

# 语音时长调整快速算法

莫双燕, 关海欣, 郑可欣

(法源司法科学证据鉴定中心声像部, 北京 100043)

**摘要:** 针对目前使用的时域语音时长调整算法计算量大、难以保证实时性的问题, 深入研究语音时长调整算法的原理, 并结合语音自身的短时平稳性、准周期特性和信号的频率特性, 提出三种解决途径(只搜索叠加部分原则、隔点搜索原则、隔点计算相似性原则), 在保证语音质量不降低的同时, 大幅减少冗余的计算量, 实验结果表明, 该方法调整后的语音质量高、计算速度快, 通过与原始算法的对比证明了以上结论, 该方法能广泛应用于实际中, 尤其是应用于实时性要求较高的场合。

**关键词:** 语音; 时长调整; 快速算法

中图分类号: TN912.3

文献标识码: A

文章编号: 1000-3630(2010)-05-0507-05

DOI 编码: 10.3969/j.issn1000-3630.2010.05.012

## Fast speech time-scale modification algorithm

MO Shuan-yan, GUAN Hai-xin, ZHENG Ke-xin

(Speech and Image Department of FAYUAN Forensic Science Center, Beijing 100043)

**Abstract:** Three approaches to solving the problem of the large computation amounts of the time-scale modification algorithm have been proposed through deeply studying the principle of the algorithm and combining with the short time stationary, quasi-periodic and frequency characteristics of the speech signal. The new proposed fast algorithm could reduce the computation significantly, and meantime maintain the quality of speech. The experimental results show that the modified speech could maintain the quality, and the new algorithm increases the speed of computation and can be widely used in practice.

**Key words:** speech; time-scale modification; fast algorithm; speed and quality

## 1 引言

语音时长调整, 就是要在不改变语音的音调并保证良好音质的情况下, 使语音在时间轴上被压缩或者拉伸, 即通常所说的变速不变调。该技术可广泛应用于语言学习、司法听辨、电影制作等领域。

近些年针对语音时长调整技术的研究主要集中在两个方面, 一是调整后的语音质量问题, 二是语音调整算法的计算量问题。

语音时长调整算法, 可分为时域方法和频域方法, 时域方法以重叠区波形相似性(WSOLA)<sup>[1]</sup>算法为代表, 对于语音应用来说可以得到较高的语音质量, 且相对于频域算法计算量略小, 而对于音乐数据, 由于其频谱变化剧烈, 时域方法通常难以获得较高语音质量, 此时通常采用计算量较大的频域方法, 如子带 WSOLA 算法<sup>[2,3]</sup>。国内相关领域学者近些年针对计算量和语音质量问题做了相应研究, 如

针对提高语音质量的方法<sup>[4]</sup>和减少算法计算量的方法<sup>[5-8]</sup>, 其中文献[4]将信号分为瞬态成分、稳态成分、静音成分, 再通过不同的比例系数进行压扩, 从而在总语音长度上满足时长调整要求, 但其并不是真正意义上将语音信号按照比例严格进行压扩, 所以其语音部分达不到比例要求, 从理论上来说并不满足语音时长调整的要求。文献[5, 6]采用了诸如过零率匹配等方法, 这两种算法的优点是减少了互相关系数计算带来的计算量, 提高了算法的速度, 但其准确度不如使用互相关系数的方法, 从而导致调整后的语音质量不高。文献[7]采用的快速搜索算法, 本质上是一种多层二分法搜索, 理论上可以提高搜索速度, 减少计算量, 然而其并未考虑语音自身的一些固有特性, 搜索速度的提高有限。

本文在深入研究 WSOLA 算法的基础上, 结合语音信号自身的周期性和信号的频率特性, 先采用按周期的搜索算法, 粗略查找最小值区域的位置, 然后再搜索具体位置, 并且根据语音信号基频明显低于采样率的事实使用隔点计算, 大幅降低了算法计算量, 且同时保证了语音信号的质量。本文介绍了 WSOLA 算法的基本原理及快速算法的理论依据

收稿日期: 2009-11-10; 修回日期: 2010-02-04

基金项目: 广西自然科学基金资助项目(0639028)

