函数拟合实现带声调的语音合成

李建文 王咿卜
(陕西科技大学电子信息与人工智能学院 陕西 西安 710021)

摘 要 语音的声调最能突显个人语气与情感状态,通过加入声调参数这一特征,能够有效提高语音合成的逼真度,使得话语更加具有区分度,并且有效提高了情感识别和语音识别的准确度,弥补了语音合成的结果缺乏情感特征以及语音演唱的不足。采用基频提取的方式分别对汉语一声、二声、三声、四声声调进行分析、研究,最终采用多项式函数拟合的方法进行声调的重新构建,从数学角度对语音声调进行分析、重构,采用三角函数模拟不同时间的语音基频曲线,将不同频率的曲线进行叠加,达到了 95.91% 的满意识别结果。合成结果表明:采用函数拟合方法进行带声调的语音合成,更好地还原了语音的数理本质,使得抽象化的语音表现得更为具体、可控。

关键词 软件工程 语音合成 函数拟合 基频提取 声调 情感

中图分类号 TP3 TN912.33 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2022.09.029

SPEECH SYNTHESIS WITH TONE BY FUNCTION FITTING

Li Jianwen Wang Yibo
(*School of Electronic Information & Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an 710021, Shaanxi, China*)

Abstract The tone of voice can highlight personal mood and emotional state. By adding the feature of tone parameter, it can effectively improve the fidelity of speech synthesis, make speech more distinguishable, improve the accuracy of emotion recognition and speech recognition, and make up for the lack of emotion feature and voice singing in the result of speech synthesis. The first tone, second tone, third tone and fourth tone of Chinese were analyzed and studied by the way of fundamental frequency extraction, and the method of polynomial function fitting was used to reconstruct the tone. From the mathematical point of view, the tone was analyzed and reconstructed. Trigonometric function was used to simulate the fundamental frequency curves of different times, and the curves of different frequencies were superposed. It achieved 95.91% satisfactory recognition results. The synthesis results show that the function fitting method for tone speech synthesis can better restore the mathematical essence of speech and make the abstract voice performance more concrete and controllable.

Keywords Software engineering Speech synthesis Function fitting Fundamental frequency extraction Tone Emotion

0 引 言

语音作为人与人交往最常用的方式之一,是传递情感最有效的手段。中国汉字大约有十万个,是一种独特的声调表意语言[1-2]。在计算机研究领域中,由于汉字种类繁多,在众多语言中所占存储空间最大,导致编码过程中极为不便,但若对相同拼音按不同声调的汉字进行归类,却可以把汉字数目缩减到约原始容量的四分之一,极大地减小了编码空间的占用且保证了语音的逼真度。在如今人工智能高速发展的时代,语音识别及语音合成要做的不仅是算法准确度的提高,还应该注重智能化和逼真度的提高[3]。大部分人仅仅着重于识别与合成的结果,而忽略了语音是否具有合适的声调,尚未实现个性化的语音合成,并没有把话语中声调所表达的情感状态作为考察的特征之一[4-7]。人类是富有情感的,不同的环境、心理状态会导致交谈的音调、声调所传达的情感千差万别[1]。同

收稿日期: 2020-04-24。国家自然科学基金项目(60672001)。李建文,教授,主研领域:皮肤听声,嵌入式开发,计算机网络通信,多媒体编程技术。王咿卜,硕士。

样的语言，使用不同声调所表达的态度也各有所异。在医学中，针对听力障碍者推出的人工耳蜗产品也并未考虑声调、语调等特征的感知[1-2]。因此，从数学角度出发，考虑汉语四种声调的特征参数以及之间参数的变换很有必要。

刘梦媛等[8]设计了基于 HMM 的语音合成系统，选取缅甸语事物声母及带声调事物韵母作为合成基元，解决了变音和变调问题；王国梁等[9]设计了端到端的语音合成系统 Tacotron 2，在语料不足的情况下使用预训练解码器，并通过多层感知机代替线性变化对停止符进行预测；宋刚等[10]基于 Target 模型进行语调分析，总结了四种声调的基频曲线变化规律，采用分段拟合方法，将各个声调分为两段来研究，拟合过程中所需特征参数有各段音调的斜率、音高变化的调域及所占时间；薛健等[11]采用线性多项式进行声调模型的构建，主要从归一化的规范模型出发，建模的参数需要从原始语音得到中值频率、不同音调基频变化的调域、同一音调但调型不同的变化调域。上述研究中，前两者基于深度学习进行语音合成，但深度学习需要极大容量语料包，过程繁琐，且失去了对语音音调的数理本质的探究，而基于 Tacotron 的方法现在更适合对英语的处理，目前对汉语等多文字的语言应用尚不成熟。后两篇论文从基频轨迹出发，讨论了基频曲线与汉语四种音调的关系，并未涉及到基频轨迹拟合四种声调在语音合成方面的实际应用。本文研究旨在从语音声调的角度出发，基于归一化模型的思想，从不同汉语的四声调的基频共性出发，对汉语的四声声调进行分析研究，提取基频轨迹的共性，将其用高次多项式进行拟合，最终以函数形式实现一种音高和音长变化可控、所需参数少且适应于各种发音的声调变换模型的语音合成，以期在语音合成、情感分析领域对语音逼真度和情感度的提高方面提供参考，以及在医学领域对人工耳蜗的构造和声调康复训练方面提供参考[12]。

