第 33 卷 第 6 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.6 Computer Engineering March 2007
·多媒体技术及应用· 文章编号：1000—3428(2007)06—0225—03 文献标识码：A 中图分类号：TP391.42

语音驱动人脸动画中语音参数的提取技术

陈 新[1]，周东生[1,2]，张 强[1,2]，魏小鹏[1]
(1. 大连大学辽宁省智能信息处理重点实验室，大连 116622；2. 大连理工大学机械工程学院，大连 116028)

摘 要：语音特征参数的提取是语音驱动人脸动画中语音可视化的前提和基础，该文立足于语音驱动的人脸动画技术，较为系统地研究了语音参数的提取。在参数精度方面，引入了用小波变换重构原始信号的思想，对重构后的信号进行参数提取，从而为语音驱动人脸动画系统建立良好的可视化映射模型奠定了基础。

关键词：语音可视化；语音特征参数；小波变换

Obtainment Technology of Speech Signal Feature in Speech Driven Face Animation System

CHEN Xin[1], ZHOU Dongsheng[1,2], ZHANG Qiang[1,2], WEI Xiaopeng[1]
(1. Liaoning Key Lab of Intelligent Information Processing, Dalian University, Dalian 116622;
2. School of Mechanical Engineering, Dalian University of Technology, Dalian 116028)

【Abstract】The obtainment of speech signal feature is the premises and foundation of audio-visual conversion in speech driven face animation. The obtainment technology of speech signal feature is studied. As for parameter precision, wavelet transform is introduced to reconstruct original speech signal. This research lays the foundations of audio-visual conversion model in the speech driven face animation system.

【Key words】Audio-visual conversion; Speech signal feature; Wavelet transform

1 概述

近年来，随着电脑多媒体交互的发展和网络广播及教学的普及，人们希望看到具有声情并茂特点的仿真人脸动画，并且这种真实度高的人脸动画在电影制作、虚拟主持人、可视电话、虚拟会议等方面存在着广泛的应用前景。语音驱动的人脸动画合成技术[1,2]正是利用语音为原始驱动力，致力于合成能随语音和语调变化的人脸动画。此项技术主要分为 3 大模块：语音处理模块，图形处理模块和人脸动画合成模块。其中语音处理模块中的语音可视化是整个语音驱动人脸动画合成技术的关键部分，其目的是将语音特征参数映射转化为 MPEG-4 标准[2]中的可视化参数来驱动脸部运动。

在这一步工作之前，首先就需要从AVI文件中分化出音频流，对语音信号进行一系列处理提取有效的语音特征参数，用来作为确定语音视频映射模型的学习训练样本。参数的选取及提取的精度将直接影响到映射模型的好坏，文献[3]的神经网络模型仅计算了混合的LPC 以及RASTA-PLP 语音特征，文献[4]仅分析了短时平均能量和LPC复倒谱的计算。本文立足于语音驱动的人脸动画合成技术，较为系统、全面地研究了语音信号的特征以及特征参数的提取技术。同时，为了有效抑制共振峰、大部分基音谐波和高频噪声对参数提取的影响，引入了小波变换[5]重构原始信号的方法。通过对重构信号进行参数提取，提高了参数的精度，为语音驱动人脸动画的后续工作建立良好的语音可视化映射模型奠定了基础。

2 语音信号的特征分析

语音信号是一种典型的随时间而发生改变的一维非平稳信号，根据所分析的参数不同，语音信号分析可以分为时域、频域、倒谱域等方法。时域特征主要有短时平均能量、过零率、线性预测系数等，频域特征主要有线性预测倒谱系数、Mel 频率倒谱系数。语音信号本身就是时域信号，因此时域分析是最为直观最为广泛的一种方法，它实现起来简单、运算量少。但是数字化的语音信号是一个时变信号，不能直接分析，这时可以在一个短时间范围内将其看作是一个准稳态过程，它将语音信号划成一段一段来分析，每一段称为一帧。实践证明，语音在 10m2~30ms 内是保持相对平稳的，即帧长取 10ms~30ms。然后，对一帧信号加窗(直角窗、海明窗等)进行特征提取，接着把帧做一个偏移(通常为一帧的 1/2 或 1/3)，进行下一帧的处理和提取。

$$W_H(n)=0.54 - 0.46\cos[2\prod n/(N-1)] \quad \text{(海明窗) } n=0,1,,N-1 \quad (1)$$

