将输入的文字信息按照喜欢的角色转换成特定语音信息；本发明还可用于使电脑机器人模拟真人声音播报新闻。

另外，上述各个操作过程可以由存储在计算机程序产品中的可执行程序实现。该程序产品定义各实施例的功能，并承载各种信号，包括（但不局限于）：1）永久存储在不可擦写存储媒体上的信息；2）存储在可擦写存储媒体上的信息；或 3）通过包括无线通信在内的通信媒体（如，通过计算机网络或电话网络）传送到计算机上的信息，特别是包括从因特网和其它网络下载的信息。

本发明的各种实施例可以提供许多优点，包括已经在发明内容中列举的，和能够从技术方案本身推导出来的。但是无论一个实施例是否取得全部优点，并且也无论这样的优点是否被认为是取得实质性提高，都不应构成对本发明的限制。同时，上文中提到的各种实施方式，仅仅是出于说明的目的，本领域的普通技术人员可以对上述实施方式做出各种修改和变更，而不偏离本发明的实质。本发明的范围完全由所附权利要求书限定。

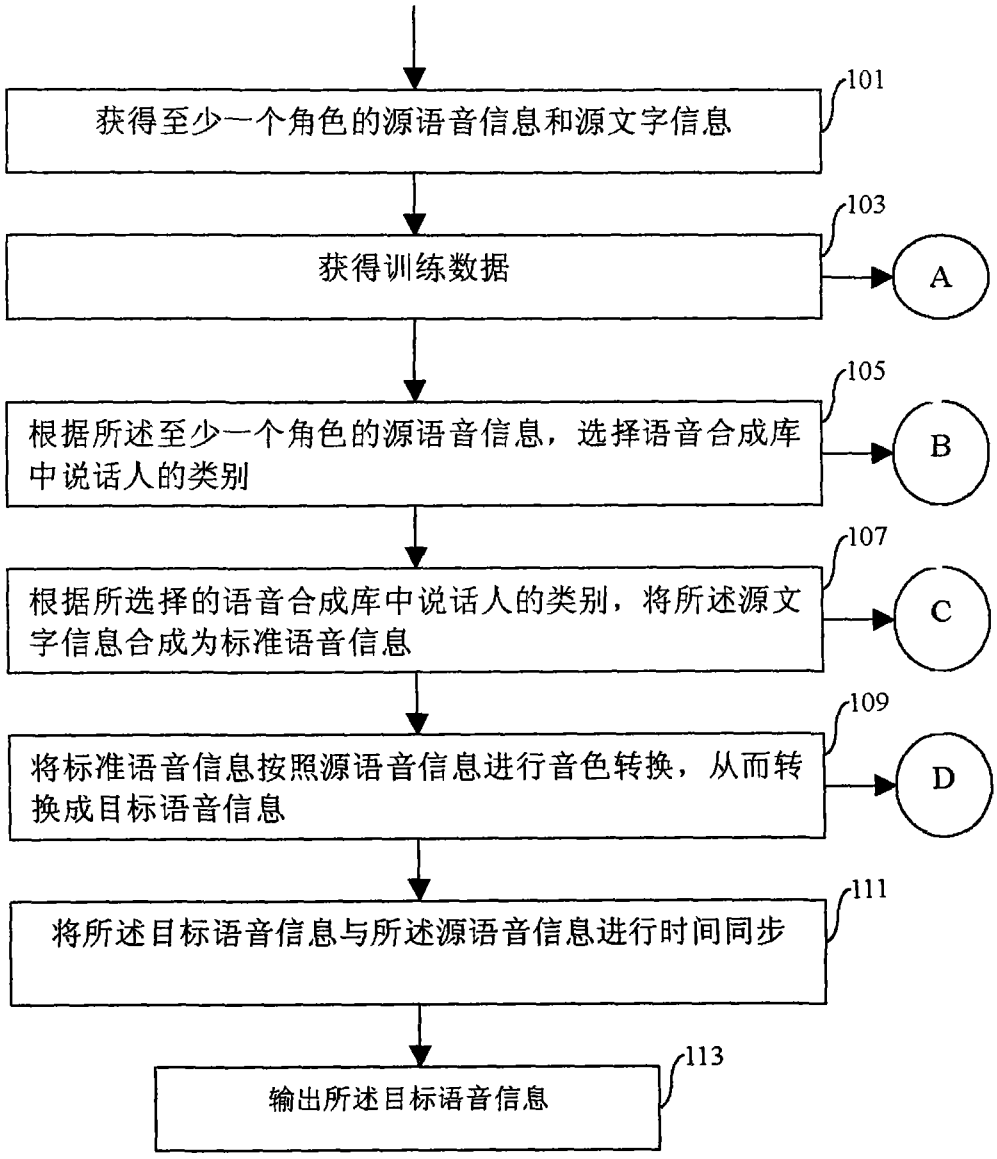

101
