(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108449629 A
(43)申请公布日 2018.08.24

(21)申请号 201810277980.8

(22)申请日 2018.03.31

(71)申请人 湖南广播电视台广播传媒中心
地址 410000 湖南省长沙市开福区三一大道455号湖南广播电视台广播传媒中心技术大楼

(72)发明人 牛嵩峰 周晓民 唐炜

(74)专利代理机构 长沙市融智专利事务所 43114
代理人 颜勇

(51)Int.Cl.
H04N 21/439(2011.01)
H04N 21/43(2011.01)
H04N 21/8547(2011.01)
H04N 5/262(2006.01)
G10L 15/26(2006.01)
G10L 15/22(2006.01)

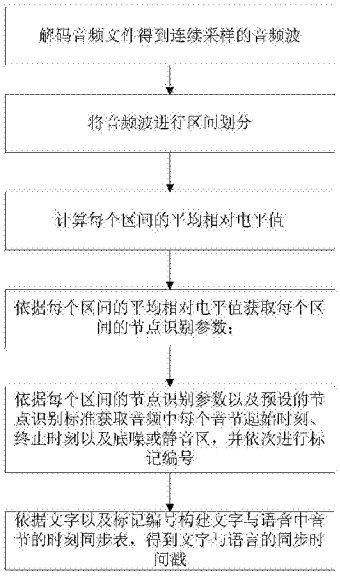
权利要求书2页 说明书10页 附图4页

(54)发明名称

一种音频的语音与文字同步方法及剪辑方法和剪辑系统

(57)摘要

本发明公开了一种音频的语音与文字同步方法及剪辑方法和剪辑系统，同步方法包括：步骤1：解码音频文件得到音频波形，并将音频同步翻译为文字；步骤2：将音频波形进行区间划分；步骤3：计算每个区间的平均相对电平值；步骤4：依据每个区间的平均相对电平值获取每个区间的节点识别参数；步骤5：依据每个区间的节点识别参数以及预设的节点识别标准获取音频中每个音节起始时刻、终止时刻以及底噪或静音区，并依次进行标记编号；步骤6：构建文字与语音中音节的时刻同步表，得到文字与语音的同步时间戳。本发明通过上述方法可以极大地提高音频剪辑效率。



CN 108449629 A

1.一种音频的语音与文字同步方法,其特征在于:包括如下步骤:

步骤1:解码音频文件得到连续采样的音频波形,并将音频同步翻译为文字;

其中,所述音频波形以时间为横轴坐标、幅度为纵轴坐标,所述音频波形上分布离散采样点;

步骤2:将步骤1中的音频波形进行区间划分;

其中,每个区间包括x个采样点,音频波形的采样频率低于或等于48kHz,x的取值范围为50-150,音频波形的采样频率为96kHz或88.2kHz,x的取值范围为100-300,音频波形的采样频率为192kHz,x的取值范围为200-600;

步骤3:计算每个区间的平均相对电平值;

其中,区间的平均相对电平值计算公式如下:

$$Y_x^n(dB)=20\log\left\{\left[\left(\left|A_1^n\right|+\left|A_2^n\right|+...+\left|A_x^n\right|\right)/x\right]/X\right\}$$

X=0000,0000,0000,0001

式中,$Y_x^n(dB)$表示第n个区间的平均相对电平值,$\left|A_1^n\right|$、$\left|A_2^n\right|$、$\left|A_x^n\right|$分别表示第n个区间中第1、2、x个采样点的幅度的绝对值,X表示量化比特数为16位的预设的取样信号,n为正整数;

步骤4:依据每个区间的平均相对电平值获取每个区间的节点识别参数;

其中,所述节点识别参数包括电平参数和电平变化参数:

$$\mu_x^n=Y_x^n(dB)$$

$$\beta_x^n=Y_x^n(dB)-Y_x^{n-1}(dB)$$

式中,μ_x^n表示第n个区间的电平参数,β_x^n表示第n个区间的电平变化参数;

步骤5:依据每个区间的节点识别参数以及预设的节点识别标准获取音频中每个音节起始时刻、终止时刻以及底噪或静音区,并依次进行标记编号;

其中,一个音节的终止时刻至后一相邻音节的起始时刻为音节的底噪或静音区;

步骤6:依据步骤1中的文字以及步骤5中标记编号构建文字与语音中音节的时刻同步表,得到文字与语音的同步时间戳;

其中,所述同步时间戳是所述同步表中每个文字与对应音节在起始时刻、终止时刻、底噪或静音区的标记编号。

2.根据权利要求1所述的方法,其特征在于:所述预设的节点识别标准为:

A:若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB,$-1dB\le\beta_x^n\le1dB$,则表示满足μ_x^n<36dB,$-1dB\le\beta_x^n\le1dB$的所述α个区间内存在底噪或静音区的起始时刻;

B:若区间的电平参数和电平变化参数满足μ_x^n>36dB,$\beta_x^n>1dB$,且随后相邻的α-1个区间的电平变化参数均大于1dB,则表示满足μ_x^n>36dB,$\beta_x^n>1dB$的所述区间内存在一个音节的起始时刻;

C:若连续α-2个区间的电平参数和电平变化参数均满足μ_x^n>36dB,$-1dB\le\beta_x^n\le1dB$,则表示满足μ_x^n>36dB,$-1dB\le\beta_x^n\le1dB$的所述α-2个区间内存在音节的高潮处时刻;

D:若区间的电平参数和电平变化参数均满足μ_x^n>36dB,$\beta_x^n<-1dB$,且随后相邻的α-1

个区间的电平变化参数均小于-1dB，则表示音节的幅度下降；

E：若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$<36dB，$-1dB \le \beta_x^n \le 1dB$的所述α个区间内存在音节终止时刻；

其中，3≤α≤7。

3.根据权利要求2所述的方法，其特征在于：节点识别标准A中，满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$的连续α个区间中的第一个区间内存在底噪或静音区的进入时刻；

节点识别标准E中，满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$的连续α个区间中第一个区间内存在音节终止时刻。