时域分析方法完全是在时间域中分析信号，理论上时间分辨率可以达到无穷大，但频率分辨率为零，而频域分析恰恰相反。一般在频域中分析信号可以使某些特性变得明显从而得到更多的信息，因此在语音信号的分析中更多地依赖频域分析。傅里叶频谱分析是频域分析广泛采用的一种方法，其基础是傅里叶变换。通过傅里叶变换和傅里叶逆变换可以求得傅里叶谱、自相关函数、功率谱、倒谱等。信号 x(n)的短时傅里叶变换如式(2)及图 1 所示。

$$Xn(e^{jw})=\sum_{m=-\infty}^{\infty} x(m)W(n-m)e^{-jwn} \quad n=0,1,...,N-1 \quad (2)$$

基金项目：辽宁省高等学校优秀人才支持计划基金资助项目(RC-05-07)；辽宁省教育厅科学研究计划基金资助项目(05L020)
作者简介：陈 新(1981－)，男，硕士生，主研方向：计算机动画；周东生，博士生；张 强，博士、教授；魏小鹏，教授、博导
收稿日期：2006-03-21 **E-mail：**magic_217@163.com

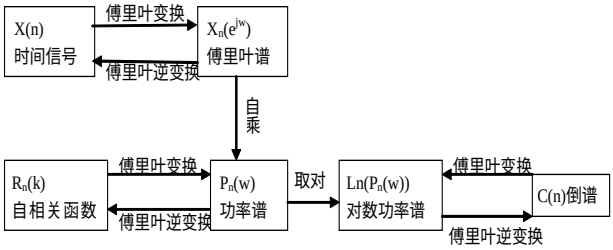


图 1 几种基于短时傅里叶变换谱之间的关系

3 参数的选取和提取

目前，在语音识别领域已被实践证明了有效的特征参数主要有线性预测编码倒谱系数(LPCC)、Mel 频率倒谱系数(MFCC)，在无噪声实验室采集语音视频录像分化出的音频流常用 LPCC，因 MFCC 需要多次利用 FFT，计算复杂度较高。但是在环境较差的情况下多采用 MFCC，它是一个基于人的听觉的语音特征参数，鲁棒性较 LPCC 要好。在语音人脸动画中，也常用到韵律参数，它能在某种程度上反映语音在音高、音强和音长方面显示出来的抑扬顿挫的特性。

3.1 短时能量

短时能量反映了语音振幅或能量随着时间缓慢变化的规律。对于信号 x(n)，短时能量定义如下：

$$En=\sum_{m=-\infty}^{\infty}[x(m)W(n-m)]^2=\sum_{m=n-N+1}^{n}[x(m)W(n-m)]^2 \qquad (3)$$

其中 w(n)为选择的窗函数，根据经验 N=150 或 250 时，基本上能反映出语音信号的特征。由于短时能量的计算涉及到平方运算，有时常用短时幅度来代替语音能量。短时幅度定义如下：

$$En=\sum_{m=-\infty}^{\infty}|x(m)W(n-m)|=\sum_{m=n-N+1}^{n}|x(m)W(n-m)| \qquad (4)$$

3.2 短时平均过零率

过零分析是时域分析中最简单的一种，针对离散信号，相邻的取样值改变符号就称为过零。如果离散时间信号的包络是窄带信号，过零率可以较为准确地反应该信号的频率。在语音信号这种宽带信号的情况下，它却只能粗略地反映信号的频谱特性。短时平均过零率可以用下式计算：

$$Zn=(1/2N)\sum_{m=-\infty}^{\infty}|\mathrm{sgn}[x(m)W(n-m)]-\mathrm{sgn}[x(m-1)W(n-m+1)]| \qquad (5)$$

