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
中图法分类号：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)[1]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[2,3]。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[4]和减少算法计算量的方法[5-8]，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 αL 进行合成。其中 α 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。

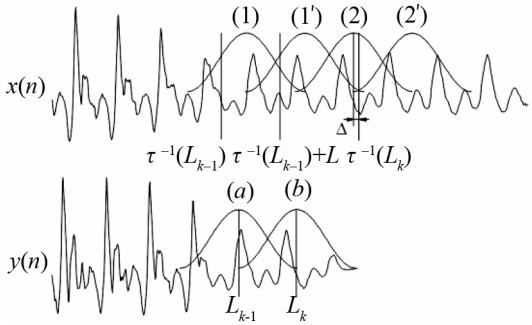

图 1 WSOLA 算法示意图

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，Δ_{k-1} 为上次搜索的最佳相似性位置偏差，δ 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 $T_N=20$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 $T_L=10$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_{sum}

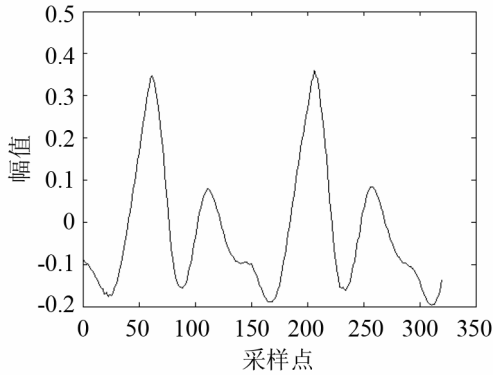

图 2 长为 20ms 的男性发音波形

Fig.2 The 20ms male voice waveform