4.根据权利要求1所述的方法，其特征在于，aac格式的音频文件的采样频率为48kHz或44.1kHz时，x的取值为90。

5.一种采用权利要求1-4任一项所述方法的剪辑方法，其特征在于：包括如下步骤：

S1：获取音频，以及将音频翻译为文字；

S2：与S1同步构建音频中文字与语音的同步时间戳；

S3：在剪辑窗口剪辑文字或者音频，再依据所述同步时间戳剪辑对应音频或文字；

a：在剪辑窗口内剪辑文字，并获取剪辑操作类型以及所选定的文字，并依据同步时间戳获取所选定的文字在音频中的标记编号，再依据获取的标记编号对音频进行相应的剪辑操作；

b：在剪辑窗口内剪辑音频，并获取剪辑操作类型以及音频剪辑范围，并依据所述同步时间戳获取音频剪辑范围对应的标记编号，再依据获取的标记编号对文字进行相应的剪辑操作；

S4：将剪辑完的文字与音频进行封装导出。

6.根据权利要求5所述的方法，其特征在于：S3中依据文件剪辑操作来剪辑音频时，每个文字对应的音频剪辑范围为：与文字对应的音节起始时刻至底噪或静音区的中间位置。

7.一种采用权利要求5所述方法的剪辑系统，其特征在于：包括音频文字转换模块、同步时间戳构建模块、文字剪辑模块、音频剪辑模块、编码模块；

其中，所述同步时间戳构建模块、文字剪辑模块均与所述音频文字转换模块连接，所述音频剪辑模块与所述同步时间戳构建模块、文字剪辑模块连接，所述编码模块与所述音频剪辑模块、文字剪辑模块连接；

所述音频文字转换模块，用于将音频翻译为文字；

所述同步时间戳构建模块，用于构建音频中文字与语音的同步时间戳；

所述文字剪辑模块，用于对文字进行剪辑；

所述音频剪辑模块，用于剪辑音频；

所述编码模块，用于将剪辑完的文字与音频进行封装导出。

一种音频的语音与文字同步方法及剪辑方法和剪辑系统

技术领域

[0001] 本发明属于语音音频剪辑、编辑的技术领域，具体涉及一种音频的语音与文字同步方法及剪辑方法和剪辑系统。

背景技术

[0002] 传统广播电台、电视台的语音类音频内容编辑，主要是利用通用的音频编辑软件实现(如audition软件，samplitude软件)，此类软件本身没有同步文本编辑窗口，需要完全依靠人耳的听觉控制来进行人工剪切和修饰，效率及准确率不高。其中，一般是利用现将音频稿件生成文件稿，然后再用wps软件和audition软件的不停切换方式工作，边修改文件边编辑音频，关于音频的修改则依托人工监听，通过人脑记忆，逐字逐句的进行听写和校对，即文字的删减调整和音频的删减调整需要分两步完成，导致在重大节目播出时，音频稿和文字稿因审稿、不同渠道发布的需要，必须一一对应时，编辑的工作将消耗大量的工作时间，且工作过程繁琐而枯燥。

发明内容

[0003] 本发明的目的是提供一种音频的语音与文字同步方法及剪辑方法和剪辑系统，能够建立音频中语音与文字的同步机制，再利用同步机制实现语音与文字的同步剪辑，进而快速地完成音频和对应文稿的同步剪辑。

[0004] 第一方面，本发明提供一种音频的语音与文字同步方法，包括如下步骤：

[0005] 步骤1：解码音频文件得到连续采样的音频波形，并将音频同步翻译为文字；

[0006] 其中，所述音频波形以时间为横轴坐标、幅度为纵轴坐标，所述音频波形上分布离散采样点；

[0007] 步骤2：将步骤1中的音频波形进行区间划分；

[0008] 其中，每个区间包括x个采样点，音频波形的采样频率低于或等于48kHz，x的取值范围为50-150，音频波形的采样频率为96kHz或88.2kHz，x的取值范围为100-300，音频波形的采样频率为192kHz，x的取值范围为200-600；

[0009] 步骤3：计算每个区间的平均相对电平值；

[0010] 其中，区间的平均相对电平值计算公式如下：

[0011] $$Y_x^n(dB)=20\log\left\{\left[\left(\left|A_1^n\right|+\left|A_2^n\right|+\ldots+\left|A_x^n\right|\right)/x\right]/X\right\}$$

[0012] X＝0000,0000,0000,0001

[0013] 式中，$Y_x^n(dB)$表示第n个区间的平均相对电平值，$\left|A_1^n\right|$、$\left|A_2^n\right|$、$\left|A_x^n\right|$分别表示第n个区间中第1、2、x个采样点的幅度的绝对值，X表示量化比特数为16位的预设的取样信号，n为正整数；

[0014] 步骤4：依据每个区间的平均相对电平值获取每个区间的节点识别参数；

[0015] 其中，所述节点识别参数包括电平参数和电平变化参数：

[0016] $\mu_x^n = Y_x^n(dB)$

[0017] $\beta_x^n = Y_x^n(dB) - Y_x^{n-1}(dB)$

[0018] 式中，μ_x^n表示第n个区间的电平参数，β_x^n表示第n个区间的电平变化参数；

[0019] 步骤5：依据每个区间的节点识别参数以及预设的节点识别标准获取音频中每个音节起始时刻、终止时刻以及底噪或静音区，并依次进行标记编号；

[0020] 其中，一个音节的终止时刻至后一相邻音节的起始时刻为音节的底噪或静音区；

[0021] 步骤6：依据步骤1中的文字以及步骤5中标记编号构建文字与语音中音节的时刻同步表，得到文字与语音的同步时间戳；

[0022] 其中，所述同步时间戳是所述同步表中每个文字与对应音节在起始时刻、终止时刻、底噪或静音区的标记编号。

[0023] 本发明中预设的节点识别标准是以节点识别参数为依据而设定的，而节点识别参数中电平参数和电平变化参数其实质上是体现了区间内采样点幅度的变化以及变化快慢，进而通过研究得到一个统一标准来判定音节起始、终止等时刻，该统一标准即本发明的节点识别标准。

[0024] 本发明通过将音频进行解码，再针对进行区间划分以及计算出每个区间的节点识别参数，再确定每个音节起始时刻、终止时刻以及底噪或静音区，进而构建文字与语音中音节的时刻同步表。如下表1所示：

[0025] 表1

[0026]

序号	文字	音节起始时刻（标记编号）	音节终止时刻（标记编号）	底噪或静音区（标记编号）
1	不	T0001	T0002	T0002-T0003
2	忘	T0003	T0004	T0004-T0005
3	初	T0005	T0006	T0006-T0007
4	心	T0007	T0008	T0008-T0009
5	方	T0009	T0010	T0010-...
6	得	...	...	...