1 汉语声调规范

1.1 发音原理

声音的形成主要由肺、气管、喉和声道等器官参与。图 1 所示为语音发音原理，空气通过肺器官输出直流气流，产生发音的动力，进入喉，喉部位的声带作为声源，产生振动，输出交流气流，再通过声道对交流气流产生谐振，对声音进行调整，从声道输出的速度波最终经过口唇辐射输出声压波，产生了人耳中听到的声音[7]。

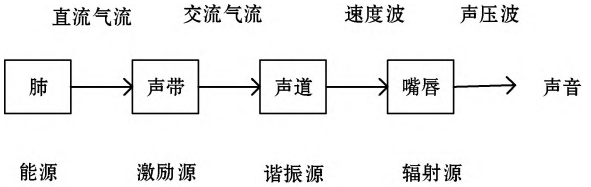

图 1 语音的发音原理

从图 1 可得，声音的发出主要是由声带周期性的振动产生。无论是汉语还是其他语言，语音都可按照声带的参与分为浊音和清音。浊音的发出伴随着声带的振动，清音是气流与空气摩擦产生，没有声带振动的参与，因此本文从浊音角度出发进行语音声调研究。

1.2 五度制音高标记

语言之所以能够体现人类的情感，最主要的特征就在于说话人对于声音声调的选择。相同的话语，不同的抑扬顿挫也会使得情感的偏重点有所差别。虽然每个人说话的腔调与讲话节奏都不同，但相同声调在走向上都是大体一致的。图 2 和图 3 分别为拼音 a 和拼音 o 的四种声调的语音频谱图(称语谱图)，其中 a1 代表拼音 a 一声，a2 代表拼音 a 二声，a3 代表拼音 a 三声，a4 代表拼音 a 四声。

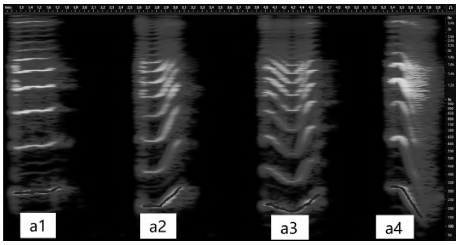

图 2 拼音 a 四种声调的语谱图

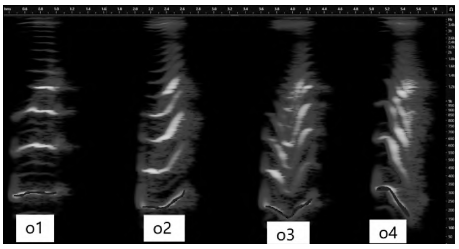

图 3 拼音 o 四种声调的语谱图

将图 2 和图 3 相同的音调进行对比，可以看出相同音调语谱图的曲线走向大致相同。在汉语中，普通话可以按照声调分为四种，分别是阴平、阳平、上声、去声四种音调[10]。图 4 所示为汉语的五度制音高标记法。

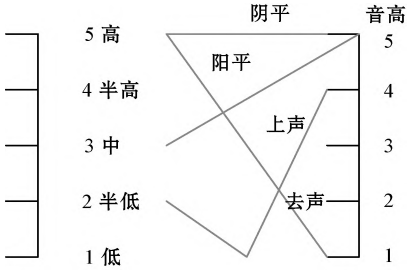

图 4 五度制音高标记法

可以看出,五度最高,一度最低。根据声调的不同,选择的音高也不同,每个汉语都有其对应的音调,即相应的音高走向,但相同音调曲线走向具有同样的共性。

1.3 基　频

在分析语音信号时,主要考察两个重要的参数,其中之一为基频。已知声音的发声源是由声带的周期性振动产生,声带一次的开启与闭合称为一个周期,这种周期的倒数称为基音频率(简称基频)[7,12]。人们所说的声调指的是基频关于时间的曲线。在语谱图上,横坐标为时间,纵坐标为频率,基频指的是位置最低的一条横线对应的纵坐标的值,该值称为基音频率[13-14]。

提取基频,首先要对语音信号进行加窗与分帧的处理,连续信号被分为时域离散信号,m 为起始时间量,得到第 i 帧的语音信号为 $x_i(m)$,长度为 M,对第 i 帧的语音信号 $x_i(m)$ 进行自相关运算[15-16],得到 $R_i(k)$:

$$R_i(k) = \sum_{m=1}^{N-m} x_i(m)\, x_i(m+k) \tag{1}$$

式中: k 是时间的延迟量; N 为语音信号经过分帧处理后每一帧的长度; $x_i(m+k)$ 是移位 k 步的语音信号。已知周期性函数进行自相关计算后,得到的函数同样具有周期性,一个周期内自相关函数图像为递增函数,在周期的整倍数位置处获得最大值[16-18]。由于语音信号的基频具有周期性,周期值为 P,因此采用自相关计算得到的函数也具有周期性,周期仍为 P,且在 P 的整倍数位置处自相关函数会达到最大值 $\max(R_i(k))$。

$$R_i(k) = R_i(k+P) \tag{2}$$

$$\max(R_i(k)) = R_i(\pm nP) \quad n=1,2,\cdots \tag{3}$$

由式(3)知,当 $k=0$ 时,$R(P)$ 为最大值[16]。根据这一原理,采用式(1)对语音信号进行自相关函数运算,在 $R(k)$ 中通过寻找最大值的周期性来确定每一帧语音信号的周期值 P[13]。

