语音表现力的感知分析研究*

杨鸿武[1,2+]，蔡莲红[1]，蒋丹宁[1]

[1](清华大学计算机系,北京市 100084)

[2](西北师范大学物理与电子工程学院,甘肃省兰州市 730070)

摘 要: 为了增加 TTS 系统的表现力，本文利用“有效性”和“一致性”实验研究了 5 种情感和 3 种风格的语音特征在感知上的作用，并统计分析了 3 种风格语音的时域特征。结果表明，基频和时长等时域特征在风格和情感的表达中起到显著的作用，为合成出具有风格和情感表现的语句，需要优先对风格和情感的韵律特征建模。

关键词: 表现力语音合成；情感；风格；TTS

1. 引言

对于人机交互来说，理想的合成语音不仅要能将“字”读正确，而且要能够表达出“言外之意”，不仅仅是“说什么”，更重要的是“说话的时候意味着什么”，也就是能够根据情境表达出合适的语义信息。在这里，“情境”不仅指文本内容，还应包括交互内容、交互背景、交互对象、交互场景、语体风格等多方面的因素，所有这些都影响着言语的表达。比如，如果通过 TTS 陈述一些事实、道理（播报一般新闻），希望听者明知，则需要合成语音的语调平稳，声音清晰；要解说体育比赛，则合成语音应该表现出激情；而机场的信息广播系统，则需要语气平缓，字真意切。

语言的语音表现丰富多彩，我们认为，可以从三个方面研究语音表现：情感、风格和个性。情感是由心理－生理变化引起的情绪反应，这导致说话人的音质以及韵律特征的改变；风格是由说话者要表达的内容以及说话的情境所决定的某种表达模式；个性是由说话人的生理特征和发音习惯决定的，这反映在音质参数上。情感、风格和个性形成的语音表达是相对独立又不可分割的三个方面。通常，个性和风格是指相对稳定的、全局的语音表现；而情感重在表现变化，它更多是一种短时的、局部的语音现象。情感、风格和个性的相互联系表现为：人们可以借助风格传递情感；同一个人也可以不时变换说话风格与情感；不同人表达情感内涵的语音表征不尽相同。因此，如果合成语音能够合成不同人、不同风格的情感语音，则会使人机交互更为和谐。

---

* 本工作由国家自然科学基金（60433030,60418012）资助

* 杨鸿武, Email:yang-hw03@mails.tsinghua.edu.cn

在语音合成的研究中，人们关注可懂度和清晰度，追求自然度。目前，数据驱动的拼接式文语转换（TTS）系统虽然可以达到很高的可懂度和清晰度，但还缺乏自然度。因此，这一领域的研究重点开始转向提高合成语音的自然度，试图为合成语音增加情感、变换风格，以促进人机交互的发展[1]。拼接式语音合成系统参数的修改范围小，难以满足语音表现的需求。为了进一步提高拼接合成系统的表现力，一方面研究直接修改语音信号的算法。利用 PSOLA、HNM[2]、HMM[3]、STRAIGHT[4]等方法，对中性的语音信号进行修改以产生目标风格和情感的语音。信号处理的方法造成音质的下降，而且频域参数的修改难以达到理想的目标。另一方面，人们期望通过录制不同人的不同情感、风格的目标语音库来实现富有表现力的合成语音。这虽然能维持信号的质量，但是录制不同风格和情感的语音库的代价太高[1]。如果只录制小规模的情感和风格目标语音库，就不能保证找到合适的基元。

为了增加 TTS 系统的表现力，本文利用“有效性”和“一致性”实验研究了 5 种情感和 3 种风格的语音特征在感知上的作用，并统计分析了 3 种风格语音的时域特征。结果表明，基频、时长等时域特征在风格和情感的表达中起到显著的作用，因此，只对风格和情感语音的韵律特征建模，就可以在一定程度上合成出具有风格和情感表现的语句，这样可以减小录制语音库的代价。