[0027] 从上述表格可知，时刻同步表包括每个文字与对应音节在起始时刻、终止时刻、底噪或静音区的标记编号，进而根据时刻同步表确定了文字与语音的对应关系，为文字与音频的同步剪辑提供了基础。

[0028] 本发明中将音频波形进行区间划分，不同音频的音频波形中每个区间的采样点所允许的范围不同，例如音频波形的采样频率低于或等于48kHz，每个区间包含50-150的采样点。其目的是可以保证，每个区间可以采集到正常人声的1个以上的谐振周期，保证了数据模型的可靠性，且同时也使谐振周期数量不会过多，而降低数据模型的精度。

[0029] 本发明基于人耳对声音的响度判断与音量幅度之间是呈对数变化的以及基于变化趋势的分析会提高识别结果的准确度，故选用对数计算出的电平参数和电平变化参数作为节点识别参数，可以更加准确地识别出音节的起始、终止时刻。

[0030] 进一步优选，所述预设的节点识别标准为：

[0031] A：若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$，

则表示满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$的所述α个区间内存在底噪或静音区的起始时刻；

[0032] B：若区间的电平参数和电平变化参数满足μ_x^n>36dB，$\beta_x^n > 1dB$，且随后相邻的α-1个区间的电平变化参数均大于1dB，则表示满足μ_x^n>36dB，$\beta_x^n > 1dB$的所述区间内存在一个音节的起始时刻；

[0033] C：若连续α-2个区间的电平参数和电平变化参数均满足μ_x^n>36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$>36dB，$-1dB \le \beta_x^n \le 1dB$的所述α-2个区间内存在音节的高潮处时刻；

[0034] D：若区间的电平参数和电平变化参数均满足μ_x^n>36dB，$\beta_x^n < -1dB$，且随后相邻的α-1个区间的电平变化参数均小于-1dB，则表示音节的幅度下降；

[0035] E：若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$<36dB，$-1dB \le \beta_x^n \le 1dB$的所述α个区间内存在音节终止时刻；

[0036] 其中，3≤α≤7。

[0037] 其中，B也表示音节的上升沿开始，D表示音节的下降沿开始。

[0038] 识别音频中各个音节的位置是从第一个区间依次往后进行判断的。基于各个区间是以时间轴依次排列的，每个区间对应在模拟音频中的时刻点是可确定的，因此每个区间的对应标记编号的时刻点是可确定的。

[0039] 基于研究发现：人耳对于声音的响度判断与音量幅度之间是呈对数变化的，音量每增加一倍，电平上增加6dB，由广电标准的静音门限-60dBFs转换为本系统的相对电平值为36dB，按16bit的量化情况(音质为CD和广播级别)，可换算为有60dB(96-36)的动态范围，也就是有10倍的音量变化。在一个单字平均发音时长里，会有一次从静音(36dB)到8至9倍(90dB左右)的音量变化过程，1dB表示的变化为：后者音量是前者的1.15倍，其为基本的一个音量开始增长的趋势，这种趋势在发音时的变化在上升沿的位置会变得更大，在本系统定义的时间戳抽样间隔下，通常是3-6dB甚至更高的前后坡度差，下降沿同样如此。

[0040] 基于上述原理可知，使用一次获得的启动门限1dB不足以判断已经进入上升周期，应该是多次计算多次判断后才能确定已经进入单音开启点，因为整体处于上升周期的某些时间戳抽样间隔内不一定会有后者比前者大很多的情况，有些部分会出现缓坡，因此不能将启动门限定义的太高。至于1dB这个门限值是不是门限太低，在需要更为锋利的时间戳启动点位置的时候，可以将门限值提高，抽取的单音节语素会更紧凑，但会导致音节之间的过度处理平滑度更差，更难处理。因此，基于上述分析，本发明将电平变化参数的参数标准设置为1dB和-1dB，同时进行多次判断才能得出节点结论，即α的取值为3-7。

[0041] 一般而言，男声发声的基础频率较低，谐振波形的时间周期较长，可调整x(步进为5)和α(步进为1)以个位数的适当增加；如果是女声，女声发声的基础频率较低，谐振波形的时间周期较短，可调整x(步进为5)和α(步进为1)以个位数的适当减少；童声接近女声，方法类似。其中，若是男女童声均采用同一α取值时，本发明将α为5作为最优的。

[0042] 进一步优选，节点识别标准A中，满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$的连续α个区间中的第一个区间内存在底噪或静音区的进入时刻；

[0043] 节点识别标准E中，满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$的连续α个区间中第一个区间内存在音节终止时刻。

[0044] 均选择第一个区间可以提高最终结果的准确率，降低选择不同区间而导致的偏差。

[0045] 进一步优选，aac格式的音频文件采样频率为48kHz或44.1kHz时，x的取值为90。

[0046] 通过研究发现，aac格式的音频文件采样频率为48kHz或44.1kHz时，若x的取值为90，每个区间的时长大概为2ms，按照一个正常人声带的发声则可以取到3个左右的谐振周期，而通过实验证明，3个左右的谐振周期其结果的准确率高，能够更加精确地确定时间戳点位。

[0047] 第二方面，本发明还提供一种采用上述方法的剪辑方法，包括如下步骤：

[0048] S1：获取音频，以及将音频翻译为文字；

[0049] S2：与S1同步构建音频中文字与语音的同步时间戳；

[0050] S3：在剪辑窗口剪辑文字或者音频，再依据所述同步时间戳剪辑对应音频或文字；

[0051] a：在剪辑窗口内剪辑文字，并获取剪辑操作类型以及所选定的文字，并依据同步时间戳获取所选定的文字在音频中的标记编号，再依据获取的标记编号对音频进行相应的剪辑操作；

[0052] b：在剪辑窗口内剪辑音频，并获取剪辑操作类型以及音频剪辑范围，并依据所述同步时间戳获取音频剪辑范围对应的标记编号，再依据获取的标记编号对文字进行相应的剪辑操作；

[0053] S4：将剪辑完的文字与音频进行封装导出。

[0054] 通过上述一种音频的语音与文字同步方法来构建音频中文字与语音的同步时间戳，来实现文字与音频的同步剪辑。

[0055] 例如，删除时，选定某一文字，根据音频中文字与语音的同步时间戳则可以确认该文字对应的音节的起始时刻、终止时刻以及前后的底噪或静音区，进而可以将该音节起始时刻至终止时刻范围的音频帧删除或者是该音节起始时刻至底噪或静音中某一时刻的范围的音频帧删除。

[0056] 其他实现方式中，在音频编辑窗口内显示已进行标记编号的音频波形，再依据广播编辑人员在音频编辑窗口，以及标记标号选择单个音节或者以音节为单位的词语或句子进行剪辑，同步依据标记标号剪辑文字稿中的文字，也实现了音频文件与文字稿的同步剪辑。