作者简介: 莫双燕(1985-), 女, 研究方向为声像信息处理。

通讯作者: 关海欣, E-mail: lantian\_guan@yahoo.com.cn

和实现方法, 并给出提出的快速算法与原始算法在计算量和语音质量方面的比较结果。

## 2 WSOLA 算法原理

WSOLA 算法采用的是分解合成的思想, 将原始语音信号以帧间距  $L$ , 帧长  $N$  进行分帧, 以帧间距  $\alpha L$  进行合成。其中  $\alpha$  为时长调整因子 (若  $\alpha > 1$ , 表示语音被压缩, 反之则被拉伸)。为克服各相邻帧合成时出现频谱断裂和相位不连续, 合成时在原始语音信号的  $\tau(L_k)$  ( $L_k = kL, \tau(L_k) = \alpha L_k$ ) 采样点处的邻域  $[-\Delta\max, \Delta\max]$  内移动, 以寻找与分解后的第  $k$  帧信号波形最大相关的波形, 确定合成帧的起始位置。其算法示意图如图 1 所示。

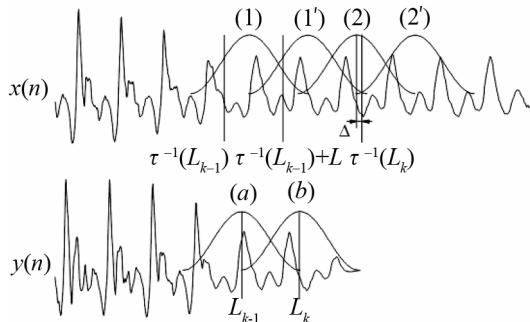


图 1 WSOLA 算法示意图  
Fig.1 Illustration of WSOLA algorithm

假设(1)帧是从原始语音信号中提取的前一帧, 把(1)帧叠加到输出信号的  $L_{k-1}$  处, 这样(a)帧=(1)帧。WSOLA 算法需要从原始语音信号  $\tau(L_k)$  处的邻域范围内找到(b)帧, 与(a)帧叠加形成自然连续, 那么 WSOLA 算法要在原始语音信号的  $\tau(L_k)$  处的最大公差范围  $[-\Delta\max, \Delta\max]$  内, 找到一帧尽可能相似(1')帧。根据(1')帧与原始语音信号帧之间的相似性计算公式, 互相关系数或短时平均幅度差 (AMDF) 系数, 可找到最佳帧(2)帧与(1')帧具有最大相似性, 这样(b)帧=(2)帧, 与(a)帧在输出信号中叠加, 形成自然连续。继续处理下一帧, 这时(2')帧代替了(1')帧的作用。

本文中重叠相加使用的窗函数为汉宁窗, 利用短时平均幅度差系数计算相似性, 如公式(1)所示:

$$c_A(k, \delta) = \sum_{n=0}^{N-1} |x(n + \tau(L_{k-1}) + \Delta_{k-1} + L) - x(n + \tau(L_k) + \delta)| \quad (1)$$

其中,  $\Delta_{k-1}$  为上次搜索的最佳相似性位置偏差,  $\delta$  为此次搜索要寻找的最佳偏差。

## 3 快速 WSOLA 算法

此节通过分析语音信号的特点, 确定 WSOLA

算法的参数, 再以此为基础, 分析如何降低 WSOLA 算法的计算量, 提出了三种途径以减少计算量。

### 3.1 WSOLA 算法参数选择

WSOLA 算法中需要确定的参数包括帧长  $N$ , 帧间距  $L$ , 最大搜索偏差  $\Delta\max$ 。

由于语音信号的短时平稳性在 10~30ms 之间, 本文选择帧长为  $T_N = 20\text{ms}$ 。对于男性发音, 其基音周期主要集中在 9ms 左右, 而对于女性发音, 其基音周期主要集中在 5ms 左右。帧间距应该至少包含语音信号的一个基本周期, 才能有效防止合成时基频断裂发生, 所以取帧间距  $T_L = 10\text{ms}$ , 同理, 搜索范围  $[-\Delta\max, \Delta\max]$  内也应该至少包含一个基音周期, 所以,  $\Delta\max$  设为 5ms 较为合理。

