韵律参数和频谱包络修改相结合的情感语音合成技术研究

邵艳秋 韩纪庆 王卓然 刘挺
(哈尔滨工业大学计算机学院，哈尔滨 150000)

摘 要：情感语音合成可以增强合成语音的表现力、人情味，是近年来的新兴课题。除了韵律特征之外，音质类和发声器官类参数对情感语音的表达也有着至关重要的影响，而通常的研究大多都是基于规则或者预先为某种情感设计的滤波器来进行这两类参数的修改。本文提出了通过频谱包络综合地调整音质类和发声器官类参数来合成情感语音的方法，并通过实验验证了这一方法的有效性。另外，实验结果也显示了当韵律参数和频谱包络同时得到修改时，相对于单独修改某类参数可以获得更好的情感合成效果。

关键字：情感语音合成；频谱包络；韵律修改

Emotional speech synthesis based on the modification of prosody parameters and spectral envelope

SHAO Yan-qiu HAN Ji-qing WANG Zhuo-ran LIU Ting
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150000)

Abstract: Emotional speech synthesis, a recently developed research subject, is expected to make the synthesized speech more expressive and human-like. Besides prosody features, voice quality and articulatory parameters are also the important factors that should be considered in emotional speech synthetic systems. Generally, rules and filters are designed to process these two kinds of parameters respectively. This paper presents that by modifying spectral envelope, the voice quality and articulatory could be adjusted as a whole. The experiments results also show that when the prosody features and spectral envelope are all modified, the best synthetic emotional speech could be got.

Key words: emotional speech synthesis; spectral envelope; prosody modification

1 引言

语音作为语言的声音表现形式，是人类交流信息最自然、最有效、最方便的手段。人类的语音中不仅包含了语言学信息，同时也包含了人们的感情和情绪等非言语信息。传统的语音处理更注重语音词汇传达的准确性，而忽略了对非言语信息的研究。如何在人机交互过程中使计算机能够具备不但能听懂带有情感的语音，而且还能够表达出情感语音的能力是真正实现人机和谐的关键。研究情感语音合成的目的就是要使合成语音听起来更自然，更有表现力和人情味。

情感语音合成技术的发展同语音合成方法的研究是密不可分的，主要包括共振峰合成法和拼接合成法[1]。比较著名的共振峰合成器如 Cahn 采用 DECtalk 共振峰合成器研制的 Affect Editor 情感语音合成器[2]，以及 Iain Murray 的 HAMLET 情感语音合成器[3]。共振峰合成器可以通过设计规则，灵活地进行参数调整，但缺点是合成语音有明显的机器音，不够自然。而波形拼接的合成方法不必像共振峰合成器那样为激励源和声道建立参数模型，而是直接从音库中选择具备合适的韵律特征的基元进行拼接来生成合成语音。Iida 和 Campbell 等开发的情感语音合成系统就是使用的这种方法[4][5]。如果基元片断较长，采用该方法合成出的语音自然度会很高，但缺点是如果语音库中没有合适的满足合成需要的韵律的基元，那么合成效果就会很差，另外该方法对音质修改能力也比较有限。

对合成系统而言，无论采用何种语音合成方法，根据影

收稿日期：2005年12月8日；修回日期：2006年2月25日

响情感表达的特征参数建立情感模型都是很关键的。然而，当说话人处在不同的情感状态时，语音的各方面特征都会有相应的改变，而人对情感的感知又是一个极其复杂的过程，所以我们通常只能提取出语音一部分特征来对某种情感进行刻画。通常这些参数可以分为三类：韵律类、音质类和发声器官类[1][6][7]。韵律类参数主要反映了不同情感下语气的变化，包括基频均值、基频曲线、音强和语速等。发声器官类参数和声道的状态相关，反映了不同情感中元音质量的变化，如鼻音、元音质量和声道肌肉紧张程度等。音质类参数则用来表征不同情感下语音音质发生的变化，主要有呼吸声、吱嘎声和斯哑声等。