[0057] 进一步优选，S3中依据文件剪辑操作来剪辑音频时，每个文字对应的音频剪辑范围为：音节的起始时刻至底噪或静音区的中间位置。

[0058] 其中，选择剪辑音节的起始时刻至底噪或静音区的中间位置，是为了留有空间进行修整，例如删除某个音节后，留有一段静音区，可以采用幅度调整函数和淡入淡出函数进行修饰，使剪切或删除的位置过渡自然，贴近自然语义逻辑和文字语感环境。

[0059] 第三方面，本发明还提供一种采用上述剪辑方法的剪辑系统，包括音频文字转换模块、同步时间戳构建模块、文字剪辑模块、音频剪辑模块、编码模块；

[0060] 其中，所述同步时间戳构建模块、文字剪辑模块均与所述音频文字转换模块连接，所述音频剪辑模块与所述同步时间戳构建模块、文字剪辑模块连接，所述编码模块与所述音频剪辑模块、文字剪辑模块连接；

[0061] 所述音频文字转换模块，用于将音频翻译为文字；

[0062] 所述同步时间戳构建模块，用于构建音频中文字与语音的同步时间戳；

[0063] 所述文字剪辑模块，用于对文字进行剪辑；

[0064] 所述音频剪辑模块，用于剪辑音频；

[0065] 所述编码模块，用于将剪辑完的文字与音频进行封装导出。

[0066] 有益效果

[0067] 与现有预测技术相比，本发明的优点有：

[0068] 1、本发明采用上述方法，可以实现音频中的语音与文字的同步剪辑，依托文字稿直接剪辑同期声音频，将音频的时间线性剪辑转换为空间的目测直接剪辑，极大地提高了剪辑效率，且本发明采用上述方法可以建立精确地的文字与语音的同步时间戳，再依托AI翻译技术，能够实现准确剪辑，相较于人工监听，其准确率更高。

[0069] 2、由于建立了精确的时间戳，传统广播编辑人员也可以单个音节或者以音节为单位的词语、句子或段落为单位在音频区进行监听的同时直接进行剪辑，剪辑完成对应的文字稿件，也能同步导出音频和文字稿，同样极大提高剪辑的效率，方便音频专业领域里的效率提升。

[0070] 3、本发明基于人耳对声音的响度判断与音量幅度之间是呈对数变化的以及基于变化趋势的分析会提高识别结果的准确度，故选用对数计算出的电平参数和电平变化参数作为节点识别参数，可以更加准确而快速地识别出音节的起始、终止时刻，进而建立精准的文字与语音的同步时间戳，时间戳的精度为ms级，为同步剪辑提供基础。

[0071] 4、本发明将电平变化参数的参数标准设置为1dB和-1dB，是充分考虑门限定义的高低要求而得出的，其使得得出的节点结论更加准确。

附图说明

[0072] 图1是本发明提供的一种音频的语音与文字同步方法的流程示意图；

[0073] 图2是本发明提供的音频波形形离散采样点的示意图；

[0074] 图3是本发明提供的音频波形的波形图，其中(a)图显示为一个汉子音节对应的音频波形，(b)图为(a)图中间白色部分的放大图；

[0075] 图4是本发明提供的音频波形上的节点标记的示意图；

[0076] 图5是本发明提供的语音与文字同步剪辑示意图。

具体实施方式

[0077] 下面将结合实施例对本发明做进一步的说明。

[0078] 本发明提供的一种音频的语音与文字同步方法及剪辑方法和剪辑系统，主要是应用于音频剪辑，其中音频文件不限制于aac，mp3，s48(mp2)，wav等格式，据统计，使用aac和mp3格式的情况占所有采访的90％，使用wav格式的情况大约为5％，其余5％为其它音频格式。例如记者外出采样一般使用采访机进行访问，利用sony，tescam，infomedia等专业音频采访机采集音频素材回新闻中心，采访完成后，在采访机里会形成一个音频文件，它的后缀名可能是.aac、.mp3或者.wav等。记者会把这个音频文件拷贝到编辑电脑，在PC操作系统环境下，将音频文件导入本发明的剪辑系统，进行音频解析、翻译和时间戳建立以及剪辑的工作。

[0079] 其中，利用AI技术实现音频翻译，将采用AI技术将音频同步翻译为文字并显示在文字窗口中。

[0080] 其中，一般在在专业广播领域，音频采样频率采用48kHz，CD的采样频率为44.1kHz。

[0081] 采样频率为48kHz时：

[0082] 当前AAC一帧的播放时间是＝1024*(1000000/48000)＝21.33ms(单位为ms)。

[0083] 采样频率为44.1kHz时：

[0084] 当前AAC一帧的播放时间是＝1024*(1000000/44100)＝22.32ms(单位为ms)。

[0085] 而以汉语为母语的人平均一秒钟可以说3到5个字。研究人员计算出汉语为母语的人的平均说话速率为每秒5.18个音节，即单个字的发音平均需时间为：1000/5.18ms＝193.05ms。

[0086] 即在48kHz采样频率下，AAC单字发音需要平均帧数为：

[0087] 193.05/21.33帧＝9.05帧。

[0088] 在44.1kHz采样频率下，AAC单字发音需要平均帧数为：

[0089] 193.05/22.32帧＝8.65帧。

[0090] 而针对mps格式的音频文件，在44.1kHz采样频率下，每帧播放时间固定为26ms，单字发音需要平均mp3帧数为：

[0091] 193.05/26.122帧＝7.39帧。

[0092] 在48kHz采样频率下，每帧时长为24.00ms，需要平均mp3帧数为：

[0093] 193.05/24.00帧＝8.04帧。

[0094] 由上可知，对于平均单个汉字发音而言，无论是aac格式，还是mp3格式，还是ac3格式，还是无论是9帧或是6帧，完全按照伴音帧的时间度量，还是无法统一个标准精确的对单个音节进行时间打点操作。因此在建立文字和伴音的时间同步机制时，需在更底层的界面上进行精细化分析。故本发明采用以下方法进行精细化分析来确定单个音节的起始、终止时刻，进而构建与文字的同步机制。

[0095] 其中，如图1所示，本发明提供的一种音频的语音与文字同步方法，即语音与文字的时间戳构建过程，包括如下步骤：

[0096] 步骤1：解码音频文件得到连续采样的音频波形；

[0097] 其中，所述音频波形以时间为横轴坐标、幅度为纵轴坐标，所述音频波形上分布离散采样点。如图2所示为音频波形形离散采样点的示意图。

[0098] 音频解析的首要目的在于将采访的音频内容通过算法进行区隔和打点，即首先是对音频格式进行解码，还原到连续采样的标准状态，本实施例中aac音频文件的解码采用aac解码插件完成。