1.4 共振峰

语音信号另一个重要的参数为共振峰。在发音过程中,基频由声带振动产生,由于传输到声道发生谐振会产生各次谐波,这些谐波同一时刻所对应的频率值为相应基频的整倍数[7]。在语谱图上,各次谐波有亮有暗,亮区域的波对应的频率值便是共振峰的频率值[14]。由图 1 可知,当不考虑口唇辐射作用时,语音信号是由 n 时刻的声门脉冲激励 $u(n)$(即基频的周期信号)经声道响应 $v(n)$ 滤波得到,即:

$$x(n) = u(n) \times v(n) \tag{4}$$

将式(4)中三个分量求倒谱,得:

$$\hat{x}(n) = \hat{u}(n) + \hat{v}(n) \tag{5}$$

由式(5)可得,在倒谱域中,声门脉冲激励与声道响应两者相分离[19-20]。为了提取共振峰,本文采用倒谱法来获取共振峰的频率值,具体操作如下。

$x(n)$ 是一个长度为 M 的语音信号,将第 i 帧的语音信号 $x_i(n)$ 进行 N 点傅里叶变换,其中 j 为复数的虚部单位,k 为傅里叶变化的第 k 个频谱,得到:

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n)\, \mathrm{e}^{\frac{-j2\pi kn}{N}} \tag{6}$$

取 $X_i(k)$ 的幅值 $|X_i(k)|$,经过对数运算,得到:

$$\hat{X}_i(k) = \log(|X_i(k)|) \tag{7}$$

对 $\hat{X}_i(k)$ 进行逆傅里叶变换,得到倒谱序列,通过低通窗函数 $window(n)$,即矩形窗:

$$window(n) = \begin{cases} 1 & n \leqslant n_0 - 1 \text{ 和 } n > N - n_0 + 1 \\ 0 & n_0 - 1 \leqslant n < N - n_0 + 1 \end{cases} \tag{8}$$

式中: $n=0,1,\cdots,N-1$; n_0 为窗函数宽度。

将式(8)中的窗函数与倒谱序列 $\hat{x}(n)$ 相乘得到 $h_i(n)$,如式(9)所示;再进行 FFT 变换得到 $X_i(k)$ 包络线,如式(10)所示,N 为傅里叶变化的区间长度 $N \geqslant M$,在包络线上取最大值,即得共振峰频率值[20-22]。

$$h_i(n) = \hat{x}_i(n) \times window(n) \tag{9}$$

$$H_i(k) = \sum_{n=0}^{N-1} h_i(n)\, \mathrm{e}^{\frac{-j2\pi kn}{N}} \tag{10}$$

图 5 所示为某一帧信号进行共振峰提取步骤图;图 6 为最终获得的一声拼音 a 语音包络线,其中虚线对应的横坐标的值为共振峰频率。

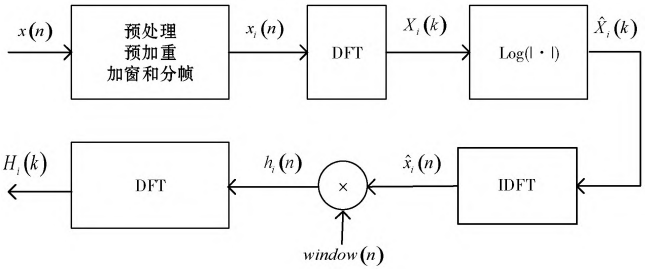

图 5　倒谱法获取语音包络线

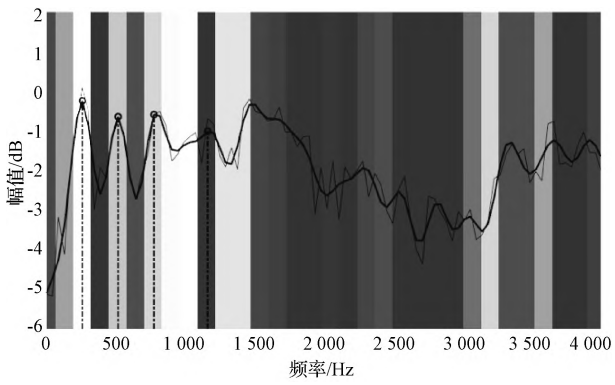

图 6　语音包络线

2 四声声调分析

2.1 声调提取

由图 2 与图 3 可以看出，语音的声调由基频曲线的频率走向决定，因此采用基频提取的方式对声调进行分析。图 7 为实际情况下提取出来拼音 a 的四种声调基频散点图。

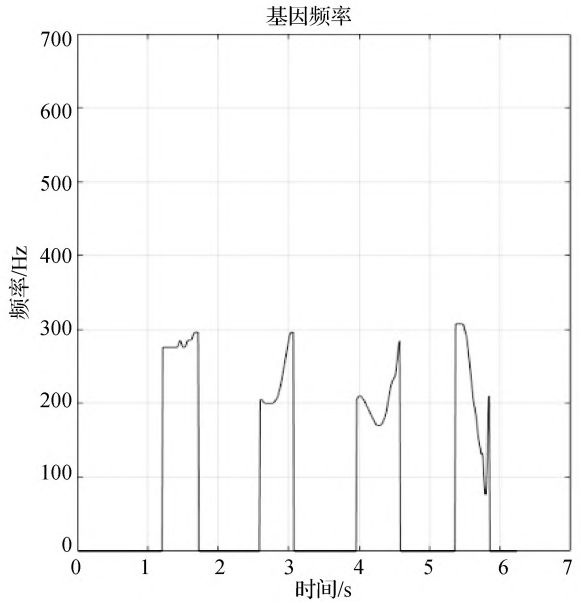

图 7 实际情况下拼音 a 的四声基频曲线

从图 4 与图 7 对比可得，实际情况下提取出来的四声调散点图与理论上的音高走向差异很大。主要区别有以下几点：

(1) 阴平声调的基频走向并不是简单的直线，在开始与结束位置存在小幅度的起伏变换。

(2) 实际情况下，阳平声调的基频变化值由起初 $F_0 \cdot \frac{4}{5}$ 到最终 F_0，与理论上 $F_0 \cdot \frac{3}{5}$ 到 F_0 不同。曲线趋势分为上升段与下降段，拐点更接近前端[10]。