式中，当 x(m)>0 时，sgn(x(m))=1 ；当 x(m)<0 时，sgn(x(m))=-1。

3.3 LPC 系数的提取

线性预测的基本原理[6~8]是某一个语音的抽样，能够用过去若干语音抽样的线性组合来逼近，在一个有限的时间间隔上，使预测抽样和实际抽样之间差值的平方和达到最小，从而确定唯一的一组线性预测系数。

根据语音声管模型建模假设：给定一 n 时刻采样的语音信号

$$x(n)=\sum_{i=1}^{p}a_i x(n-i)+Ge(n)$$

式中，e(n)为高斯白噪声序列，G 为约束 e(n)的增益系数。设线性预测值为

$$\tilde{x}(n)=\sum_{i=1}^{p}a_i x(n-i)$$

常数a_i(i=1,2,...,p)即为LPC系数，于是可得到实际抽样和预测抽样之间差值e(n)的平方和为

$$En=\sum_{n}e^2(n)=\sum_{n}[x(n)-\sum_{i=1}^{p}a_i x(n-i)]^2 \qquad (6)$$

要使E_n最小，则需

$$\partial En/\partial a_k=-(2\sum_{n}x(n)x(n-k)-2\sum_{i=1}^{p}a_i\sum_{n}x(n-k)x(n-i))=0 \qquad (7)$$

得到

$$\sum_{n}x(n)x(n-k)=\sum_{i=1}^{p}a_i\sum_{n}x(n-k)x(n-i) \qquad k=1,2...,p \qquad (8)$$

其中阶数p常取 12~14。求解方程组(8)比较经典的两种方法是自相关法和协方差法[6~8]，它们都有各自的局限性。本文采用 Burg算法[8~9]计算每一帧的LPC系数，直接通过语音样本递推得到结果，较好地解决了前两种方法中精度和稳定性之间的矛盾。首先建立前向预测误差和反向预测误差的概念，前向预测误差就是通常意义上的线性预测误差e(n)，它用i个过去的样本x(n-1)到x(n-i)来预测x(n)时的误差。反向预测误差为

$$b^{(i)}(n)=x(n-i)-\sum_{j=1}^{i}a^{(i)}_j x(n-i+j) \qquad (9)$$

它可以看作是用时间上延迟时刻的样本 x(n-i+1)到 x(n)来预测 x(n-i)时的误差。Burg 算法是将前向预测和反向预测的平均值作为它的预测误差来计算。以下是 Burg 算法的计算步骤：

(1)初始化：$e^{(0)}(n)=b^{(0)}(n)=x(n)$ n=0 ,1 2,...,N-1

(2)令 i=1；

(3)计算反射系数k_i和预测系数$a^{(i)}_j$和$a^{(i)}_i$：

$$k_i=\frac{2\sum_{n=0}^{N-1}[e^{(i-1)}(n)*b^{(i-1)}(n-1)]}{\sum_{n=0}^{N-1}[e^{(i-1)}(n)]^2+\sum_{n=0}^{N-1}[b^{(i-1)}(n-1)]^2} \qquad (10)$$

$a^{(i)}_j=a^{(i-1)}_j-k_i a^{(i-1)}_{i-j}$, j=1,2,...,i-1 $a^{(i)}_i=k_i$

(4)计算两个误差$e^{(i)}(n)$和$b^{(i)}(n)$：

$e^{(i)}(n)=e^{(i-1)}(n)-k\,b^{(i-1)}(n-1)$

$b^{(i)}(n)=b^{(i-1)}(n-1)-k_i\,e^{(i-1)}(n)$

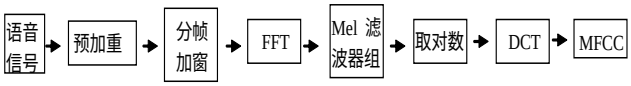
(5)让$i=i+1$，如果i<p，则返回第(3)步，否则结束。最后，仅需要i=p时的系数$a_j=a^{(p)}_j$,j=1,2,3,...,p。