[0099] 步骤2：将步骤1中的音频波形进行区间划分；

[0100] 其中，每个区间包括x个采样点，音频波形的采样频率低于或等于48kHz，x的取值范围为50-150，音频波形的采样频率为96kHz或88.2kHz，x的取值范围为100-300，音频波形的采样频率为192kHz，x的取值范围为200-600。一般优选x取值为对应范围的中间点。

[0101] 本实施例中选用aac格式的音频文件且采样频率为48kHz或44.1kHz时，x的取值为90。如图3中的(a)图显示为一个汉子音节对应的音频波形，其中该音节的时长约为200ms，

(b)图为(a)图中间白色部分的放大图，其为汉子音节开启部分的上升沿位置，音频波形的波形图上90个点构成一个区间T_x^n，区间T_x^n中有近3至4次的波形周期起伏。

[0102] 本实施例中优选x为90是基于采样频率为48kHz时，aac单字发音需要平均帧数为9.05帧，采样频率为44.1kHz时，aac单字发音需要平均帧数为8.65帧。例如采样频率为48kHz时，汉语单个字的发音平均采样点数为：

[0103] 1024*9.05＝9267.2

[0104] 如果每个区间包括90个采样点，则是将单个字的发音平均采样点数分为约100分，每区间对应的时间约为2ms，按照一个正常人声带的发声则可以取到3个左右的谐振周期，而通过实验证明，3个左右的谐振周期其结果的准确率高，能够更加精确地确定时间戳点位。

[0105] 采样频率为44.1kHz时，将单个字的发音平均采样点数分为约100分，每区间对应的时间也约为2ms，采样点数为88个，近似90，在此不再赘述。

[0106] 其他可行的实施例中，音频波形的采样频率低于或等于48kHz，x取值可以是50-150中的任意一个，或者是音频波形的采样频率为96kHz或88.2kHz，x的取值范围为100-300，音频波形的采样频率为192kHz，x的取值范围为200-600，这是基于为了更加精准地确定时间戳，则每个区间至少能够取到一个完整的谐振周期且要高于一个谐振周期，因为考虑到每个人发音频率有差异，需要留有一定的容错空间，因此必须高于一个谐振周期，例如每个区间多30个采样点来缓冲，但是每个区间提取的谐振周期也不能过多，过多则将造成数据选取不够精细，降低预测结果的可靠性。因此，本发明将音频波形的采样频率低于或等于48kHz，x取值选为50-150的范围。按48kHz和44.1kHz的广播专业级采样频率，x的取值为50时，能提取到6个谐振周期，x的取值为150时，能提取到1.5个谐振周期。

[0107] 步骤3：计算步骤2中每个区间的平均相对电平值；

[0108] 其中，区间的平均相对电平值计算公式如下：

[0109] $$Y_x^n(dB)=20\log\left\{\left[\left(\left|A_1^n\right|+\left|A_2^n\right|+...+\left|A_x^n\right|\right)/x\right]/X\right\}$$

[0110] X＝0000,0000,0000,0001

[0111] 式中，$Y_x^n(dB)$表示第n个区间的平均相对电平值，$\left|A_1^n\right|$、$\left|A_2^n\right|$、$\left|A_x^n\right|$分别表示第n个区间中第1、2、x个采样点的幅度的绝对值，X表示量化比特数为16位的预设的取样信号，n为正整数。

[0112] 其中，上述实质为将模拟信号转换成数字信号，把输入信号量化成一个个离散的数据序列其中，每个区间对应一个平均相对电平值$Y_x^n(dB)$。

[0113] 步骤4：依据步骤3计算出的每个区间的平均相对电平值获取每个区间的节点识别参数；

[0114] 其中，所述节点识别参数包括电平参数和电平变化参数：

[0115] $$\mu_x^n=Y_x^n(dB)$$

[0116] $$\beta_x^n=Y_x^n(dB)-Y_x^{n-1}(dB)$$

[0117] 式中，μ_x^n表示第n个区间的电平参数，β_x^n表示第n个区间的电平变化参数。以此类

推，记录n取值后的μ_x^n和β_x^n的状态表。其中β_x^1为0。

[0118] 步骤5：依据步骤4计算出的每个区间的节点识别参数以及预设的节点识别标准获取音频中每个音节起始时刻、终止时刻以及底噪或静音区，并依次进行标记编号。

[0119] 其中，一个音节的终止时刻至后一相邻音节的起始时刻为音节的底噪或静音区。由此可知，可以确定音节对应的时间节点，即音节的起始时刻、终止时刻以及底噪或静音区。

[0120] 其中，预设的节点识别标准如下：

[0121] A：若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$<36dB，$-1dB \le \beta_x^n \le 1dB$的所述α个区间内存在底噪或静音区的起始时刻；

[0122] B：若区间的电平参数和电平变化参数满足μ_x^n>36dB，$\beta_x^n > 1dB$，且随后相邻的α-1个区间的电平变化参数均大于1dB，则表示满足μ_x^n>36dB，$\beta_x^n > 1dB$的所述区间内存在一个音节的起始时刻；

[0123] C：若连续α-2个区间的电平参数和电平变化参数均满足μ_x^n>36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$>36dB，$-1dB \le \beta_x^n \le 1dB$的所述α-2个区间内存在音节的高潮处时刻；

[0124] D：若区间的电平参数和电平变化参数均满足μ_x^n>36dB，$\beta_x^n < -1dB$，且随后相邻的α-1个区间的电平变化参数均小于-1dB，则表示音节的幅度下降；

[0125] E：若连续α个区间的电平参数和电平变化参数均满足μ_x^n<36dB，$-1dB \le \beta_x^n \le 1dB$，则表示满足$\mu_x^n$<36dB，$-1dB \le \beta_x^n \le 1dB$的所述α个区间内存在音节终止时刻；

[0126] 其中，3≤α≤7。本实施例中将选择α为5。

[0127] 步骤6：依据步骤1中的文字以及步骤5中标记编号构建文字与语音中音节的时刻同步表，得到文字与语音的同步时间戳。

[0128] 其中，所述同步时间戳是所述同步表中每个文字与对应音节在起始时刻、终止时刻、底噪或静音区的标记编号。其中，依据每个单音节所对应的起始、结束时刻，可以确定每个单音节在语音中的位置，进而与文字可以构成时刻同步表。如上表1所示，从表格中可知，一个音节的终止时刻至后一相邻音节的起始时刻为音节的底噪或静音区，例如表1中“不忘初心方得”中的文字“不”对应在音节的起始时刻、终止时刻的标记编号为T0001、T0002，而“不”相邻的字“忘”应在音节的起始时刻、终止时刻的标记编号为T0003、T0004，即相邻的“不”、“忘”之间的底噪或静音区为T0002-T0003。