(3) 实际情况下，上声声调的基频变化值由起初 $F_0 \cdot \frac{3}{5}$ 到 $F_0 \cdot \frac{1}{5}$ 再到最终 F_0，与理论上 F_0 到 $F_0 \cdot \frac{1}{5}$ 不同。曲线趋势分为上升段和下降段，拐点位置居中，其幅度变化比阳平变化幅度大[10]。

(4) 实际情况下，去声声调的基频变化值由起初 $F_0 \cdot \frac{6}{5}$ 到最终 $F_0 \cdot \frac{3}{5}$，与理论上 $F_0 \cdot \frac{2}{5}$ 到 $F_0 \cdot \frac{1}{5}$ 再到 $F_0 \cdot \frac{3}{5}$ 不同。曲线趋势变化快，时间短。

2.2 声调拟合

2.2.1 函数最高次数选择

为了使得拟合曲线更接近实际情况下的声调，采用 n 次多项式对实际情况下提取出来的各个音调基频进行拟合：

$$y_l(x) = \sum_{i=1}^{n} a_i x^i \quad l = 1,2,3,4 \tag{11}$$

式中：y_l 为第 l 音调的拟合结果($l=1$ 为阴平，$l=2$ 为阳平，$l=3$ 为上声，$l=4$ 为去声)；i 为次数；a_i 为次数为 i 次的系数；x 为时间序列；a_i 为 x 的系数。

对于次数 n，由多项式性质可得，n 选择越高，函数拟合效果越好，误差越小，但过高会导致过拟合越来越严重。为了防止过拟合且保证有较小的误差，本文统一采用相同的有限次数对四种声调进行拟合。在四种声调中，由于上声声调的基频曲线变化最复杂，因此选择上声调为例进行不同次数拟合。表 1 为多项式不同次数拟合结果。

表 1 多项式使用不同次数拟合结果对比

拟合结果指标	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
SSE	3.088e+04	3 471	2 722	752	734.9
R-square	0.170 7	0.906 8	0.926 9	0.979 8	0.980 3
Adjusted R-square	0.156 4	0.903 5	0.923 0	0.978 3	0.978 4
RMSE	23.070	7.803	6.972	3.698	3.689

其中，误差平方和(SSE)越小，说明函数拟合效果越好；确定系数(R-square)越接近 1，表明拟合函数中的变量对原函数 y 有越强的解释能力，即模型对数据拟合效果越好。RMSE(Root mean squared error)为均方根标准差。Adjusted R-square(Degree-of-freedom adjusted coefficient of determination)为自由度确定系数。

综合分析各种次数的拟合结果，确定了当次数 n 大于等于 4 时拟合效果较好，由于当 n 大于 4 时，各项次数的系数值过于大，基本在 e+04 以上，且拟合效果的提高程度很小。因此，在拟合函数时，选择 $n=4$ 来进行函数拟合，不仅可以有效保证声调的匹配程度，而且简化了参数，减小了运算量。不同拼音的四声调走向有其共性，选择 $n=4$ 来进行拟合，也可以更好地使拟合函数适应不同的语音，避免过拟合。

2.2.2 函数系数

由于本文采用多项式函数进行曲线拟合，因此在拟合过程中，采用最小二乘法进行 n 次拟合。

从原始曲线得数组 (x_i, y_i)，$i=0,1,\cdots,m-1$，x_i 为第 i 点的时间值，y_i 为对应的频率值。以多项式最高次数 n 为 4 进行四次拟合。令拟合函数为：

$$y_l(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 \quad l = 1,2,3,4 \tag{12}$$

$$\boldsymbol{A}=\begin{bmatrix}a_0\\a_1\\a_2\\a_3\\a_4\end{bmatrix}\quad \boldsymbol{X}=\begin{bmatrix}1 & x_1 & x_1^2 & x_1^3 & x_1^4\\1 & x_2 & x_2^2 & x_2^3 & x_2^4\\1 & x_3 & x_3^2 & x_3^3 & x_3^4\\1 & x_4 & x_4^2 & x_4^3 & x_4^4\\1 & x_5 & x_5^2 & x_5^3 & x_5^4\end{bmatrix}$$

则式(12)可化为线性代数形式:

$$y_l(x)=\boldsymbol{XA}\quad l=1,2,3,4 \tag{13}$$

为了保证拟合效果,寻找与原基频曲线样本点(x_i,y_i)距离平方和最小的拟合曲线,采用均方误差Q求极小值来进行系数求解[23-24]:

$$Q(a_0,a_1,a_2,a_3,a_4)=\min\left(\sum_{i=0}^{m-1}\left(\begin{matrix}a_0+a_1x_i+a_2x_i^2\\+a_3x_i^3+a_4x_i^4-y_i\end{matrix}\right)^2\right) \tag{14}$$