获得至少一个角色的源语音信息和源文字信息
103
获得训练数据
A
105
根据所述至少一个角色的源语音信息，选择语音合成库中说话人的类别
B
107
根据所选择的语音合成库中说话人的类别，将所述源文字信息合成为标准语音信息
C
109
将标准语音信息按照源语音信息进行音色转换，从而转换成目标语音信息
D
111
将所述目标语音信息与所述源语音信息进行时间同步
113
输出所述目标语音信息

图 1

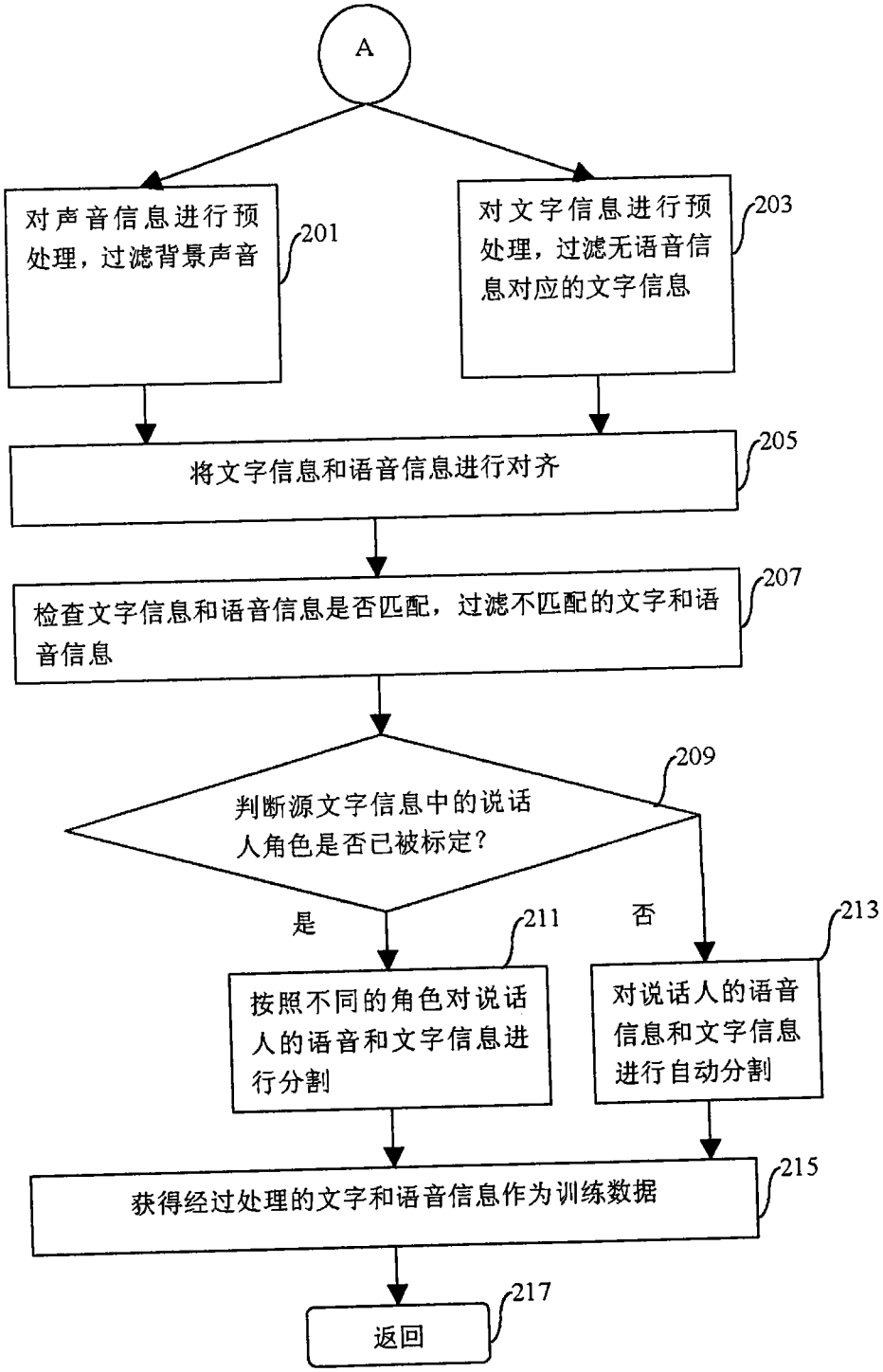

A
对声音信息进行预处理，过滤背景声音
201
对文字信息进行预处理，过滤无语音信息对应的文字信息
203
将文字信息和语音信息进行对齐
205
检查文字信息和语音信息是否匹配，过滤不匹配的文字和语音信息
207
判断源文字信息中的说话人角色是否已被标定？
209
是
否
按照不同的角色对说话人的语音和文字信息进行分割
211
对说话人的语音信息和文字信息进行自动分割
213
获得经过处理的文字和语音信息作为训练数据
215
返回
217