从 WSOLA 算法流程不难看出, 计算量大的根源在于相似性的计算, 如果能更加快速地计算相似性, 加快搜索算法, 则算法总体计算量必然会迅速下降。

### 3.2 只搜索叠加部分原则

只搜索叠加部分原则, 即在计算相似性的时候, 仅计算叠加部分的相似性, 对后半部分信号不予考虑, 依据如下。

WSOLA 采用的是重叠相加法, 重叠部分长度为  $N/2$ , 本文中  $N = T_N \times fs$ , 重叠部分为 10ms, 其中  $fs$  为语音信号采样率, 无论对于男性还是女性语音信号, 均涵盖一个周期以上的语音信号, 不会造成基频断裂发生, 图 2 是一帧长为 20ms 的男性发音信号, 含有 2 个基音周期。并且, 针对一帧信号的叠加, 后半部分并不参与叠加, 也不会造成基频断裂发生, 所以该部分语音信号若参与计算不仅不会对提高语音质量起到作用, 反而使计算量增加一倍, 因此公式(1)中求和上限改为  $N/2 - 1$ , 相应的计算量减少一半。

针对于个别基音频率很低的人, 也可以适当增加公式(1)中的求和上限, 此处定义求和上限为  $T_{\text{sum}}$

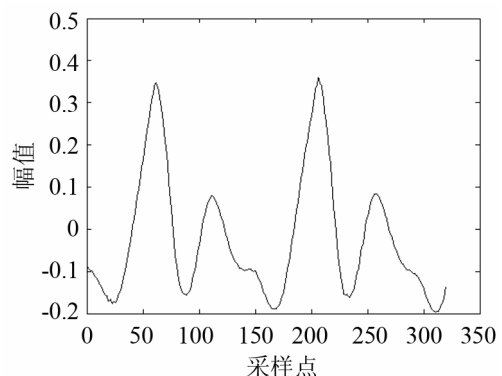


图 2 长为 20ms 的男性发音波形  
Fig.2 The 20ms male voice waveform

(其中,  $N/2-1 \leq T_{sum} \leq N-1$ ), 本文中  $T_{sum}$  取  $N/2-1$ 。

### 3.3 隔点搜索原则

隔点搜索原则, 即在最大公差范围  $[-\Delta_{max}, \Delta_{max}]$  内, 寻找最佳相似帧的过程中, 并不是每一点都搜索, 而是先采用粗略的隔点搜索, 然后在某一范围内进行精细搜索, 以达到快速搜索的目的。

由于语音信号的短时平稳和准周期的特性, 相关性函数应为连续函数, 且变化平缓, 以此为基础可知并不需要一次搜索所有点, 可先隔点搜索出大约的相似性最大位置, 然后再在小范围内继续搜索最佳匹配位置。本文将粗略搜索间隔定义为  $T_s$ , 精细搜索则在第一次粗略搜索的最佳位置邻域范围, 即在  $[-T_s/2, T_s/2]$  范围内做精细搜索, 本文中  $T_s$  取  $0.5ms$ 。图 3 为一段语音某一帧全部采样点均搜索的结果, 可见, 相似性函数存在最小值, 且变化平缓, 最佳相似性在采样点 32 位置。而使用隔点粗略搜索的结果, 如图 4 所示, 先搜索到粗略位置, 经折算后也在 32 位置, 为了保证其准确性再次使用精细搜索, 见图 5, 最终可确定相似性最大的位置。算法思想如图 6 所示。

第一次搜索, 每隔  $T_s=0.5ms$  计算一次相似性, 得到相似性最大点 A, 然后在 A 邻域内精细搜索, 搜索所有采样点位置, 得到最佳位置 B, 搜索结束。