共振峰合成器在修改音质类和发声器官类参数时相对于拼接式合成系统要容易些，但合成效果比较差。如果仅仅调整韵律参数，而不调整音质类和发声器官类参数，是否足以表达情感目前尚没有清晰的定论。Nagasaki 认为当韵律类参数发生变化时，很容易辨别出情感变化[8]；Moriyama 和 Ozawa 认为只凭借韵律参数可以较好的表现惊奇，生气和悲伤这三种情感，但对区分高兴和恐惧这两种情感是不够的[9]；Gobl[10] 通过对比实验证明利用合适的音质合成的语音情感远比只用基频参数明显，从而强调了音质类参数在表达情感中的重要作用。

到目前为止，多数研究者在合成情感语音时，尤其是在修改音质类参数方面，采用的方法多为经验地添加某种音质，或根据规则，或预先为某种情感设计滤波器来修改共振峰等。而没有用统计或机器学习的方法，为不同的语音自动生成音质类和发声器官类相关的参数来自动合成情感语音。本文提出了通过修改语音信号的频谱包络，来整体调整音质类和发声器官类参数，使其表现出特定的情感的假设。这样就避免了对每个参数的作用和效果单独、细致地讨论。其意义在于，如果能够通过机器学习的方法，生成合适的频谱包络，如生成 LPC 系数，便可以实现一个灵活的，自动的情感语音合成系统。除此之外，我们也在实验中证明了当综合调整韵律参数和频谱包络进行合成时，要比单独调整韵律参数或频谱包络的效果好得多。

2　情感类型选择及语料库的建立

2.1　情感类型选择

在进行情感语料库的建立之前，首先需要确定情感类型。目前情感类型的划分主要有离散的表示和连续维度的表示两种[11]，而离散表示和维度表示在某种程度上是可以相互转化的。离散的情感表示是将情感划分为基本类和扩展类[12]，扩展情感是由基本情感变化混合而成的，至于基本情感的数量从两种至八种不等，学术界尚未达成共识。综合考虑维度空间常见的激发度和评价度以及离散情感表示，本文选择了三种情感进行了语料的录制，即高兴、生气和悲伤，生气和高兴属高激发度，悲伤属低激发度，同时悲伤和生气属负评价范畴而高兴属正评价。

2.2　语料采集

为了避免句子本身语义所带有的情感倾向性对情感判别造成的影响，本文选择了 310 句具有较高情感自由度，在语义上较为中性的句子进行情感语料的录制。这些句子的平均长度为 7.5 个汉字，由一青年女发音人在专业的录音室录制完成，保存为 16khz，16bit 量化的单声道波形文件。每个句子均采用 4 种情感方式来朗读，即中性，高兴，生气和悲伤。为了保证发音人能有效、正确地表达出相应的情感，在录制某种情感语音之前，首先要调动发音人的情绪，比如在朗读高兴的句子前，发音人先要设想一些高兴的事情，或看一些开心的影像资料，朗读一些能令发音人感到高兴的句子，待发音人的情绪被充分调动起来之后再朗读相应情感的正式句子。有 10 名学生参加的对随机抽取的 100 个情感句子进行的听辨测试表明，其中表示高兴和生气的句子被 100% 的正确识别，悲伤的识别率也高达 98%。这说明我们录制的情感语料库是能够正确地表达相应的情感的。另外，这些句子都进行了基频曲线的提取和音节边界的切分。

3　情感语音合成的实验

为了验证我们提出的频谱包络的修改可以综合修改音质类和发声器官类参数的假设，以及为了考察各参数对情感合成所起的作用，我们进行了三组实验。实验的内容包括：(1) 单纯通过修改韵律参数来合成情感语音；(2) 单纯通过修改频谱包络来合成情感语音；(3) 同时修改韵律参数和频谱包络来合成情感语音。最后，我们对合成的结果进行了人工评价。通过评价的结果，可以看出同时修改韵律参数和频谱包络可以合成出情感较为明显的语音，同时也验证了我们前面所提出的假设。实验的原理和具体方法将在 3.1 ~ 3.3 小节中详细介绍。而在 3.4 小节中我们对三组实验合成语音进行了人工的比较分析和情感识别率的测试。