图 2

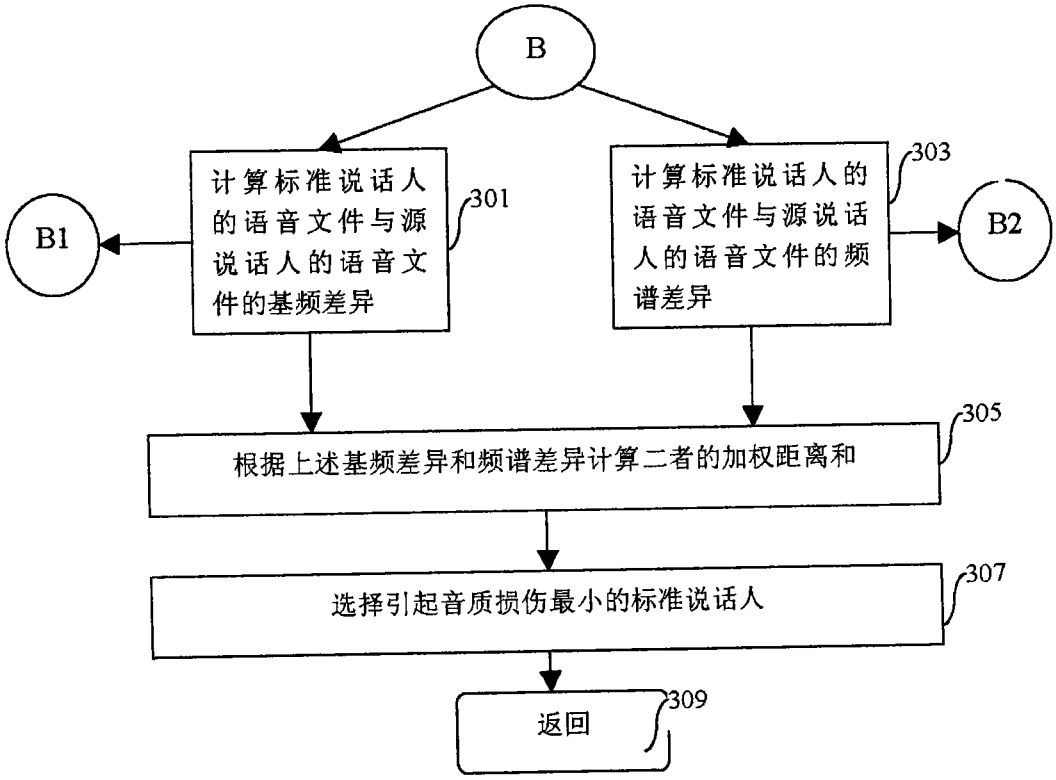

B
计算标准说话人的语音文件与源说话人的语音文件的基频差异
301
计算标准说话人的语音文件与源说话人的语音文件的频谱差异
303
B1
B2
根据上述基频差异和频谱差异计算二者的加权距离和
305
选择引起音质损伤最小的标准说话人
307
返回
309