2．情感与风格的表示

对于情感，目前有两种主要的表示方法，一是离散表示方法，二是维度表示方法[5]。离散表示方法将情感空间划分为若干个离散的范围，并用不同的名称表示每个范围的情感。目前普遍认为存在少数几种基本情感，大多数研究者认可的基本情感为害怕、愤怒、悲伤、高兴。维度表示方法在连续变化的维度上表示情感。最常见的维度为激发度（arousal）和评价度（evaluation）。激发度主要从生理的角度描述情感。高激发度的情感（如愤怒，高兴等），在语音的声学信号上表现为基频升高，能量加强和语速加快。与此相反，低激发度的情感（如悲伤），在声学上表现为基频降低，能量减弱和语速减慢。评价度描述的是情感在认知方面的属性。例如，高兴是正性的情感，而愤怒，悲伤则是负性的情感。本文采用了情感的离散表示方法，研究了愤怒、害怕、高兴、悲伤、惊讶五种基本情感以及中性语音。

不同于情感，对于语音的风格，目前在学术界尚无统一的定义。不同的研究者从不同的角度定义了风格。赵元任[6]将语调分为词调、中性语调和情感语调。Blado[7]等人将风格定义为正式-口语维，Abe[7]研究了小说，广告和百科全书风格，Higuchi[7]等人对比了紧张、愤怒和礼貌三种风格与未标注的风格, 而 Gibbon[7]等人将说话风格归类为朗读风格以及几种音高和时长上有差异的几种自然风格。我们认为，在言语工程中，风格是由说话者要表达的内容以及说话的情境所决定的某种表达模式。

Schlosberg[8]按照愉快-愤怒，紧张-松弛以及激活水平把情绪排列在一个倒立的锥体上。而风格的模式，我们也可借鉴 Schlosberg 的情绪模式，把风格也看成是由这三个维度构成。我们分析了三种风格：第一种是***松弛风格***，这种风格的语音在愉快-愤怒维上

表现为中性，在松弛-紧张维上表现为松弛，激动水平低；第二种是***中性风格***，这种风格在三个维度上均表现为中性；第三种是***活跃风格***，这种风格的语音在三个维上的变化都较大。

3．语音表现的声学特征分析

语音最主要的感知特征包括音高、音强、音长和音质。从语音的产生模型来看，语音的声学参数分为两类：声源参数和声道参数。这些参数的静态和动态特征与语音的生理特征以及心理/社会特征相关。声源的声学特征包括基频、时长等时域特征，声道的声学特征包括共振峰等频域特征。

为了考察时域特征和频域特征对不同情感和风格语音的感知作用，我们设计了“一致性”和“有效性”的感知实验。所谓“一致性”，是指语句在声学上的表现应与它的文本内容和出现的具体情境相一致。例如，好消息应当高兴地表现，坏消息应当悲伤地表现。图 1 中（a）给出了一个“一致性”测试的例子。“有效性”是指当语句的文字内容不具情感倾向时，声学表现出的情感信息是否被听者感知。例如，“我已经知道自己的考试结果了”，若高兴地表现，则传递出“考的还不错”的信息；若悲伤地表现，则很有可能是考的不理想。若语句被判断为与文字内容一致，或者有效地传递出了情感信息，则认为语句具有情感表现。图 1 中（b）给出了一个“有效性”测试的例子。

(a) ***“一致性”***测试示例文本：
语句：如今我的生活变得十分艰难。
问题：请问该语句的表现方式是否与文本内容一致？
选项：A. 一致　B. 不大一致　C. 很不一致

(b) ***“有效性”***测试示例文本：
语句：我已经知道自己的考试结果了。
问题：考试结果最有可能的是下面哪一个？
选项：　A. 考的不好　B. 我不知道　C. 考的很好

图 1 “一致性”测试和“有效性”测试的例子

3．1 情感和风格语音的一致性和有效性测试

本文对愤怒、害怕、高兴、悲伤、惊讶 5 种基本情感和松弛、活跃两种风格进行了测试。为每种情感和风格设计了 4 个语句（2 句具有情感倾向，2 句中性）。由一位女性播音员录制。另外，对松弛、中性和活跃三种风格又各录制了 100 句语句。所有的数据均以 16 位 16k Hz 采样率的 wav 文件格式保存，并用 Visual Speech 手工标注。