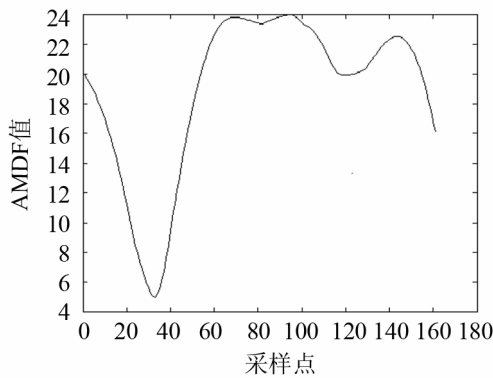


图 3 全部采样点搜索结果  
Fig.3 The search results of all data

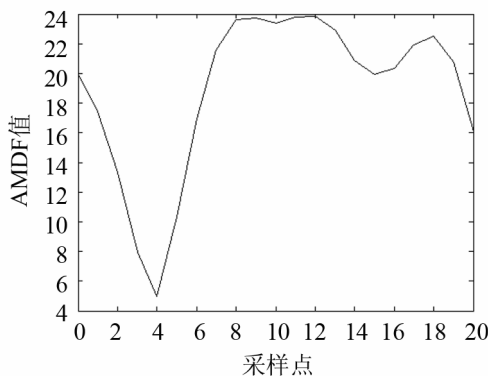


图 4 粗略搜索结果  
Fig.4 The rough search results

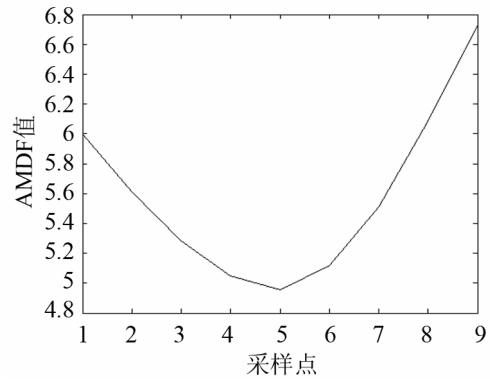


图 5 精细搜索结果  
Fig.5 The detailed search results

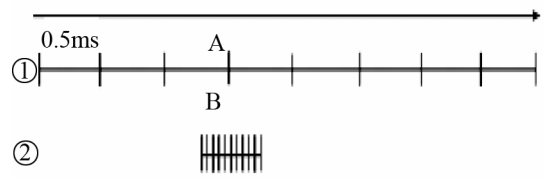


图 6 搜索过程示意图  
Fig.6 The diagram of search process

### 3.4 隔点计算相似性原则

隔点计算相似性原则, 是指在计算相似性的时候, 并不是所有采样点均参与计算, 只有其中一部分参与计算, 这样能够减少大量的计算。

由于语音信号的基音频率远小于语音信号的采样率, 在这种情况下, 并不需要所有采样点均参与计算也应当可以准确地计算出相似性, 出于这种考虑, 将公式(1)中的  $n$  隔点取出参与计算, 隔点间隔定义为  $T_c$ , 本文中  $T_c$  取  $0.5ms$ 。按此方法, 并结合隔点搜索方法, 得到的粗略搜索和精细搜索结果分别如图 7、8 所示, 对比图 4、5 可看出, 这种方法的确在减少计算量的同时, 保证了相关性计算的准确性。