3.1　修改韵律参数合成情感语音的实验

实验以中性情感的句子作为原始语音，并用上述三种情感语音的韵律参数对该语音相应的韵律参数进行修改，从而得到合成的情感语音。具体方法为，在中性情感的句子中截取每个音节，然后提取三种情感语音中该句子的对应音节的基频曲线，音节长度和音节平均能量，用 PSOLA 技术按照这些值对原中性情感语音的音节进行修改然后拼接，最终生成合成的情感语音。其过程如图 1 所示。

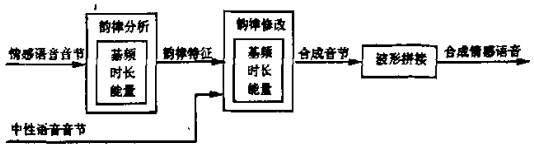


图 1　修改韵律参数合成情感语音过程

由于实验提取的基频曲线，音节长度和音节平均能量来自于真实的情感语音，因此可以认为此韵律参数是理想的，

并且,基频曲线、音节长度和音节平均能量基本能全面地反映韵律类参数中的基频变化、语速和音强。图2展示了原始的情感语音的波形和基频曲线以及合成情感语音的波形和基频曲线。可以看出,修改后的基频曲线和音节长度基本和人工朗读的语音的基频曲线和音节长度一致。也就是说可以认为PSOLA技术对信号的修改也是理想的。

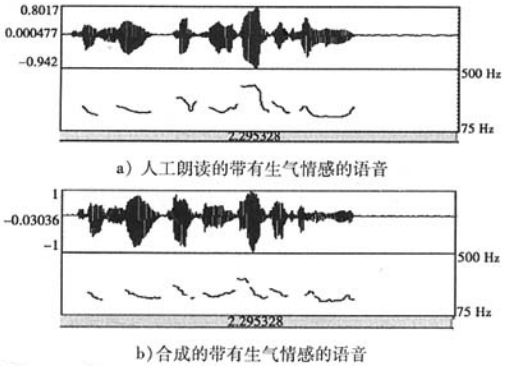


图2 句子"可是,这绝对不可能!"带有生气情感时的波形及基频曲线

韵律参数完全从人工朗读语音中提取,而音节的选择也来自对应句子的人工朗读语音,是为了排除韵律生成模块和单元选择模块的不理想带来的干扰,从而单纯地在理想条件下讨论韵律参数对情感语音合成的作用。但在听觉效果上,单纯修改韵律参数后合成的语音并没有体现出明显的情感,只是一些句子的语气语调上略带有特定情感的特征。比如,高兴和生气会伴随着基频的升高,而悲伤会有明显的基频降低。但是从听觉上并不能感觉到明显的情感倾向。这说明在汉语的情感语音合成中单纯利用韵律参数是不足够表现情感的。于是我们继续尝试加入其他的参数来体现情感特征。

3.2 通过修改频谱包络合成情感语音的实验

实际上,发声器官类的参数都是声道状态相关的参数,因此与频谱包络的形状和共振峰位置,共振峰带宽等密切相关[13][14]。而音质类参数则主要是与喉部发声状态相关的,但它们的一些特征也会在频谱上有所体现。如产生呼吸声的时候,会增加高频部分谐波之间的噪声,而且声门通过的气流量会影响第一谐波分量和第一共振峰的带宽;由于产生吱嘎声音质现象时声门的脉冲很窄,通常会有一个比较平滑的频谱[15];而产生嘶哑声时也会伴随有频谱的噪声[14]。可见,频谱包络记载了很多与发声器官和音质类参数相关的特性。因此在本实验中,我们采用通过修改频谱包络来综合调整发声器官类参数和音质类参数,从而避免对每个参数进行单独修改。