得到LPC系数以后，我们就可以利用a_i和倒谱系数c(n)间的递推关系求出LPCC。递推公式如下：

$$c(n)=a_n+\sum_{i=1}^{n-1}(1-i/n)a_i c(n-i) \qquad n=1,2,...,p \qquad (11)$$

3.4 Mel 倒谱系数的提取

MFCC 系数是 Mel 尺度频率域提取出来的倒谱参数。Mel 尺度反映了人耳对频率感知的非线性特性，它需要设置若干滤波器。以下是 MFCC 系数提取的流程：



第 i 个滤波器为

$$H_i(k)=\begin{cases}\dfrac{2(k-f[i-1])}{(f[i+1]-f[i-1])(f[i]-f[i-1])} & f[i-1]\le k\le f[i]\\ \dfrac{2(f[i+1]-k)}{(f[i+1]-f[i-1])(f[i+1]-f[i])} & f[i]\le k\le f[i+1]\\ 0 & \text{其他}\end{cases} \qquad (12)$$

滤波器的个数 M 通常取为 24，其中 k=0:1:(N/2-1)。具体的提取步骤如下：

(1)预加重，预加重网络的输出 y(n)和输入的语音信号 x(n)

的关系可用方程 y(n)=x(n)-a*x(n-1)表示，其中 a 为预加重系数，$0.9 \le a \le 1$。

(2)对预加重后的信号 y(n)分帧加窗，再做快速傅里叶变换，取模的平方得到离散功率谱 S(k)。

(3)将S(k)通过设计好的滤波器组$H_j(k)$，求功率值。

(4)对滤波器的输出取对数，做离散余弦变换(DCT)，即可得到 Mel 倒谱。

$$c_i = \sqrt{\frac{2}{M}} \sum_{j=1}^{M} \log(S(k) * H_j(k)) \cos[\frac{i*\pi}{M}(j-\frac{1}{2})] \quad i=1,2,...,p \qquad (13)$$

普通的 MFCC 参数只反映了语音参数的静态特性，为获得反映语音动态变化的参数，可以用差分倒谱参数来描述这种动态特性：

$$d(n)=(1/\sum_{i=-k}^{k} i^2) \sum_{i=-k}^{k} i*c(n+i) \qquad (14)$$

其中 k 为常数，经验值为 2 或 4，它实质上就是当前帧的前 k 帧和后 k 帧的线性组合。

3.5 基音频率

基音是指发浊音时声带振动的周期性，是声带基本振动频率的表现，它是语音信号中最重要的韵律参数[6~8]之一。提取基频的方法主要有波形估计法、相关处理法、变换法，权衡计算的精度和复杂度，本文选择了自相关函数法。短时自相关函数在基音周期的各个整数倍点上有很大的峰值，只要找到第一最大峰值点的位置，并计算它到原点的距离便能估计出基音周期，基音周期的倒数就是基音频率。基音频率估计的步聚如下：

(1)对语音信号进行滤波预处理，去掉开头 20 个输出值不用，得到{S(n)}；

(2)求{S(n)}的前 100 个采样点和后 100 个采样点的最大幅度，取其中较小的一个并乘以 0.68 作为门限电平 L；

(3)对{S(n)}进行中心削波得到{x(n)}和三电平量化得到{y(n)}；

(4)求{x(n)}和{y(n)}的互相关值 R(k)：

$$R(k)= \sum_{i=21}^{N} x(i)*y(i+k) \qquad 20 \le k \le N/2 \qquad (15)$$

(5)找到R(20)到R(N/2)中的最大值R_{max}，如果R_{max}小于0.25R(0)，则认为本帧为清音，基频为零；否则基音频率为

$$F_p = \frac{1}{N_p} = \frac{1}{\arg \max\limits_{20 \le k \le N/2} R(k)} \qquad (16)$$