[0129] 应当理解，时刻同步表的精度取决每个区间所选定的x大小，因此，本发明基于对音频文件的分析将x取值为50～150以提高时刻同步表的精度。如图4所示，在音频波形上可以标记出音节的节点标记，如：标记01、标记02、标记03、标记04等位置。其中，标记01至标记03为第一个音节，标记02至标记03之间为底噪声区，标记02至标记04为第二个音节。得到的时间戳即为：表1中文字中“不”与“起始时刻”(T0001对应标记01)，“终止时刻”(T0002对应标记03)，“底噪或静音区”(T0002-T0003对应标记03-标记02)。

[0130] 基于上述方法，本发明还提供一种音频的剪辑方法，包括如下步骤：

[0131] S1：获取音频，以及采用AI技术将音频翻译为文字并显示在文字窗口。

[0132] S2：与S1同步构建音频中文字与语音的同步时间戳。

[0133] 其中，将音频翻译为文字的同时，同步进行同步时间戳构建，即得到上述步骤6中文字与语音中音节的时刻同步表。

[0134] S3：在剪辑窗口剪辑文字或者音频，再依据所述同步时间戳剪辑对应音频或文字；其中，剪辑操作类型包括剪切、复制、粘贴和删除。

[0135] a：在剪辑窗口内剪辑文字，并获取剪辑操作类型以及所选定的文字，并依据同步时间戳获取所选定的文字在音频中的标记编号，再依据获取的标记编号对音频进行相应的剪辑操作；

[0136] b：在剪辑窗口内剪辑音频，并获取剪辑操作类型以及音频剪辑范围，并依据所述同步时间戳获取音频剪辑范围对应的标记编号，再依据获取的标记编号对文字进行相应的剪辑操作；

[0137] 本实施例中是以在剪辑窗口内剪辑文字，再自动剪辑音频为例进行说明。

[0138] 如图5所示，在文字窗口双击文字部分，可按人工智能识别的基本语义选中单句或词组，同步选中对应的音频。如需改变选择范围，再用鼠标扩大和缩小文字范围，同步关联到相应的音频，音频的切口和延时的大小由人工智能算法判断。选择完成之后，再统一在文字稿上进行剪辑或编辑，这样更进一步，提升音频编辑的工作效能。

[0139] 本实施例中优选，剪辑音频时，每个文字对应的音频剪辑范围为：音节的起始时刻至底噪或静音区的中间位置。即留有底噪或静音区，其目的是为了防止完全以音节的起始或终止时刻进行剪辑时导致接口过度不自然的情况，即接口过渡过紧或过松，底噪不一致而造成编辑后的音频听感不自然的问题。本发明优选留有一段底噪或静音区，是为了引用幅度调整函数和淡入淡出函数对语音边界进行适当修整。

[0140] 其中所采用的幅度调整函数和淡入淡出函数在剪切和删除的位置，使其过渡自然，需尽量符合人的自然语义逻辑和文字语感环境。例如，将后面讲过的一句话插入前面的某一个自然段里，则需采样自然段里的音节之间的过渡时间和音节的幅度，然后计算该自然段的平均过渡时间和平均幅度，再计算目标句子的平均过渡时间和平均幅度，得到两者的比值，最后对插入的句子进行幅度调整和过渡间隙的淡入淡出调整。并且需注意到，尽量采用自然段中的底噪片段进行修饰，以适合语境，同时也不能完全抛弃原句子的底噪片段，在音节过渡的位置，需要保留部分原位置的底噪，采用淡入淡出函数进行卷积操作，以尽量实现过渡的平滑。

[0141] 其他可行的实施例中，每个文字对应的音频剪辑范围为：音节的起始时刻至底噪或静音区的任意位置或者音节的起始时刻至同一音节的终止时刻。

[0142] 应当理解，基于上述原理，可以对音频进行剪切、复制、粘贴和删除，剪切即在音频文件中将文字对应的音节剪辑范围剪切掉，复制即在音频文件中将选择文字对应音节剪辑范围，粘贴即在音频文件中将待粘贴的音频帧插入到音频文中的对应位置，删除即在音频文件中将文字对应的音节剪辑范围删除。

[0143] S4：将剪辑完的文字与音频进行封装导出。

[0144] 在音频编辑工作完成以后，文字和音频文件就是目标文件，它们之间是严格同步的，需要将其导出。文字可直接导出为word文件或txt文件，音频可导出为.aac，.mp3，.wav等格式，如果是.aac等有损压缩文件，则必须根据国际标准对其进行重新编码，再恢复成音频帧的封装格式，重新导出。

[0145] 基于采用上述剪辑方法，本发明还提供一种剪辑系统，包括音频文字转换模块、同步时间戳构建模块、文字剪辑模块、音频剪辑模块、编码模块；

[0146] 其中，所述同步时间戳构建模块、文字剪辑模块均与所述音频文字转换模块连接，所述音频剪辑模块与所述同步时间戳构建模块、文字剪辑模块连接，所述编码模块与所述音频剪辑模块、文字剪辑模块连接；

[0147] 所述音频文字转换模块，用于采用AI技术将音频翻译为文字并显示；

[0148] 所述同步时间戳构建模块，用于构建音频中文字与语音的同步时间戳；

[0149] 所述文字剪辑模块，用于依据文字窗口内的剪辑操作对文字进行剪辑或者基于文字与语音的同步时间戳随音频剪辑操作而同步剪辑文字；

[0150] 所述音频剪辑模块，用于基于文字与语音的同步时间戳随文字剪辑操作而同步剪辑音频或者是依托在音频剪辑窗口内的剪辑操作而剪辑音频。

[0151] 所述编码模块，用于将剪辑完的文字与音频进行封装导出。

[0152] 其中，音频文字转换模块、同步时间戳构建模块、文字剪辑模块、音频剪辑模块的具体实施过程请参照上述方法的描述。为了提高系统的剪辑效率，本发明优选音频文字转换模块首先将声音转文字。其次在将声音转文字过程中，同步时间戳构建模块同步建立文字和音频的同步机制，也就是“时间戳”，其中，时间戳的建立可以不要显示在文字窗口上。