对方程$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\begin{pmatrix}a_0\\a_1\\a_2\\a_3\\a_4\end{pmatrix}=\boldsymbol{X}^{\mathrm{T}}\begin{pmatrix}y_1\\y_2\\\vdots\\y_m\end{pmatrix}$求解得到$a_i(i=0,1,2,3,4)$,最终得到拟合函数。

2.2.3　拟合步骤

由于语音波形可以分解为多个三角函数,同样也可以经过三角函数的叠加构成语音波形。三角函数的频率为基频,其各次谐波为基频的整倍数级,三角函数的幅值为基频及各次谐波的强度,由此进行曲线拟合。图 8 所示为拟合步骤。

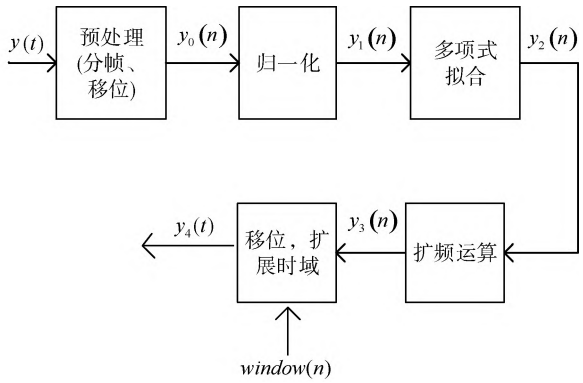

图 8　声调拟合步骤

根据图 8 流程,可将声调合成分为以下几步:

(1) 将获取的基频连续曲线$y(t)$进行预处理,首先对声调经过分帧处理,得到离散点,初始横轴位置为n_0,声调频率最高位置为y_0,将曲线移至横轴初始位置$y(n-n_0)$,为使得拟合函数统一并且方便处理,将曲线纵轴初始位置设置为 0,即$y(n-n_0)-y_0$,得$y_0(n)$。

(2) 为了使拟合函数能够根据实际情况进行音高控制,将得到的n时刻的$y_0(n)$进行归一化,让曲线的频率最高值为 1,最低点为 0,根据式(15)得$y_1(n)$:

$$y_1(n)=\frac{y_0(n)-\min(y_0(n))}{\max(y_0(n))-\min(y_0(n))} \tag{15}$$

(3) 对$y_1(n)$采用二项式定理确定多项式的系数,得到拟合函数。

(4) 由于音调的频率变化差值较大,因此需要对拟合函数进行纵轴的扩频以实现真实的幅度变化,通过获取原始语音的音高差$\max(y(n))-\min(y(n))$来对拟合函数进行扩频,以实现正确的音高变化,根据式(16)得到$y_3(n)$。

$$y_3(n)=y_2(n)\cdot(\max(y(n))-\min(y(n))) \tag{16}$$

(5) 扩频之后的拟合函数$y_3(n)$与实际曲线$y(n)$的音高仍存在差异,因此要通过移位使得拟合函数的初始频率达到原始音频的初始频率,由拟合函数$y_3(n)$的中值频率y_{3c}与实际曲线$y(n)$的频率中值y_c的差值决定移位量,更好地保证了合成的基频曲线不受原始语音基频两端不稳定点的影响。最终由式(17)得到拟合结果y_{41}。

$$y_{41}(n)=y_3(n)+(y_{3c}-y_c) \tag{17}$$

将拟合结果进行语音参数读取,得到声调变化的时域信息(初始位置为t_0,结束位置为t_1),采用矩形窗进行时域截取,如式(18)所示。为了使得声调变化时长可控,设最终发音时长为t_2,f_s为采样率,N为语音信号分帧后的长度,进行扩展最终得到$y_4(n)$,如式(19)所示。

$$y_{42}(n)=y_{41}(n)\cdot window(n) \tag{18}$$

$$y_4(n)=y_{42}(n)\left(\frac{t_2\cdot f_s\cdot(t_{01}-t_{02})}{N-1}\right) \tag{19}$$

2.2.4　pitch 模型

通过上述步骤依次可得四种声调的拟合函数模型的参数分布及拟合结果,如表 2 所示。

表 2　四种声调的拟合参数分布及拟合结果

声调	函数模型: 式(12) $l=1,2,3,4$					R-square
	a0	a1	a2	a3	a4	
阴平 y_1	0.453 0	0.983	−16.18	95.30	−181.23	0.870 3
阳平 y_2	0.000 4	0.415	3.29	103.16	−270.67	0.978 9
上声 y_3	0.498 7	1.280	−107.74	659.14	−975.25	0.980 2
去声 y_4	1.029 0	−2.400	−14.86	71.76	−117.01	0.997 5

分析表 2 中的数据可得,阴平的基频曲线变化幅度较小,阳平次之,上声和去声的基频曲线变化幅度较大。根据最终得到确定系数与极限值 1 相比可得,通过四次多项式进行语音基频拟合方法可行。