采用了中心削波和三电平量化技术，用中心削波后信号和三电平量化后信号的互相关代替自相关，可达到较好的参数估计效果。

4 小波分析与实验仿真

在语音信号分析中，传统算法都是建立在比较理想的条件下，而实际中在噪声背景下很难达到较为理想的效果。并且，由于共振峰、基音谐波的影响，使参数的精度大打折扣。为此，我们引入了小波变换，对原始语音信号利用小波进行分解提取小波系数。语音信息主要集中在小波变换尺度较大的低频部分，白噪声主要集中在小波变换尺度较小的高频部分，用含有低频特征的小波系数对语音信号进行重构，最后对重构后的信号进行分析提取需要的特征参数。但是进行小波分解时，也会产生高频系数，这样一个信号的高频系数向量是有用信号和噪声信号的高频系数的叠加，若采用强制性的消噪方式将所有高频小波系数置为零，势必丢失部分信号特征。因此，我们采用了SURE和MINIMAXI阈值选取规则[5]仅将部分系数置为零，从而使较为弱小的信号也能提取出来。这样得到的结果不仅能有效克服噪声的影响，而且稳定性好、精度高。由于Daubechies小波基[5,10]具有良好的逼近性与稳定性，实验使用Daubechies小波对原始语音信号进行分解重构，采用海明窗，窗长 256，帧移 128，对信号进行分析求解参数。数据来源于实验室环境下的一段 17s长的录音，采样率为16kHz，量化精度为 16bit。图 2 为原始语音信号的时域波形图，图 3 为用含低频特征的小波系数重构后的信号，图 4 为用Burg算法对第 2 帧原始语音提取的LPC系数曲线，图 5 则为对重构语音提取的LPC系数曲线。

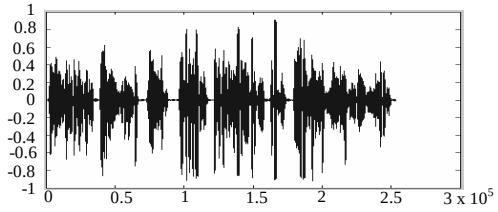


图 2 原始语音信号

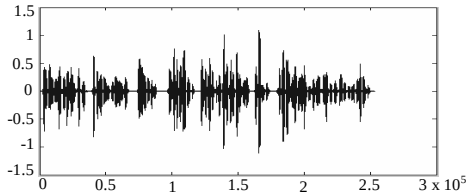


图 3 利用小波重构的信号

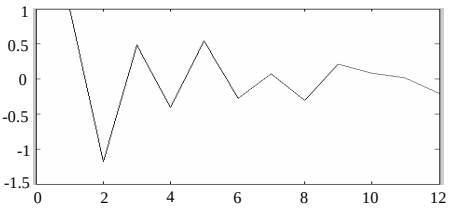


图 4 原始信号第 2 帧的 LPC 参数

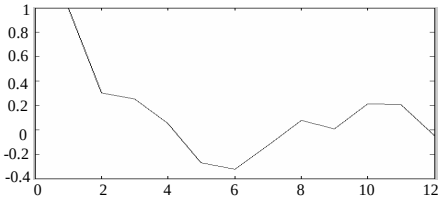


图 5 重构信号第 2 帧的 LPC 参数

5 总结

本文较为系统地研究了语音人脸动画中几种典型的、具有代表意义的语音特征参数的提取，对于 LPC 参数、MFCC 参数和基音频率，给出了详细的计算步骤和算法。并且，为降低噪声和抑制共振峰对参数提取的影响，引入了小波变换。它是一种信号的时间-尺度分析方法，具有多分辨分析的特点，能有效地从信号中提取信息。实验证明，对重构信号提取的特征参数能很好地满足语音处理模块中映射模型的需要，能够提取鲁棒的语音短时特征。下一步的工作主要是在这些特征参数的选择组合上，构成一个合适的特征矢量。另外，在使用时还可以利用聚类和主成分分析等方法来降低混合特征矢量的维数，减少计算量，提高系统的实时性。

(下转第 235 页)