[0153] “时间戳”建立的核心点在于，首先要建立音频和文本的严格同步机制。只有首先做到精准的时间同步，才有可能在完成文字剪辑时，同步处理音频内容。它涉及文字对应的语音起始和截止的具体时间点位、音频剪切的点位和剪切之后前后语音过渡能否自然，是否需要加入适当的静音时间和淡入淡出的语音处理技巧，同时直接关系到后续截取内容时的精确程度(以ms为单位)。

[0154] 通过本发明所述方法，可以极大地提高新闻工作者的剪辑效率，例如针对新闻稿中受访者会说很多，有时候会跑题，有时候会讲错一句或者几句话等诸如此类的情况，传统剪辑方式为，音频编辑坐在电脑前，带上耳机，在音频编辑软件里逐字逐句的听，然后剪掉冗余和出错的部分，达到基本的时间限制要求。然后对剪辑过的关键点进行降噪、延时和淡入淡出处理。最后再听，直到耳朵听不出明显的错误之后再送给编审去审稿。而使用本发明所述方法与系统，音频编辑不再需要逐字逐句的进行音频监听，而是通过机器将采访内容的声音转成文字，直接在文稿编辑窗里进行文字剪辑，再通过文稿时间戳子系统在文字窗口和对应的音频窗口位同时打上剪辑标记，在剪辑点位以及语义的转换点，自动进行一遍降噪、延时和淡入淡出等语音信号处理技术处理。减轻了编辑人员在语音翻译和信号处理上的工作量，让编辑能够集中精力处理关键点位，提高音频作品的美感和可听性。

[0155] 需要强调的是，本发明所述的实例是说明性的，而不是限定性的，因此本发明不限于具体实施方式中所述的实例，凡是由本领域技术人员根据本发明的技术方案得出的其他实施方式，不脱离本发明宗旨和范围的，不论是修改还是替换，同样属于本发明的保护范围。

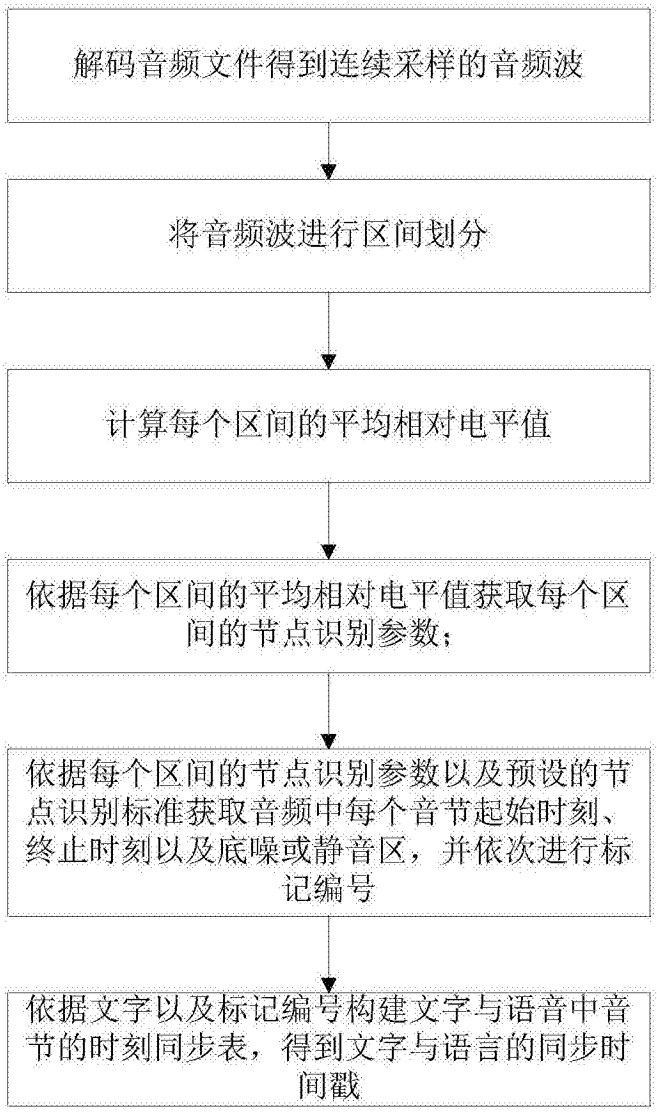

解码音频文件得到连续采样的音频波
将音频波进行区间划分
计算每个区间的平均相对电平值
依据每个区间的平均相对电平值获取每个区间的节点识别参数；
依据每个区间的节点识别参数以及预设的节点识别标准获取音频中每个音节起始时刻、终止时刻以及底噪或静音区，并依次进行标记编号
依据文字以及标记编号构建文字与语音中音节的时刻同步表，得到文字与语言的同步时间戳