图 3

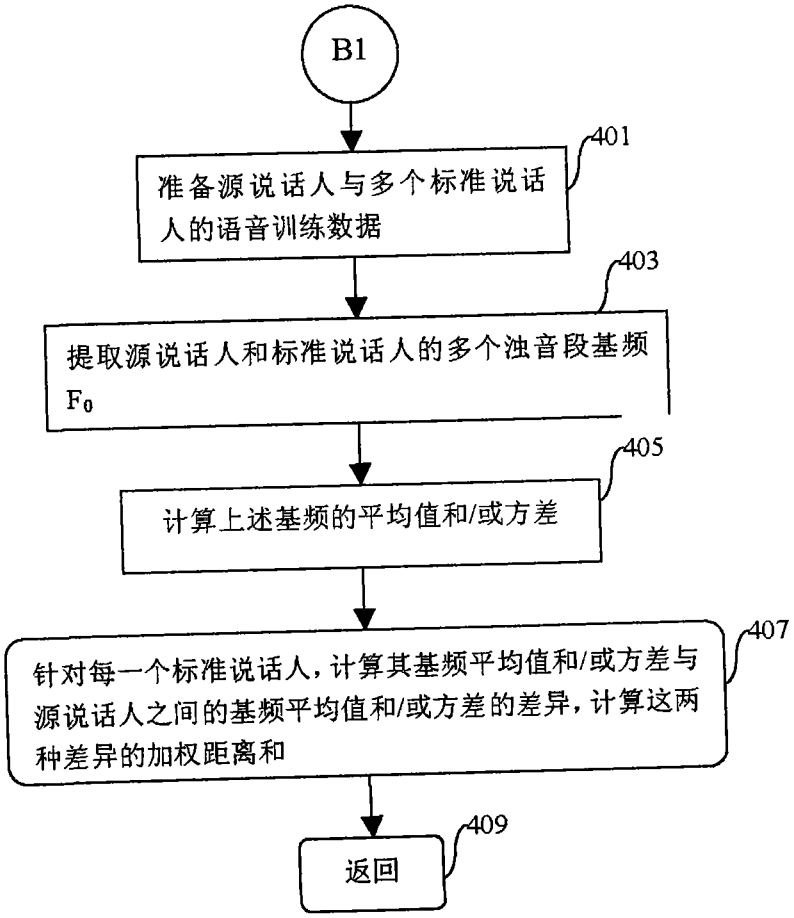

B1
401
准备源说话人与多个标准说话人的语音训练数据
403
提取源说话人和标准说话人的多个浊音段基频 F_0
405
计算上述基频的平均值和/或方差
407
针对每一个标准说话人，计算其基频平均值和/或方差与源说话人之间的基频平均值和/或方差的差异，计算这两种差异的加权距离和
409
返回