我们分析了不同语句基频、时长和能量的差异。采用拷贝合成的方法，利用 STRAIGHT 算法[3]，交换中性语句以及情感或风格语句的韵律参数，产生具有情感/风格韵律特征或中性韵律特征的合成语音，然后进行测听试验，结果如表 1 所示。表 1 列出了被判断为具有情感或风格表现的平均比率。被测听的语句是录制的情感或风格语句和中性语句，以及合成的带有情感或风格韵律特征的中性语句和带有中性韵律特征的情感或风格语句。由表 1 可见，至少 89%的录制的情感或风格语句被判断为具有情感或风格

表现（列 2），而录制的中性语句则很少被判断为具有情感或风格表现。因此，录音数据在总体上“一致性”较好。当中性语句改换为情感或风格韵律时，大部分情况下被判断为具有情感或风格表现，“有效性”较好（列 4）。但将情感或风格语句改换为中性韵律时，合成语句被判断为具有情感或风格表现的比率稍高（列 5），这说明情感或风格表达中除韵律参数外，还受频谱参数的影响，但是，与频谱特征相比较，基频、时长等时域特征起到更显著的作用。

表 1：自然语句以及合成语句被判断为具有情感/风格表现的平均比率（%）.

被判断为具有情感/风格表现的比率（%）	录制的情感/风格语句	录制的中性语句	录制的中性语句+情感/风格韵律	录制的情感/风格语句+中性韵律
愤怒	99	2	87	30
害怕	89	1	76	5
高兴	91	1	68	7
悲伤	93	16	66	49
惊讶	90	6	80	13
松弛	92	4	82	18
活跃	93	6	86	14

3．2 风格语音的时域特征分析

对三种风格的语句，计算了以音节为单位的语速和语速的变化。在语速的计算中，没有考虑停顿。表 2 显示了实验结果。由表 2 可以看出，这三种风格的语音在语速上明显不同，松弛风格的语速最慢，而中性风格的语音语速最快。此外，在这三种风格中，语速的变化也不相同，松弛风格中的变化最小，而活跃风格中的变化最大，这是松弛风格多为短句，而中性和活跃风格多为段落的原因。中性风格中的语速变化少于活跃风格的，这是由于新闻需要用平缓的调来说明事实，因而语速变化的次数少；而在活跃风格中，则利用语速的变化来达到积极的效果。

表 2：三种不同风格的时长信息

	松弛	中性	活跃
消除停顿的语速（ms/音节）	255	172	202
语速的变化（%）	17	6	37

我们也分析了三种风格的基频均值、最大值、最小值和基频变化率等特征。图 2 给出了三种风格的基频范围变化的统计结果。图 2 中，左边是松弛风格的基频范围变化率，中间是中性风格，右边是活跃风格。由图 2 可以看出，不同风格的基频变化范围不同。松弛风格和活跃风格的基频的变化范围明显不同于中性风格的基频特征，对于松弛的风格，基频其变化范围窄，这说明在这种风格中，语调平缓。而在活跃风格中，基频的变化范围大，说明在这种风格中，语调的变化丰富。此外，这两种风格的基频平均值也不

同于中性风格的。

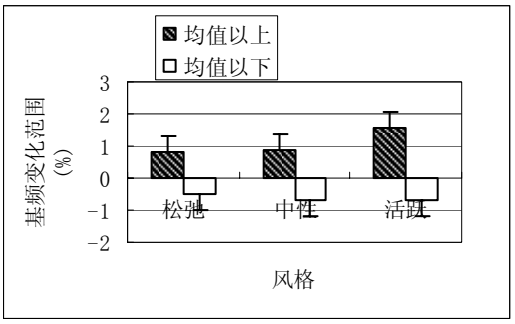

图 2 松弛、中性和活跃三种风格的基频变化范围

4. 结论

为了实现高表现力的语音合成，本文用一致性和有效性分析语言与表达的关系，并研究了 3 种风格的语音的声学表现。当合成语句不包含风格和情感时，可以通过风格和情感韵律在一定程度上表现出风格和情感，满足交互的需要。从这些研究中，我们发现在风格和情感的表达中，与频谱特征相比较，基频、时长等时域特征在风格和情感的表达中起到更显著的作用，为合成出具有风格和情感表现的语句，需要优先对风格和情感的韵律特征建模。