图1

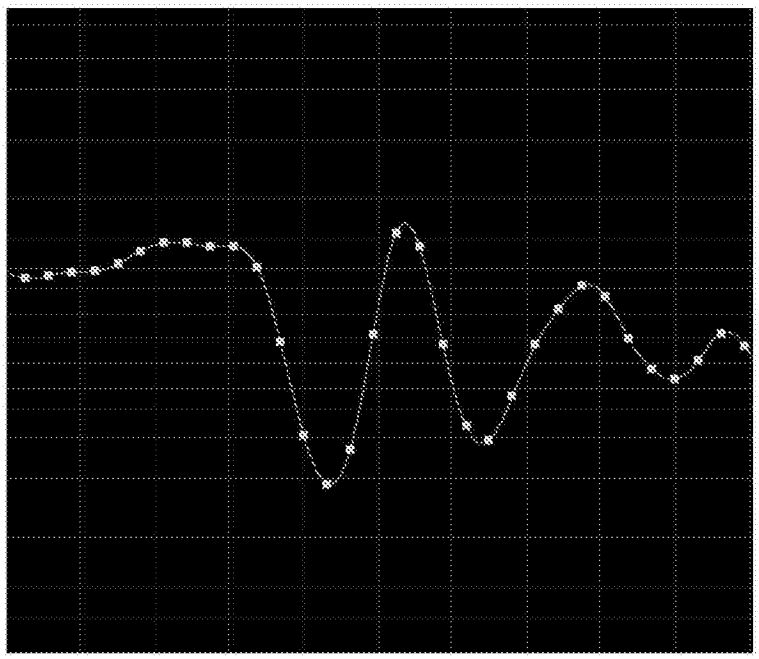
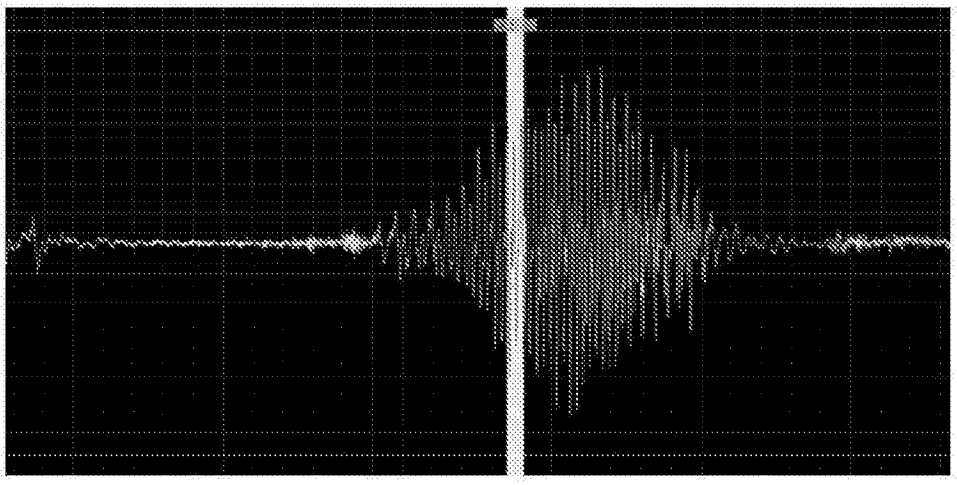
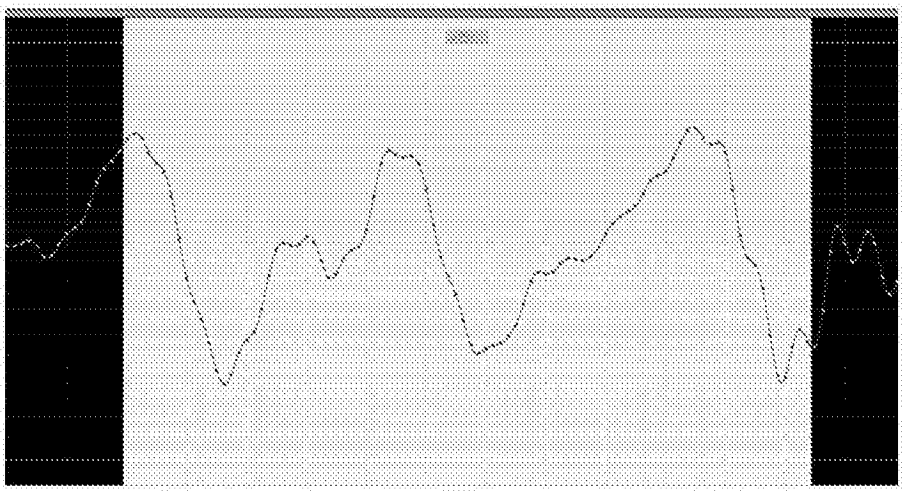


图2



(a)



(b)

图3

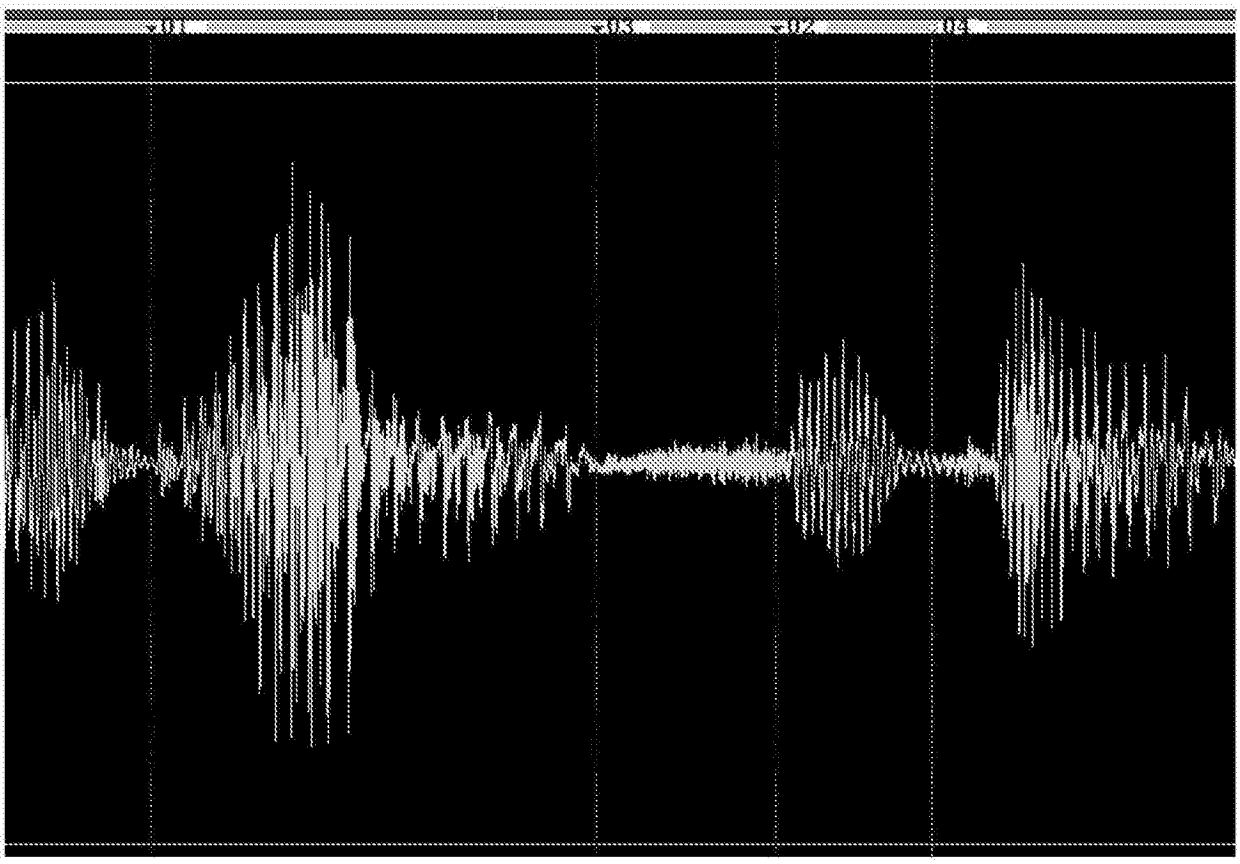


图4

图5