图 4

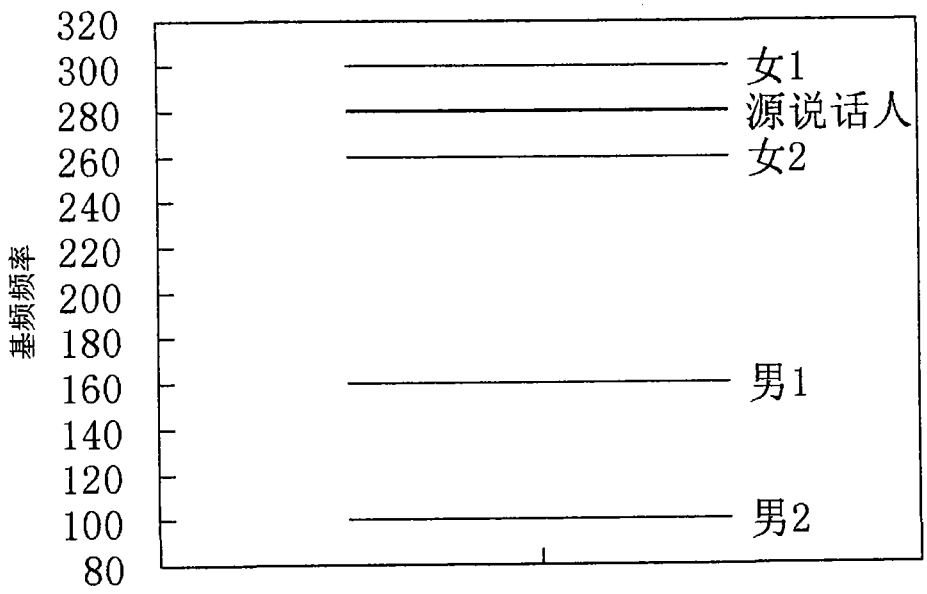

320
300
280
260
240
220
200
180
160
140
120
100
80
基频频率
女1
源说话人
女2
男1
男2

图 5

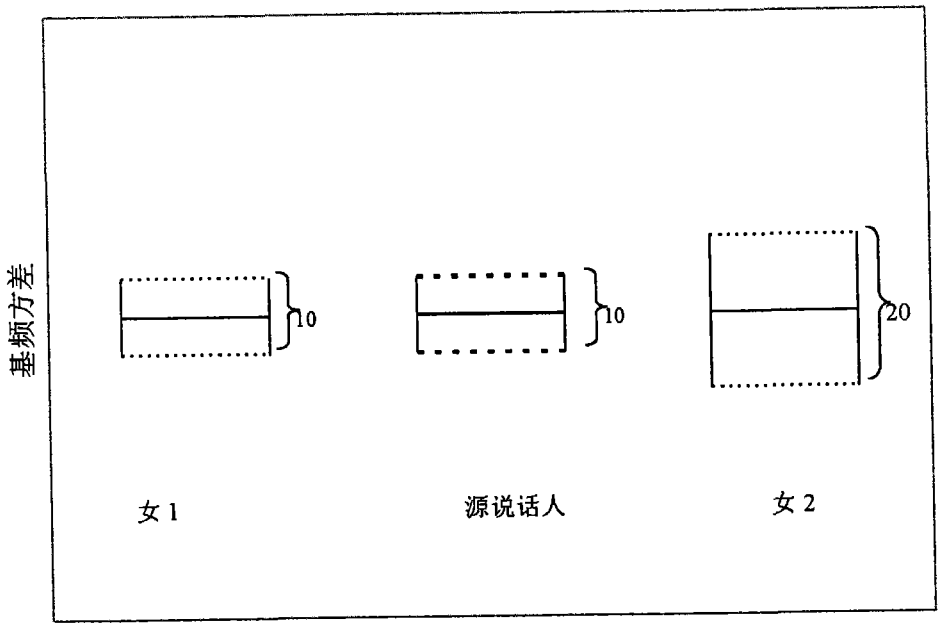

基频方差
10
10
20
女 1
源说话人
女 2

图 6

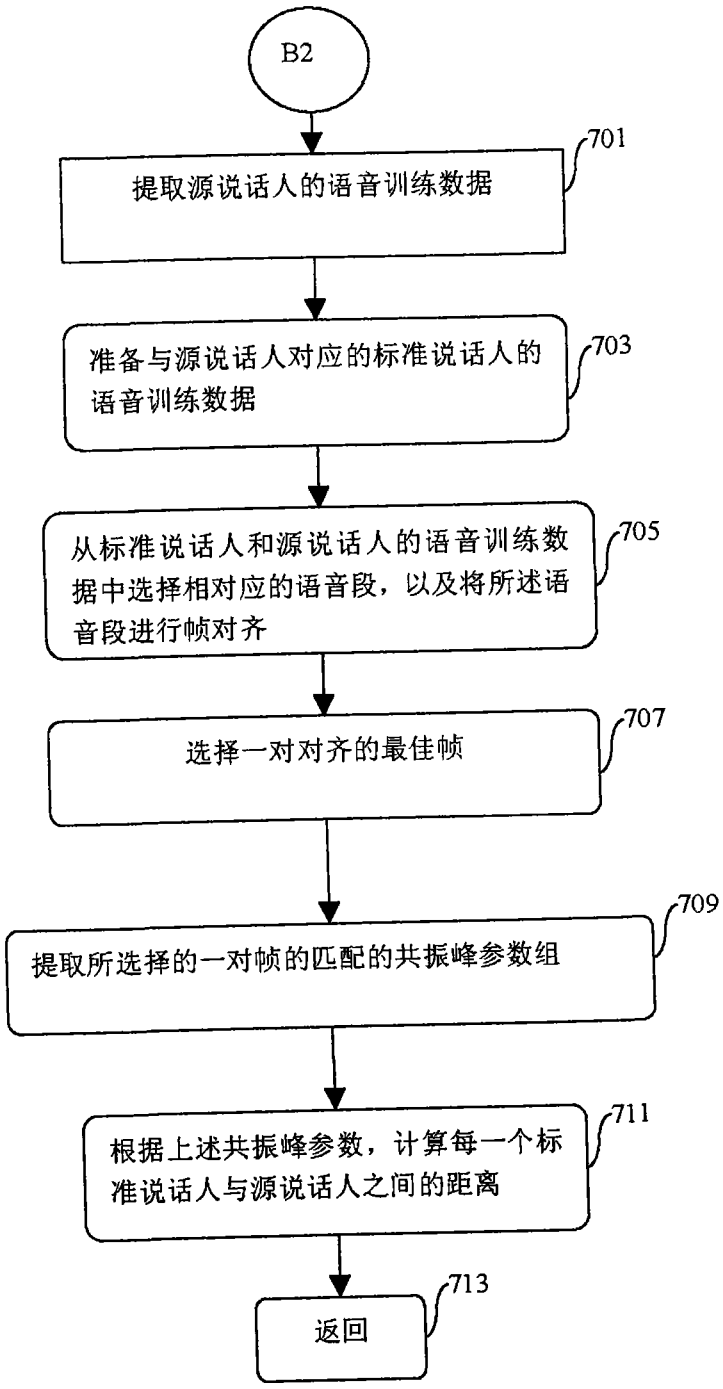

B2
701
提取源说话人的语音训练数据
703
准备与源说话人对应的标准说话人的语音训练数据
705
从标准说话人和源说话人的语音训练数据中选择相对应的语音段，以及将所述语音段进行帧对齐
707
选择一对对齐的最佳帧
709
提取所选择的一对帧的匹配的共振峰参数组
711
根据上述共振峰参数，计算每一个标准说话人与源说话人之间的距离
713
返回