3 实 验

3.1 拟合结果

根据表 2 中四种声调的拟合函数参数，令发音时长为 1，基频的频率最大值为 300 Hz，最终得到四种声调基频发音曲线，如图 9 所示。

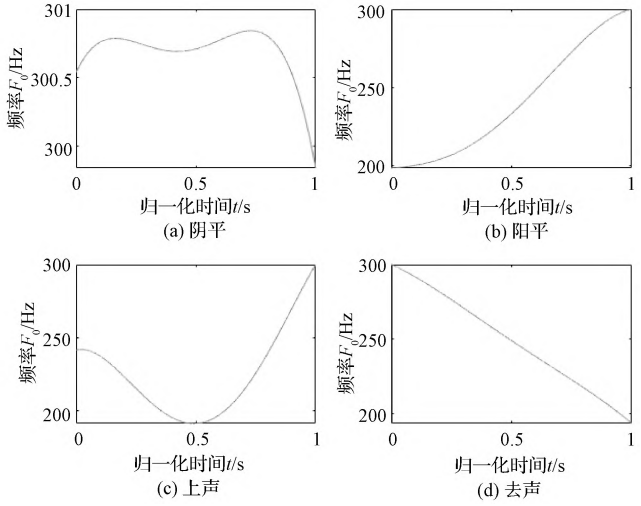

图 9 四种声调拟合的基频曲线对比图

由图 9(a) 可得，阴平的曲线在实际情况下并不是单一的直线，在最高频率 300 Hz 时，有较小幅度的波动。图 9(b) 中阳平的基频曲线有拐点，拐点之前为斜率递增，拐点之后斜率递减。图 9(c) 中上声的基频曲线有拐点，拐点之前为斜率递减，拐点之后斜率递增。图 9(d) 中去声基频曲线在发音中间阶段先有小幅度的频率波动。

由 2.2.2 节可知，语音可以经过多个三角函数叠加构成，如式(20) 所示。

$$output(t) = \sum Amp \cdot \sin(wt + \varphi) \tag{20}$$

式中：Amp 为幅度，控制声音的响度；w 为声带振动频率；t 为时间；φ 控制声音发音时间的移位。由于 $w = 2 \cdot \pi \cdot f$，f 为基频周期，则式(20) 变化为如下函数：

$$output(t) = \sum Amp \cdot \sin(2 \cdot \pi \cdot f \cdot t + \varphi) \tag{21}$$

在语音合成过程中，要实现声调控制，需要将固定的声带振动频率即式(20) 中的定值 w 变为随着时间有相应声调起伏变化的函数，即 $y_l(n)$，实验合成语音选取的采样频率为 8 kHz，因此在合成过程中，时间的间隔 n 值非常小，即离散的采样取值可以等效为连续时间变化 $y_l(t)$。

$$output(t) = \sum Amp \cdot \sin(2 \cdot \pi \cdot k \cdot y_l(t) \cdot t + \varphi) \tag{22}$$

式中：$y_l(t)$ 为式(12) 中四种声调拟合函数；k 为基频的整倍数级；$2 \cdot \pi \cdot k \cdot y_l(t)$ 为共振峰频率。

根据式(22) 最终从数学原理角度出发实现了带有音调控制的语音合成。采用 Adobe Audition 软件进行分析，将语音的原声和合成语音进行语谱图对比。图 10 – 图 13 分别为原声和合成的拼音 a 的四种音调的语谱对比图(左侧为原声语谱图，右侧为合成语音语谱图)。