## 4 快速算法与原始算法性能比较

### 4.1 计算复杂度比较

按第 3 节参数设置, 则  $N=T_N \times fs = fs \times 20/1000$ ,

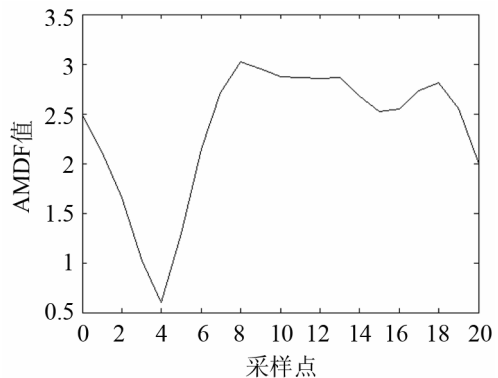


图 7 粗略搜索结果  
Fig.7 The rough search results

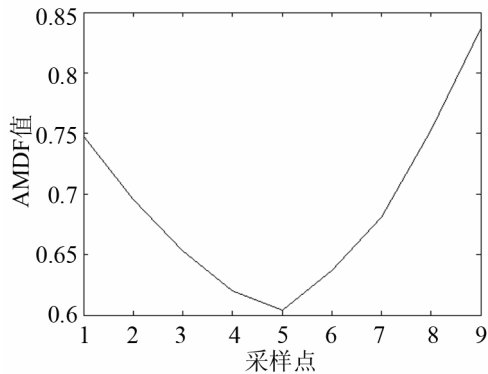


图8 精细搜索结果  
Fig.8 The detailed search results

按照原始算法搜索一帧需要计算  $N$  个相关系数，每个相关系数需要  $N$  次加法，所需的计算量为  $N^2=(fs \times 20/1000)^2$ ，若按本文快速算法，搜索一帧需要计算  $T_{sum}/(T_s \times fs)=20$  次相关系数，计算每个相关系数需要  $2\Delta_{max}/T_c+T_c \times fs=20+fs/2000$  次加法，所需计算量为  $20 \times (20+fs/2000)$ 。通常采样率越高，计算量越大，对于高采样率的语音信号，原始算法在实时性上很难保证，而本文快速算法的优势很明显，计算量增加缓慢，表 1 是三种采样率情况下原始算法与本文算法搜索一帧所需计算量的对比。

表 1 不同采样率下两种算法计算量比较

Table 1 Comparison between two algorithms' computation amounts under different sampling rates

| 采样率     | 8000Hz | 16000Hz | 44100Hz |
|---------|--------|---------|---------|
| 原始算法计算量 | 25600  | 102400  | 777924  |
| 快速算法计算量 | 480    | 560     | 841     |

很明显，从表 1 数据可以看出本文快速算法在计算量上有着明显的优势，可以实现实时快速的语音变速调整。

#### 4.2 语音质量比较

在语音质量上，原始算法和本文的快速算法差异微小，在多种语音时长调节因子  $\alpha$  下，两种算法调整后的语音经多人试听，效果都很令人满意，均感无差异。为了更客观的对比两种算法，采用了 BARK 谱失真测度评价方法 BSD<sup>[9]</sup>，该方法是一种能反应人耳听觉特征的语音质量评价方法，与人耳的主观听感有极高的相关性，可以很好地反应语音质量。该评价方法的计算公式如式(2)所示：

$$r_{NSR} = 10 \times \lg \left\{ \frac{1}{M} \sum_{i=0}^{M-1} \sum_{b=1}^{b=S} \frac{\sum_{j=bl}^{bh} |X_{ij}|^2 - |Y_{ij}|^2}{\sum_{j=bl}^{bh} |X_{ij}|^2} \right\} \quad (2)$$

式中  $X_{ij}$  是原始语音第  $i$  帧第  $j$  条谱线， $Y_{ij}$  是调

整后语音第  $i$  帧第  $j$  条谱线， $b$  是 BARK 谱临界带编号， $S$  是临界带个数， $bl$  和  $bh$  分别是临界带  $b$  的起止谱线， $M$  是总帧数。

由于调整后语音长度与原始语音长度不同，计算 BARK 谱失真测度时，将调整后的语音帧移设为原始语音帧移的  $1/\alpha$  倍，BARK 谱临界带取 25 个，在不同的时长调节因子  $\alpha$  下，原始算法与本文快速算法得到的失真测度如图 9 所示：

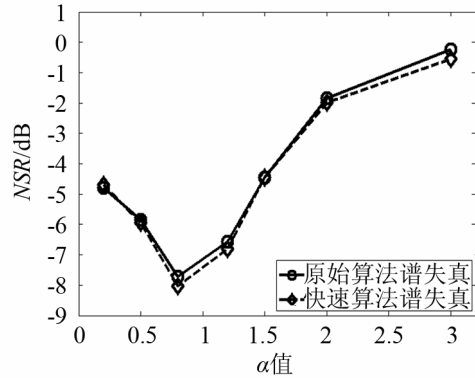


图9 两种算法失真测度比较

Fig.9 The comparison between two algorithms' distortion measures

从图 9 中可观察到，相比于原始算法，快速算法不但没有造成语音质量降低，反而在某些调节因子下语音质量有所改善。这个结果首先证明了隔点搜索和隔点计算相似性的正确性，其次说明了只搜索叠加部分相似性会使波形拼接部分更加光滑，带来更好的语音质量。

另外，从图 9 中还可看出，调节因子越偏离原始的  $\alpha=1$  位置，其语音质量越低，并且调节因子  $\alpha < 1$  时语音质量下降速度要慢于  $\alpha > 1$  的情况。这说明语音在压缩过程中丢失的信息更多。