我们仍就以3.1实验中的中性情感的语音作为原始语音,提取出三种情感语音中对应音节的LPC参数。由于语音信号的短时周期性,我们以20ms为一帧截取音节元音中的一个片断,可以近似认为在这一帧内语音信号有稳定的周期。我们将原始语音的音节和情感语音的对应音节都以帧为单位做16阶的LPC分析,然后将其频谱分解为包络和激励源谱。最后,我们用情感语音的频谱包络替换掉原始语音的频谱包络,再与原始语音的激励源谱合成新的频谱,然后逆变换生成时域波形。其过程如图3所示。

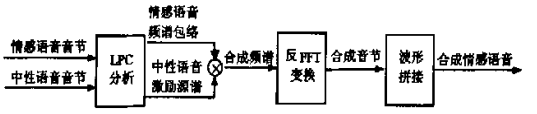


图3 修改频谱包络合成情感语音过程

从听觉效果上,合成语音能明显的感觉到音质的变化,而且合成的高兴和生气的部分句子能够感觉到说话人的情感,但合成的悲伤的句子情感表现得很不明显。由于没有修改韵律参数,此时合成语音在语气语调上并不自然,不像是在相应情感下所说的话。所以,单纯用频谱包络合成的情感语音表现力还是不够,而且当某些情感与基频的变化有很大关系时,音质和发声器官类参数是不足以表达出该情感特征的。

3.3 同时修改韵律参数和频谱包络合成情感语音的实验

通过上面的结果分析,我们结合了上述两个实验所利用的参数,即同时修改韵律参数和频谱包络来合成情感语音。对于中性情感的原始语音,我们首先按照3.1中所述的方法对其进行韵律的修改,然后再按3.2中的方法修改它的频谱包络,得到最终的合成情感语音。其过程如图4所示。

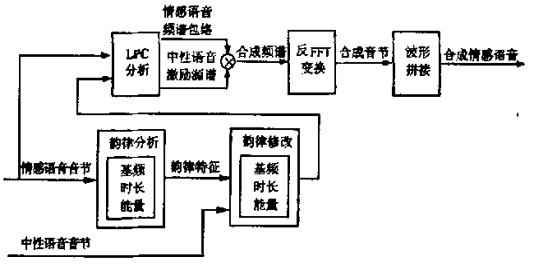


图4 同时修改韵律参数和频谱包络合成情感语音过程

图5所示为合成语音的波形和基频曲线与人工朗读语音的波形和基频曲线的对比。从图中可以看出,合成的语音的韵律和人工朗读语音是基本一致的,从而保证了合成语音自然的语气语调。

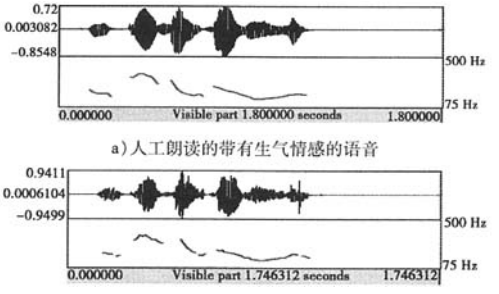
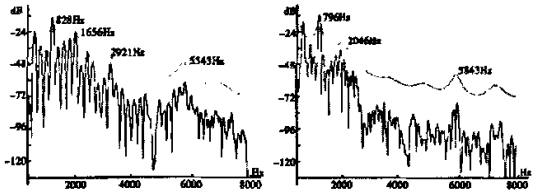
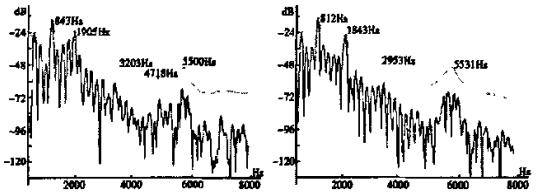