参考文献

[1] B. Murtaza, N. Shrikanth, and J. Lewis. Synthesizing expressive speech: overview, challenges, and Open Questions, In N. Shrikanth and A. Abeer. Text to speech synthesis: new paradigms and advances, Prentice Hall. 2004, P175-201

[2] Y. Stylianou, Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification, PhD. Thesis, 1997. École Nationale des Télécommunications, Paris.

[3] T. Masuko, K. Tokuda, T. Kobayashi, and I. Satoshi. Voice characteristics conversion for HMM-based speech synthesis system. Proc. ICASSP. Apr, 1997. Munich, Germany, P1611-1614.

[4] H. Kawahara, I. Masuda Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine. 2001. vol. 18(1), P32-80.

[6] Yuan-Ren Zhao. Tone and Intonation in Chinese. Bulletin national Institute of History and Philology Academia Sinica. 1933. vol. 4(2), P121-134.

[7] J. Terken, Variability and Speaking Style in Speech Synthesis, In E. Keller, B. Bailly, A. Monaghan, J. Terken and M. Harkvale. Improvement in Speech Synthesis. JOHN WELEY & SONS. 2002, P199-203

[8] H. Schlosberg. Tree dimensions of emotion. Psychological Review. Mar., 1954. vol. 1(2), P81-88.

Perceptual Analysis on Expressive Speech

YANG Hong-Wu[1,2+], CAI Lian-Hong[1], JIANG Dan-Ning[1]

[1](Department of Compute Science & Technology, Tsinghua University, Beijing 100084, China)

[2](College of Physics and Electronics Engineer, Northwest Normal University, Lanzhou, 730070, China)

+ Corresponding author: Phn: +86-10-6277-1587, E-mail: yang-hw01@mails.tsinghua.edu.cn

Key words: expressive speech synthesis; speaking style; emotion; text-to-speech

Abstract: Research efforts in the field of TTS have placed emphasis on the naturalness in synthetic speech to facilitate its various applications in Human-Computer Interaction (HCI). For better expressiveness of synthetic speech, the relation between the speech and expression is analyzed through both appropriateness and efficiency tests. Appropriateness means that the expression should be relevant for the verbal content of speech. Efficiency measures the ability of the expression to convey meanings that are literally implicit. 5 emotions—anger, fear, happiness, sadness and surprise, along with 2 styles—lax and active were studied. Then, to isolate the influence of prosodic features on expressiveness, copy-synthesis method was applied with the STRAIGHT algorism. So synthetic utterances were obtained which contain expressive prosody and neutral segment or neutral prosody and expressive segment. Finally, 25 subjects evaluated these natural utterances and synthetic speeches. The results show that over 89% recorded expressive utterances were correctly perceived as appropriate expressiveness. More than 85% synthetic utterances using expressive prosody and neutral utterances were perceived with to be expressive. However, only 13% of synthesized utterances using expressive utterances and neutral prosody were perceived with to be expressive. Another 100 sentences was made for lax, neural and active style by a female voice. An analysis of average pitch, maximal pitch, minimal pitch and various rate of pitch were made in 3 styles. The results show that the various range of pitch of the lax and active styles were different from those of the neutral style with narrower various range for lax style, which means a moderate change of intonation, and a wider various range for active style, which means a dramatic change of intonation. A calculation of the speed of speaking and differences in the speed of three styles was made on the basis of syllable as a unit. The results of the experiment show that speed varies from style to style, with the lax being the lowest and the neutral being the highest. The results also show that the differences in the speed vary from style to style, too, with the neutral being the lowest and the active being the highest. From the results of the studies on emotions and speaking styles, we have arrived at the conclusion that timing domain parameters such as pitch, energy and duration play a dominant role in the realization of speaking style and emotion, with modeling prosody being more important than modeling spectrum for expressive TTS system.