图 7

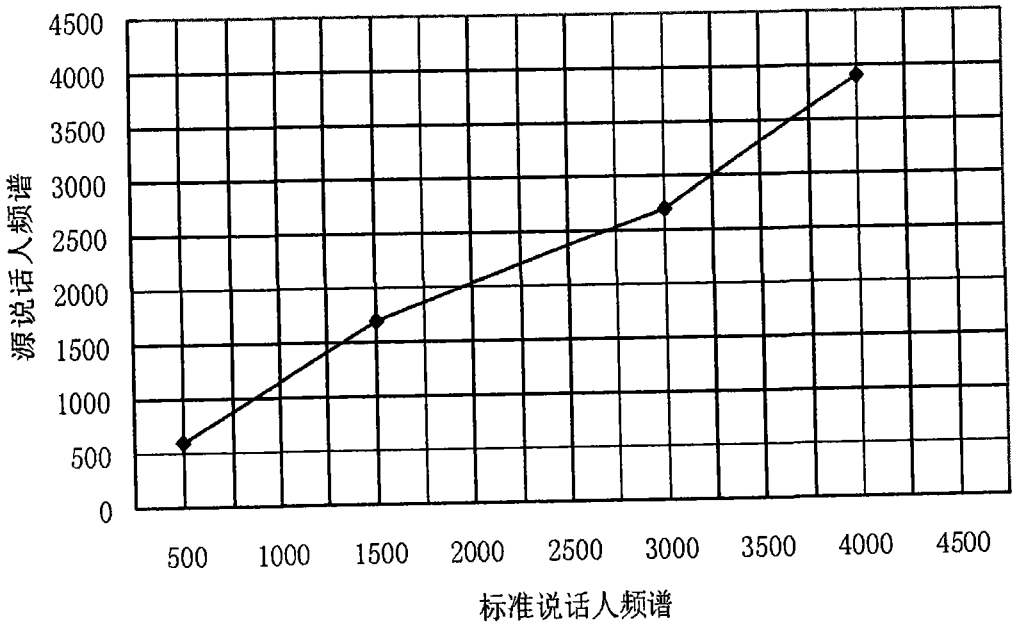

4500
4000
3500
3000
2500
2000
1500
1000
500
0
源说话人频谱
500
1000
1500
2000
2500
3000
3500
4000
4500
标准说话人频谱