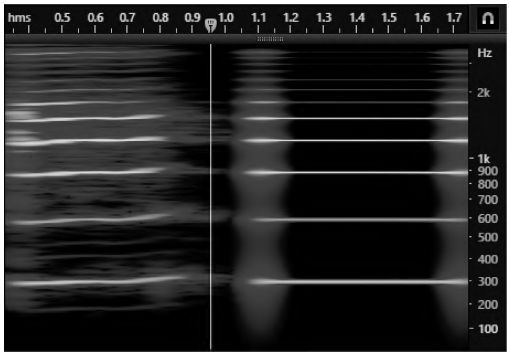

图 10 拼音 a 阴平的原声与合成结果对比

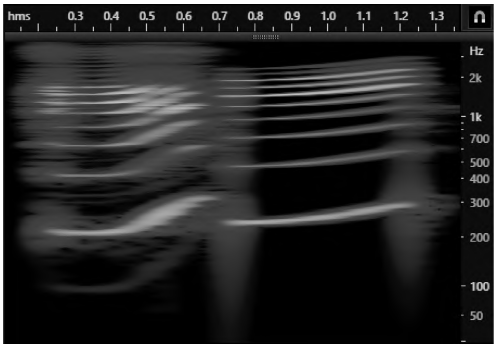

图 11 拼音 a 阳平的原声与合成结果对比

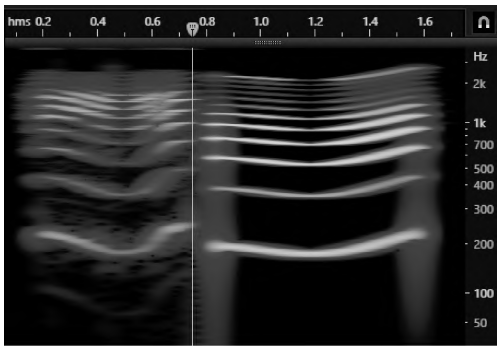

图 12 拼音 a 上声的原声与合成结果对比图

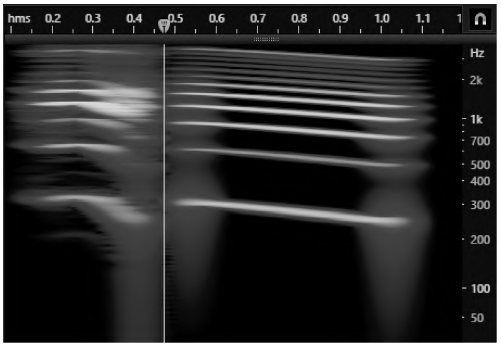

图 13 拼音 a 去声的原声与合成结果对比图

可以看出，由于现实情况下，人受身体状态与发音器官构造的差异，使得语音的发出在语谱图上会呈现

一些有干扰的阴影,影响发音效果[9]。对于越标准的发音,基频与共振峰曲线越清晰,存在的阴影越少。为合成清晰度高、干扰小的语音,本文采用函数拟合方法可以很好地去除外界对发音的影响,使得发音结果更标准。图 10 – 图 12 对应的一声、二声、三声声调都能够得到很好的拟合结果,而四声声调存在偏差是因为在实际情况下,基频的变化不是从刚开始就下降,一般先保持一段水平进而开始走低,由于这段水平发音时间很短且保持一声,因此在进行函数拟合时,可以利用平缓的下降来进行拟合,最终得到拟合结果。

为直观地检测通过人耳后合成结果与原始语音的听觉差,采用 Sound-Similar Free 软件对两种结果进行相似度检测,该软件通过做时域分析,获取频谱随着时间变化的特征向量来计算相似度,最终得到四种声调的相似度对比结果。图 14 为拼音 a 上声(a1)的合成结果与原始语音的相似度对比检测结果。

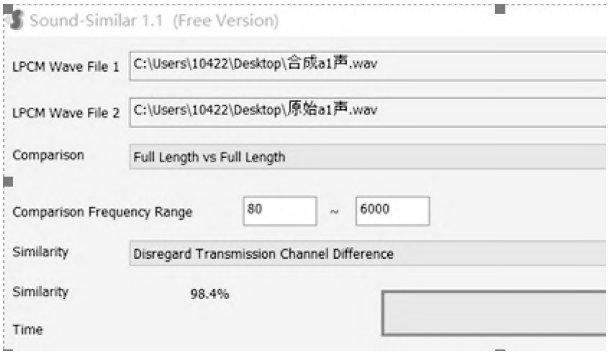

图 14　拼音 a 上声相似度检测结果

采取同种方法测试四种声调的合成结果,表 3 为四种声调的检测结果。表 3 分析结果与图 10 – 图 13 频谱分析对比结果贴合。

表 3　四种声调的相似度对比结果(%)

指标	阴平	阳平	上声	去声
相似度	98.4	93.6	89.7	83.5

3.2　实验对比

现在大部分考虑声调的语音合成系统,主要采用 Target 模型及二次曲线拟合方法。在该模型中,四种声调被简单地划分为斜率为零、上声、下降不同且变化趋势单一的直线,结合二次曲线计算基频曲线拐点位置进行拟合[10]。由于三声调曲线变化最复杂,因此以三声调为例进行实验对比。图 15 所示为拼音 a 上声的原声和以 Target 模型为基础的语音合成结果,图 16 所示为对于拼音 a 上声采用高次多项式和以 Target 模型为基础的语音合成结果。

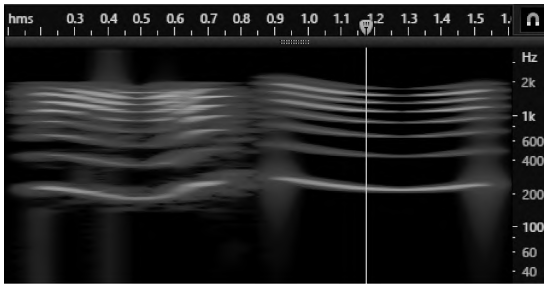

图 15　拼音 a 上声的原声与 Target 模型为基础进行合成结果对比图

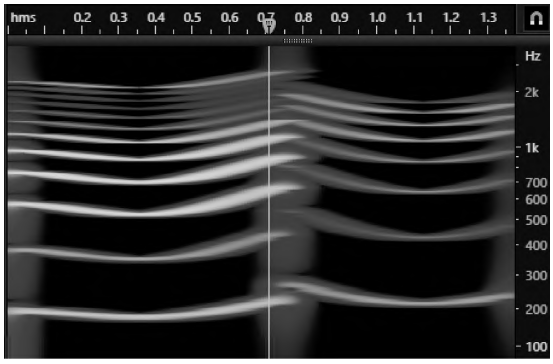

图 16　拼音 a 上声的多项式拟合与 Target 模型为基础进行合成结果对比图

可以看出,由于语调曲线变化不是单一的,而是变化复杂且拐点较多的,因此采用高次多项式,较以 Target 模型为基础进行带语调的语音合成结果得到的拟合效果更好。

利用支持向量机的方法对声调拟合参数进行训练、分类,最终得到两种方法关于四种声调的识别结果,如表 4 所示。

表 4　四种声调的识别结果(%)

模型	阴平	阳平	上声	去声	总体识别率
Target 模型	98.4	91.7	92.5	96.9	95.17
多项式曲线拟合	98.3	93.5	94.6	96.7	95.91

由表 4 可得,对于阴平和去声来说,由于两种声调的基频变化曲线都是单一的,因此识别率几乎没有差别; 而对于阳平与上声音调来说,采用多项式进行基频曲线拟合效果更好,总体识别率也更高。虽然采用多项式曲线拟合方法进行转换之后,阳平和上声较阴平和去声识别结果的正确率较低,但总体上看,此曲线拟合技术已经可以达到使用的效果。

