语音的表现力及其声学特征分析*

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

(电话: 010-62771587; E-Mail:yang-hw03@mails.tsinghua.edu.cn)

摘要：为了增加 TTS 系统的表现力，本文分析了语音的表现，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，情感，风格，语音合成，TTS

1.引言

语音不但能表达语言信息，而且也能表达情感、意向、态度和说话者特性等信息，因此，语音是人们相互交流的最重要、最自然、最便捷和最有用的手段，也是人机交互中最为方便直接的方式之一。随着人机交互技术的发展，语音技术也广泛应用到人机交互中。语音识别技术使得计算机能够 “ 听懂 ” 人的语音，而语音合成技术使计算机能够“ 说出 ” 人类的语言。目前，语音合成技术获得了飞速发展，尤其是可以将任意文字转换成语音信号的文语转换系统（TTS），其输出语音的音质已经基本接近真人的发音，在人机对话、远程语音信息服务、机器阅读、电信、娱乐方面得到了广泛的应用。但是，目前国内外的大多数语音合成系统只能以某种朗读风格将书面语言转换成口语输出，缺乏不同年龄、性别特征及语气、语速的表现，更不用说赋予个人的感情色彩和语言风格。随着信息社会的需求发展，人们试图为合成语音增加情感、变换风格，以提高合成语音的自然度，促进人机交互的发展[1]。这不仅涉及到计算机语言生成以及人类大脑的高级神经活动，也涉及到语音的表现力的分析。

目前的语音合成系统多采用数据驱动的拼接语音合成。这种方法的合成语音可达到很高的可懂度，但参数的修改范围小，难以满足语音表现力的需求。为了进一步提高拼接合成系统的表现力，一方面研究利用信号处理方法，对中性的语音信号直接进行修改以产生目标风格和情感的语音；另一方面，通过录制不同人的不同情感、风格的大语音库来实现富有表现力语音合成。信号处理的方法造成音质的下降，频域参数的修改也难以达到理想的目标。而录制不同大语音库的代价太高[1]。

为了增加 TTS 系统的表现力，需要对富有表现力语音的声学特征进行分析。本文利用“ 有效性 ”和“ 一致性 ”实验[1]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，这样可以在拼接语音合成系统中减小录制语音库的代价。

2．语音的表现

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 ”，能够根据情境表达出合适的语义信息。也就是说，理想的合成语音应该具有非常丰富的表现力。在这里，语音表达的“ 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。

* 本工作由国家自然科学基金（60433030,60418012）资助

语音的表现可分为表事、表意和表情三个层次。通常“中性语调”的语音可以用来说明事实；韵律丰富的语音可以隐含意向（称之为“逻辑语调”）；而富有表现力的语音则可以表达更丰富的情感（称之为“情感语调”）。富有表现力的语音包括了不寻常的重音、音高的整体变化、讲话速度的变化等。自然对话中，同一句话，文字虽然相近，如果语音表现不同，则表达的意向和情感也不同甚至相反。

语音表现的三个层次可以从情感、风格和个性这三个方面来研究。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。

对于情感，目前有两种主要的表示方法，一是范畴表示方法，二是维度表示方法[2]。范畴表示方法利用不同的词刻画某一类情感。目前大多数研究者认为存在害怕、愤怒、悲伤、高兴等少数几种基本情感。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；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。

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风格[3]。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。这种表达模式可以借鉴情感的表示方法来表示。我们根据 Schlosberg[4]的三维情绪模型，按照语音表达的愉快-愤怒，紧张-松弛以及激活水平来表示语音风格。例如，信息播报风格的语音，其在愉快-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，我们定义为松弛的风格；新闻播报在三个维度上均表现为中性，我们定义为中性的风格；而体育解说在这三个维上的变化都较大，我们定义为活跃的风格。

语音的个性特征分为生理特征和心理/社会特征[5]，前者与音质及情感有关，后者与说话风格有关。生理特征随性别、年龄、体重等变化，主要由说话者的声道形状和尺寸、声带的长度、质量和弹性、肺容量等决定。声道形状和尺寸影响共振，即共振峰频率和带宽，声带和肺容量分别影响基频和响度。心理/社会特征是在人的社会生活动中逐渐形成的，它们通常显示语言学、语义、情感等方面的差异，主要对韵律特征有贡献。

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。

(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

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 的上半部分给出了一个“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 的下半部分给出了一个“ 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[6]，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，但是录制的表现悲伤的中性语句被判断为悲伤的比率较高（列 3）。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。

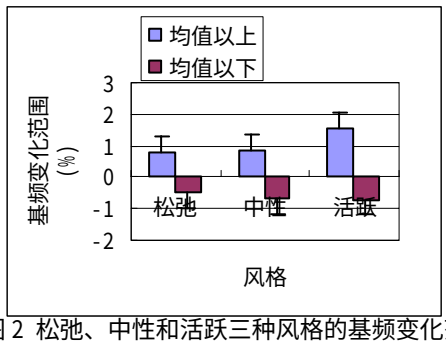

图 2 松弛、中性和活跃三种风格的基频变化范围

4．结论

为了实现高表现力的语音合成，本文探讨了语音的表现，并用一致性和有效性分析了语言与表达的关系，研究了 3 种风格的语音的声学表现。从这些研究中，我们认为语音的表现力包括情感、风格和个性。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。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。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, pp 175-201, Prentice Hall.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 2001, 18(1): pp. 32-80.

[3] 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, pp 199-203, JOHN WELEY & SONS

[4] H. Schlosberg. Tree dimensions of emotion. Psychological Review, 61(2):81-88, Mar. 1954.

[5] H. Kuwabara and Y. Sgisaka (1995), “Acoustic characteristics of speaker individuality: Control and conversion,” Speech Communication, Vol. 16, pp. 165-173.

[6] 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.

Analysis on Expressivity and Acoustic Correlation of Speech

Yang Hongwu[1),2)], Cai Lianhong[1)] Jiang Danning[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)

(Tel: +86-10-6277-1587; E-mail: yang-hw01@mails.tsinghua.edu.cn)

Abstract: For better expressiveness of synthesized speech, the relation between the speech and expression is analyzed through the appropriateness test and efficiency test. The acoustic features are studied for different speaking styles and emotions. From the results, we can draw a conclusion that timing domain parameters such as pitch, energy and duration play a dominant role in the expression of speaking style and emotion, thus modeling prosody is more important than modeling spectrum for expressive TTS system.

Keywords: expressivity; emotion; speaking style; speech synthesis; Text-to-Speech