图 8

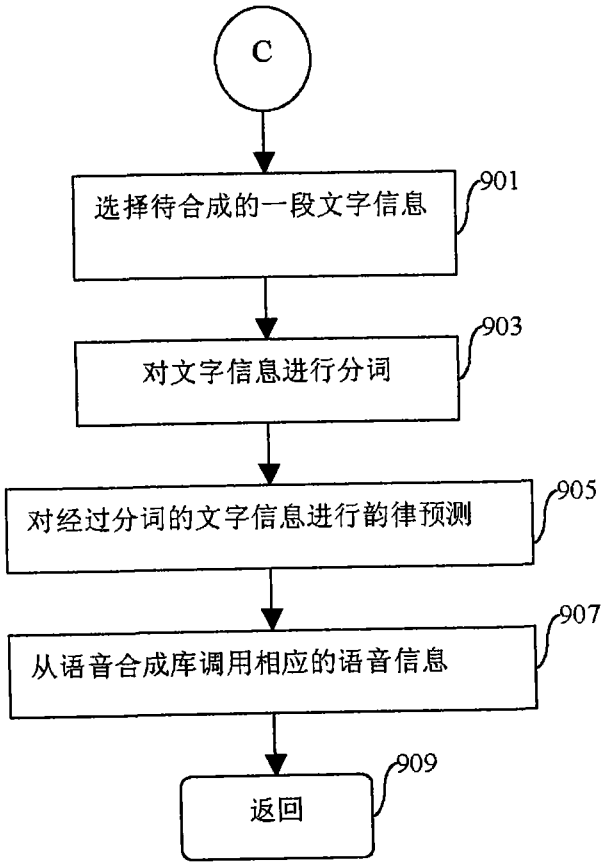

C
选择待合成的一段文字信息
901
对文字信息进行分词
903
对经过分词的文字信息进行韵律预测
905
从语音合成库调用相应的语音信息
907
返回
909

图 9

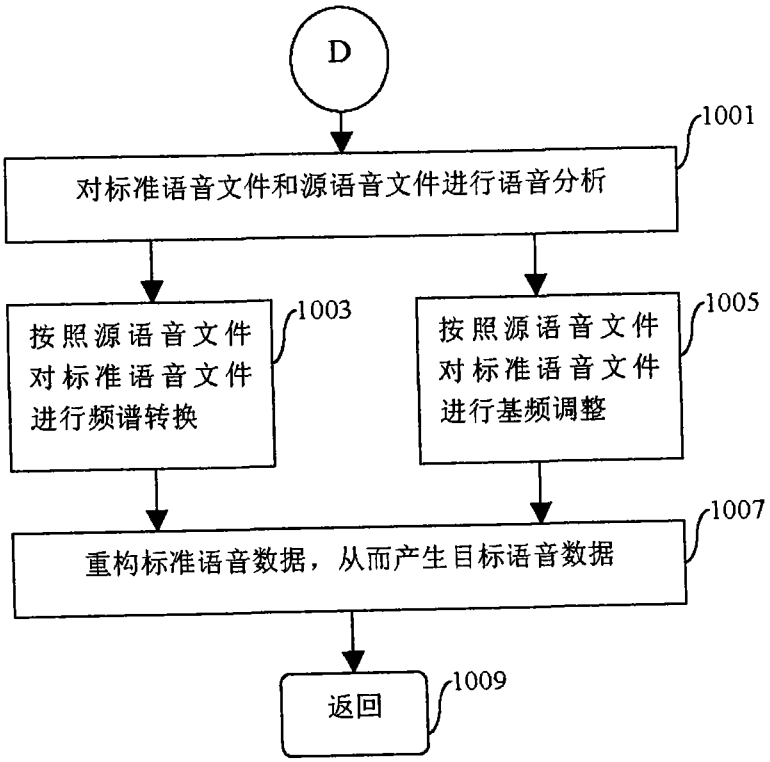

D
1001
对标准语音文件和源语音文件进行语音分析
1003
按照源语音文件对标准语音文件进行频谱转换
1005
按照源语音文件对标准语音文件进行基频调整
1007
重构标准语音数据，从而产生目标语音数据
1009
返回