4　结　语

(1) 语音发音两个重要的参数为基频和共振峰。

对语音的某一帧频率值进行自相关运算，在周期处存在极大值。基频值采用自相关运算求极大值方法求得。共振峰的频率值可根据倒谱法求得。

(2) 四种语调的基频曲线在实际情况下，一声语调存在波形变化，二声与三声语调基频与五度制音高标记法描述的音高走向不同，拐点更接近前端，四声语调基频下降趋势更快，时间更短。

(3) 采用多项式进行基频曲线拟合，选择四阶多项式拟合与原始曲线相似度可达到 97.98%，同时避免了曲线过拟合。

(4) 对发音的数学原理进行分析，提取了语音的基频及共振峰两个重要参数，最终通过三角函数的叠加以及四种音调的控制实现了声调可控的语音合成。相比传统的基频提取，本文方法能够通过函数拟合来灵活调整语调；相比机器学习，本文方法对语料包的要求更低。经过验证，本文方法达到了 95.91% 的识别率，对于今后语音合成、情感分析、语音识别的智能化、准确度有很好的参考价值，对探究发音的数学原理有参考意义。

参 考 文 献

[1] 亓贝尔，古鑫，刘子夜，等. 汉语普通话人工耳蜗使用者对声调识别的分析研究[J]. 中国耳鼻咽喉头颈外科，2017，24(4)：175－179.

[2] 杨丽萍，卢岭，刘莉，等. 人工耳蜗使用者汉语声调感知与音乐感知相关性研究[J]. 中华耳科学杂志，2019，17(6)：905－909.

[3] 张丹烽，李冠宇，赵英娣. 语音合成技术发展综述与研究现状[J]. 科技风，2017(22)：72.

[4] Luo X，Lauren H. Vibrotactile stimulation based on the fundamental frequency can improve melodic contour identification of normal-hearing listeners with a 4-channel cochlear implant simulation[J]. Frontiers in Neuroscience，2019，13(10)：1145－1158.

[5] Yu K K，Li L，Chen Y，et al. Effects of native language experience on Mandarin lexical tone processing in proficient second language learners[J]. Psychophysiology，2019，56(11)：13448.

[6] Han Y Q，Goudbeek M，Mos M，et al. Relative contribution of auditory and visual information to mandarin Chinese tone identification by native and tone-na?ve listeners[J]. Language and speech，2020，63(4)：856－876.

[7] 宋知用. MATLAB 在语音信号分析与合成中的应用[M]. 北京：北京航空航天大学出版社，2013：16－20.

[8] 刘梦媛，杨鉴. 基于 HMM 的缅甸语语音合成系统设计与实现[J]. 云南大学学报(自然科学版)，2020，42(1)：19－27.

[9] 王国梁，陈梦楠，陈蕾. 一种基于 Tacotron 2 的端到端中文语音合成方案[J]. 华东师范大学学报(自然科学版)，2019(4)：111－119.

[10] 宋刚，姚艳红. 用于汉语单音节声调识别的基频轨迹拟合方法[J]. 计算机工程与应用，2008，44(29)：239－240，244.

[11] 薛健，蔡莲红. 一种基于声调规范模型的声调变换方法[J]. 计算机工程与应用，2005，41(10)：40－43，85.

[12] Lima T A D，Costa-Abreu M D. A survey on automatic speech recognition systems for Portuguese language and its variations[J]. Computer Speech & Language，2020，62：101055.

[13] Wu H，Dong X X，Wang Q M. New principle of busbar protection based on a fundamental frequency polarity comparison[J]. PloS One，2019，14(3)：0213308.

[14] 李永，范雪，杨鸿波. 声谱图在汉语普通话声调识别中的应用[J]. 信息通信，2017(7)：89－92.

[15] Sampaio M C，Bohlender J E，Brockmann-Bauser M. Fundamental frequency and intensity effects on cepstral measures in vowels from connected speech of speakers with voice disorders[J]. Journal of Voice，2019，35(3)：422－431.

[16] 马效敏，郑文思，陈琪. 自相关基频提取算法的 MATLAB 实现[J]. 西北民族大学学报(自然科学版)，2010，31(4)：54－58，63.

[17] 曹梦霞，郑永果，郑尚新. 基于归一化自相关的语音基频特征提取[J]. 信息技术与信息化，2014(2)：49－51.

[18] 吴树兴. 一种语音信号基音周期时域估计算法[J]. 电脑知识与技术，2019，15(22)：214－216.

[19] de Carvalho C C，da Silva D M，de Carvalho A D，et al. Evaluation of the association between voice formants and difficult facemask ventilation[J]. European Journal of Anaesthesiology，2019，36(12)：972－973.

[20] 白燕燕，胡晓霞. 基于基音周期和共振峰频率检测的倒谱特征研究[J]. 电子测试，2019(17)：48－49.

[21] 王硕，Mannell R，Newall P，等. 共振峰信息在汉语声调感知中的作用[J]. 中国耳鼻咽喉头颈外科，2012，19(1)：8－11.

[22] Hu G X，Determan S C，Dong Y，et al. Spectral and temporal envelope cues for human and automatic speech recognition in noise[J]. Journal of the Association for Research in Otolaryngology，2020，21：73－87.

[23] 张勤. 最小二乘估计在曲线拟合中应用的研究[J]. 成功(教育)，2011(18)：302－303.

[24] 刘霞，王运锋. 基于最小二乘法的自动分段多项式曲线拟合方法研究[J]. 科学技术与工程，2014，14(3)：55－58.