图5 句子"我不想再玩了!"带有生气情感时的语音的波形和基频曲线

图 6 为分别采用上述三种方法合成出的语音和人工朗读语音中一个对应音节的频谱和频谱包络的对比。从图中可以看出，修改频谱包络后的合成语音，语音的共振峰的位置，共振峰的带宽和频谱包络的形状都要比未修改频谱包络的合成语音更相似于人工朗读的语音。而从听觉效果上也表明，同时修改韵律参数和频谱包络的合成语音情感表现的更明显而且更自然。



a）人工朗读语音　　b）修改韵律参数的合成语音



c）修改频谱包络的合成语音　　d）修改韵律参数和频谱包络的合成语音

图 6　带有生气情感的句子“我不想再玩了！”中音节“再”的频谱和包络

4　合成效果的人工评价

为了比较合成语音的效果，也证明通过修改韵律和频谱包络来合成情感语音的有效性，本文采用两种方法对合成效果进行了人工评价。

4.1　情感识别率实验

我们分用上述三种方法合成了高兴、生气和悲伤三种情感的语音各 20 句。然后将用每种方法合成的 60 句语音打乱顺序，不标明其情感将它们混在一起。实验雇用了 10 名未经过任何训练的学生听这三组（各 60 句）语音，然后标出每句话的情感。实验要求强制选择，即听到一句话必须在三种情感中选出一种标出。之后，我们通过统计得到三种方法合成语音的平均情感识别率，如表 1 所示。

上表中，情感的识别率（R_X）与误识别率（E_{XY}）按照以下公式计算：

$$R_X = \frac{\text{情感 X 被识别为情感 X 的句子数}}{\text{情感 X 的句子总数}} \times 100\% \quad (1)$$

$$E_{XY} = \frac{\text{情感 X 被识别为情感 Y 的句子数}}{\text{情感 X 的句子总数}} \times 100\% \quad (2)$$

从表中可以看出，修改韵律和频谱包络的方法合成的语音的情感识别率最高，而只用韵律参数合成的结果要优于只用频谱包络。

表 1　三种模型合成情感语音的平均识别率及各情感之间的误识别率

方法	实际情感	识别情感		
		高兴	生气	悲伤
韵律参数	高兴	63%	29.5%	11.5%
	生气	20.5%	65.5%	4.5%
	悲伤	16.5%	5%	84%
频谱包络	高兴	61%	25.5%	17%
	生气	19.5%	44%	10%
	悲伤	19.5%	30.5%	73%
韵律参数 & 频谱包络	高兴	79.5%	7.5%	5.5%
	生气	14%	88%	3.5%
	悲伤	6.5%	4.5%	91%

4.2　情感效果主观对比实验

由于上述情感识别率的实验为强制选择，那么就存在合成的语音本身情感不明显，而实验者通过其特征（如基频高低或发声状态等）来判断其情感的因素。为了进一步对比三种方法合成的情感语音情感的明显性，我们又做了如下的主观对比实验。实验请五名学生来听同一种情感，同一句话，分别用三种方法合成的语音，并对其进行打分。打分的规则为，情感最明显的给 3 分，其次的给 2 分，最差的给 1 分。实验的句子仍为三种方法，合成高兴、生气和悲伤三种情感各 20 句。最后，根据打分我们统计出三种方法合成的情感语音的平均支持度如图 7 所示。图中平均支持度（S）的计算公式为：

$$S = \frac{\sum_{N}\sum_{n} mark_n}{N \times n} \quad (3)$$

其中 N 为句子数，n 为评测人数，$mark$ 为评测人的打分。

从主观对比实验的结果也证明了同时修改韵律参数和频谱包络的方法合成的语音情感效果最优；只修改韵律参数合成情感语音其次；而只修改频谱包络合成的情感语音效果最不明显。这个宏观上的趋势与情感识别率的实验结果是一致的。