关于  $T_s$ 、 $T_c$  两个参数的选取规则如下：针对基频较低的人，由于语音信号和相关性系数变化均更加平缓，可适当增大  $T_s$ 、 $T_c$  的值，加快计算速度，对于基频较高的人，可适当减小  $T_s$ 、 $T_c$  的值，以防止搜索位置出现过大偏差。本文算法及其所对应参考值已在实际项目中使用，性能稳定，适用性广。

## 5 结论

本文针对 WSOLA 算法计算量大的问题，深入分析语音信号自身的特性，围绕语音信号自身短时平稳、准周期性特点，提出了三种解决途径，通过适当调整  $T_{sum}$ 、 $T_s$ 、 $T_c$  可大幅降低计算量，同时又可保证调整后语音的质量，通过与原始算法的对比证明了以上观点。本文提出的快速 WSOLA 算法可

广泛应用于实时性要求较高的场合。

### 参 考 文 献

- [1] W Verhelst, M Roelands. An overlap-add technique based on waveform similarity(WSOLA) for high quality time-scale modification of speech[J]. Proc ICASSP, 1993, 2(7): 554-557.
- [2] Dorran D, Lawlor R. An efficient time-scale modification algorithm for use within a subband implementation[C]. Proc. Digital Audio Effects, 2003, 9.
- [3] Dorran D, Lawlor, R. Time-scale modification of music using a subband approach based on the bark scale[J]. Acoustics, Speech, and Signal Processing, 2004, 6(4): 225-228.
- [4] 黄昊, 郭立, 李琳. 基于感知敏感成分划分的语音时长规整算法[J]. 数据采集与处理, 2008, 23(6): 740-745.  
HUANG Hao, GUO Li, LI Lin. Time-scale modification of segmentation based on perceptually sensitive portion[J]. Data Acquisition & Processing, 2008, 23(6): 740-745.
- [5] Yim S, Pawate B I. Computationally efficient algorithm for time scale modification(GLS-TSM)[C]. IEEE ICASSP, 1996, 2: 1009-1012.
- [6] Wong W, Au O C. Fast SOLA-based time scale modification using modified envelope matching[C]. Proc. ICASSP, 2002, 3: 3188-3191.
- [7] 毛启荣, 詹永照, 杜守富. 一种快速实时语音个人特征改变方法[J]. 电子与信息学报, 2007, 29(2): 434-438.  
MAO Qirong, ZHAN Yongzhao, DU Shoufu. A fast modification method for personal characteristics of real-time speech[J]. Journal of Electronics & Technology, 2007, 29(2): 434-438.
- [8] 黄昊, 郭立. 分段语音时长规整算法[J]. 声学技术, 2007, 26(6): 1191-1195.  
HUANG Hao, GUO Li. Time-scale modification of segmented speech[J]. Technical Acoustics, 2007, 26(6): 1191-1195.
- [9] 吴淑珍, 赵朝阳. 基于听觉模型的客观音质评价方法研究[J]. 电子学报, 1999, 27(7): 92-94.  
WU Shuzhen, ZHAO Chaoyang. A study of perceptually-based features for objective speech quality evaluation[J]. Acta Electronica Sinica, 1999, 27(7): 92-94.

## 中国船舶重工集团公司第 726 研究所姚蓝教授 到东海研究站进行学术交流活动

2010 年 9 月 1 日, 在东海站成立五十周年之际, 中船重工第 726 研究所姚蓝教授到声学所东海站进行学术交流活动。

姚蓝教授曾任哈尔滨船舶工程学院系主任、中船重工集团公司第七二六研究所总工程师, 现任《声学技术》杂志主编。姚蓝教授长期从事教学和研究工作, 培养硕士 50 余名, 博士近 10 名。完成多项国家重大科研任务, 获国家科技进步一等奖、部级科技进步一等奖等奖项, 并曾获国家有突出贡献中青年专家和部级先进教授等荣誉称号。

上午 10 点, 姚蓝教授作了有关发展水下声学技术装备方面的报告, 关平副站长主持了报告会。会后姚蓝教授与参加报告会的专家和学生就相关问题进行了探讨。

《声学技术》编辑部