图 10

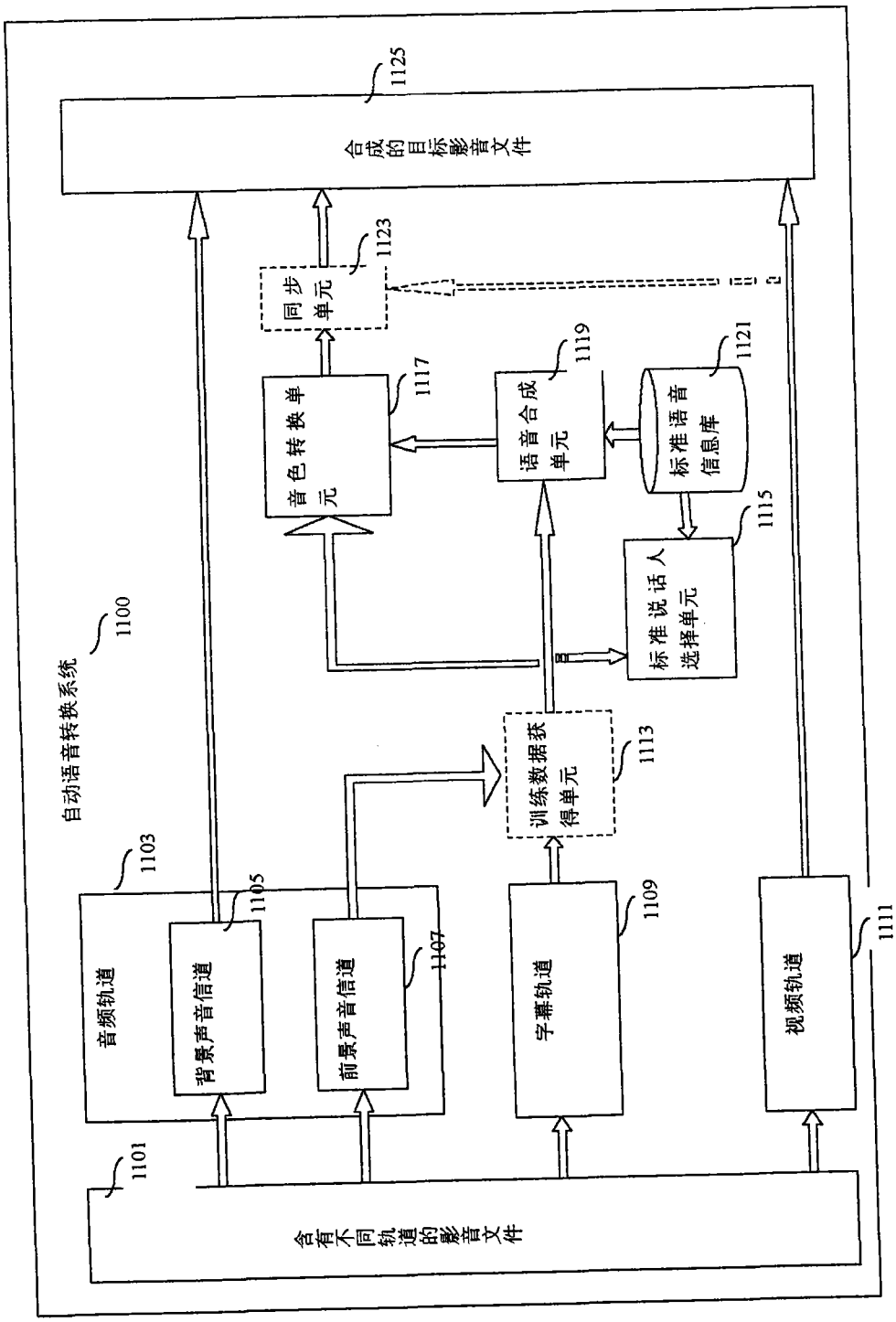

自动语音转换系统
1100
含有不同轨道的影音文件
1101
音频轨道
1103
背景声音信道
1105
前景声音信道
1107
字幕轨道
1109
视频轨道
1111
训练数据获得单元
1113
标准说话人选择单元
1115
音色转换单元
1117
语音合成单元
1119
标准语音信息库
1121
同步单元
1123
合成的目标影音文件
1125

图 11

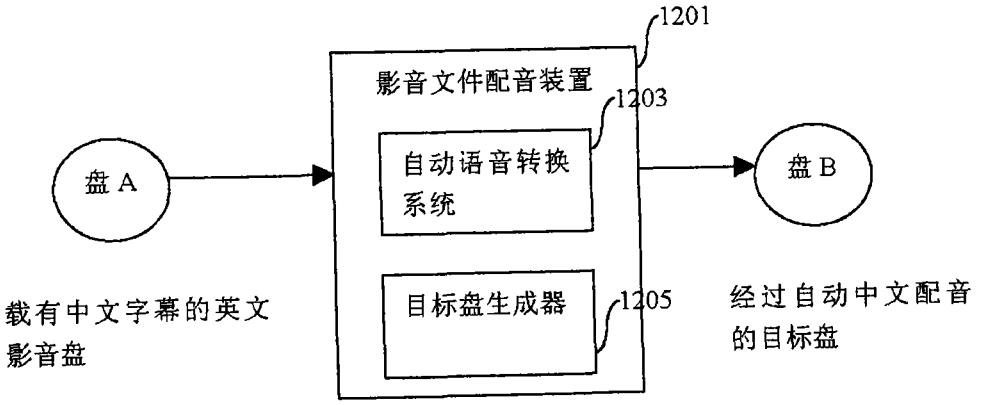

1201
影音文件配音装置
1203
自动语音转换系统
盘A
盘B
目标盘生成器
1205
载有中文字幕的英文影音盘
经过自动中文配音的目标盘

图 12

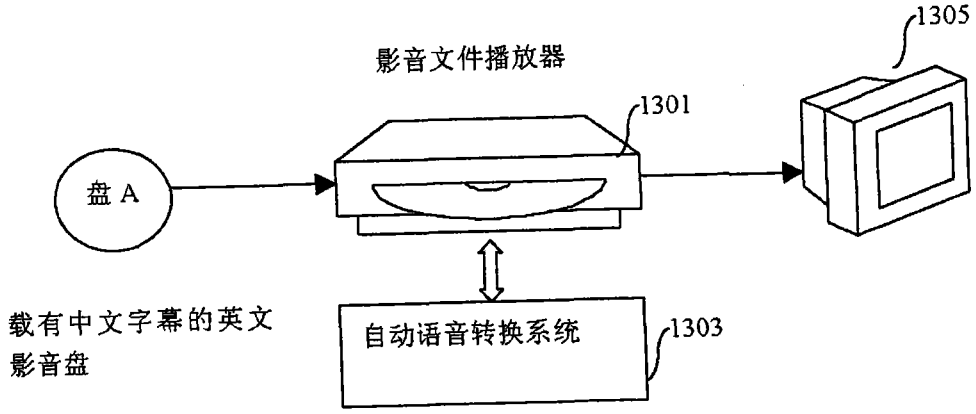

1305
影音文件播放器
1301
盘A
载有中文字幕的英文影音盘
自动语音转换系统
1303

图 13