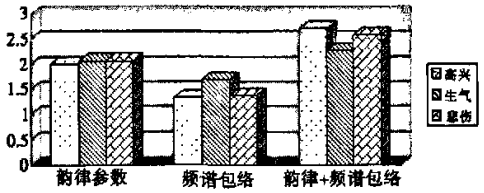


图 7　情感效果主观对比实验结果

5　结论

本文通过修改韵律参数，修改频谱包络和同时修改韵律参数和频谱包络三种方法，进行了以中性情感语音来合成特定情感语音的实验。情感识别率的实验结果表明，同时修改韵律和频谱包络的方法合成的语音情感识别率最高；而三种

方法合成的语音的人工对比实验的结果,进一步证实了这种方法合成的语音情感表达得较另两种方法更为明显。从理论上分析,通过修改频谱包络,我们修改了与音质和发声器官相关的一些参数。由此可以得到音质和韵律在刻画情感上都有至关重要的作用的结论,这与有些研究者研究结论是一致的。这组实验验证了我们的假设,即可以通过对频谱包络的修改,综合地调整与音质和发声器官相关的各种参数从而表现出不同的情感。这就意味着,在合成情感语音时,我们可以不必关心具体的每个音质或发声器官相关的参数的作用和影响,也不需通过人工的规则来为某种情感添加某种音质,或预先为某种情感设计相应的滤波器,而可以利用机器学习模型来自动为待合成的情感语音生成频谱包络。对于不同的应用或者合成不同的情感,我们将只需要用不同的训练语料来自动训练我们的频谱包络生成模型,而无需人工的修改和重新设计。这将使情感语音合成系统更加灵活。

参考文献

[1] M. Schroder. Emotional speech synthesis: A review. In: Proceedings of the 7th European Conference on Speech Communication and Technology Eurospeech 2001, Aalborg, 2001: 561 ~ 564.

[2] J. E. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, 1989.

[3] I. R. Murray, J. L. Arnott. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication. 1995, 16: 369 ~ 390.

[4] Iida A, Campbell N, Higuchi F, Yasumura M, A Corpus-based Speech Synthesis System with Emotion, Speech Communication, 2003, 40, 161 ~ 187.

[5] Iida A, Campbell N, A Speech Synthesis System with Emotion for Assisting Communication, In: Proceedings of ISCA Workshop (ITRW) on Speech and Emotion. Newcastle, Northern Ireland, 2000, 167 ~ 172.

[6] E. Rank and H. Pirker, "Generating emotional speech with a concatenative synthesizer", in Proceedings, ICSLP '98, Sydney, Australia, 1998, 3: 671 ~ 674.

[7] 陶建华, 许晓颖. 面向情感的语音合成系统. 第一届中国情感计算及智能交互会议论文集, 北京, 2003: 191 ~ 198.

[8] Nagasaki Y, Komatsu T: Can people perceive different emotions from a non-emotional voice by modifying its F0 and duration? In: Proceedings of Speech Prosody 2004. Nara, Japan (2004).

[9] T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech", IEEE ICMCS 99, June 1999.

[10] Gobl C., Bennett E., N'i Chasaide A.: Expressive synthesis: How crucial is voice quality? . In: Proceedings of IEEE Workshop on Speech Synthesis, Santa, Monica (2002).

[11] R. Cowie, R R. Cornelius. Describing the emotional states that are expressed in speech. Speech Communication. 2003, 40: 5 ~ 32.

[12] M. Schroder, R. Cowie. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. Eurospeech 2001.

[13] S. Hawkins, K. Stevens. Acoustic and Perceptual Correlates of the Non-nasal Nasal Distinction for Vowels. Journal of the Acoustical Society of America. 1985, 77: 1560 - 1575.

[14] J. Laver. The Phonetic Description of Voice Quality. Cambridge: CUP. 1980.

[15] D. Klatt, L. Klatt. Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. Journal of the Acoustical Society of America. 1990, 87: 820 - 857.

作者简介

邵艳秋, 女, 1970年生, 哈尔滨工业大学计算机学院语音处理研究室博士研究生, 主要研究方向为语音合成, 情感语音生成。