朗读语流中的喘息音段分析研究

袁楚 李爱军

中国社会科学院语言所

提要：在朗读语流和口语语流中，经常会出现一些副语言学现象 （paralinguistic information）的出现。常见的副语言学现象包括喘息、拖音、句间沉默、笑声、哭声等等。本文以常见的喘息段（breath segment）这种副语言学现象为例，通过韵律结构和情绪唤醒度（valence）、情绪活跃度（activation）等参数确定喘息段声学特性，包括强度（intensity）和时长（duration）对朗读语料中的喘息段进行了初步的统计分析。

由此我们得到了朗读语料时喘息段在不同情绪唤醒度和活跃度状态下的一些普遍性规律，并对其进行初步量化，形成了可以应用于实际合成的规则。所提取的规则主要涵盖了喘息段出现的位置、时长以及强度等三个特征。我们利用合成系统合成的部分语料，将我们获得的规则运用到了合成实验中，并对合成语料进行了听辨实验，得到比较理想的结果。

关键词：喘息段　自然朗读语音　情绪　唤醒度　活跃度

一 引言

目前的语音合成和语音识别的相关研究中，研究者大多避开了自然口语部分的某些特征，在包括录音、标注、训练，以及合成和识别的全过程中，严格控制发音人的说话方式，尽量使他们“发音规范”，尽力避免口语信息的影响。随着应用环境的快速成熟，以及 2008 年北京奥运会的临近，对于口语中的语音信号处理已经不再局限于朗读语音的识别和合成。研究者们开始更多的关注口语中丰富的副语言学现象，研究如何使用这些以前大家并未重视的信息来加强合成语音的自然度，以及如何使用这些附加的信息表达一定的情感或者态度。

在情感语音研究方面，前人已经作出了很多有益的尝试。2000年在爱尔兰召开ISCA的Workshop on Speech and Emotion国际会议第一次把致力于情感与语音研究的学者聚集在一起。Lida在《A Speech Synthesis System for Assisting Communication》 中对三种情感中的每一种(愤怒、高兴和悲伤) 同一说话人的一段连续语音被录制下来制成某种情感的语音单元选择数据库，为了合成某一种情感，只有符合该情感的语音数据库才会被选中，提取出里面的语音单元。在这种方法下合成的情感语音有50－80％的情感识别率。刘亚斌、李爱军在2002年第16期中文信息学报上发表题为《朗读语料与自然口语的差异分析》的论文，提出口语中的丰富信息对口语自然度有相当大的影响，特别是一些出现较多的口语副语言学现象，更是需要关注的重点。吴禀雅等人通过分析不同文本类型的语体色彩和感情色彩，也较好地改进了合成语音的自然度。中科院自动化所的陶建华和北京师范大学的许晓颖联合发表了题为《汉语情感系统中情感划分的研究》，在对现代汉语词库中的形容词库和动词词库的进行义类标注的基础上，对现代汉语的基于心理感受的情感系统和基于表现力的情感系统进行了梳理和分类。另外，近几年来日本ATR也在进行情绪语音识别和合成的研究和实践，以Nick Campbell为首的研究团队在情绪语音合成和识别方面获得了不错的成果。2004年的ICSLP大会上，Nick Campbell发表了题为 Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation 的论文，研究了用某一个固定的语气词表达各种情绪时的方法和合成中的可能性。2003年，Fujisaki教授提出将人们所发出的语音信息分为三个层次，包括语言学信息、副语言学信息和非语言学信息，这三者层次有时候并不是特别的清楚。其中，语言学信息能用语言文字记录和表达出来。副语言学信息不能用语言文字记录，但是可以由发音人控制其产生并以此来表达某种情绪、态度或者说话风格。而非语言学现象则是由说话人的年龄、性别以及生理结构等等决定的，无法由说话人有意识的进行控制。这个

本文得到中国社会科学院重点学科“语音与自然话语处理”资助。

理论从产生的原理和实现的方式上对语言学信息、副语言学信息和非语言学信息作出了界定。

本文试图从原本很少被人注意的副语言学信息入手对其在语流中的作用进行了一些初步的研究。常见的副语言学信息包括喘息、拖音、句间沉默、笑声、哭声等等，都能传达说话人的情感和情绪。本文以常见的喘息这种非语言信息为例，考察喘息段的出现与韵律结构和情感表达的强度、情感的正负等参数的关系，从而确定朗读合成语音中如何通过插入具有不同声学特性的喘息段，增加情感的表现力，通过合成实验听辨结果证明，有助于增加自然语音合成中的丰富情感信息。

二　语料介绍

2.1 研究的对象

在这里我们没有区分呼气和吸气，这是因为在某些人的说话习惯中，喘息时呼气的表现比较明显，而另外一些人则会将吸气表现的比较明显。因此我们在对喘息现象进行定性分析的时候，不再考虑这两者的区别。如图1所示，我们可以在两个语图上明显看到两条虚线之间就是出现在语流中的喘息音段。

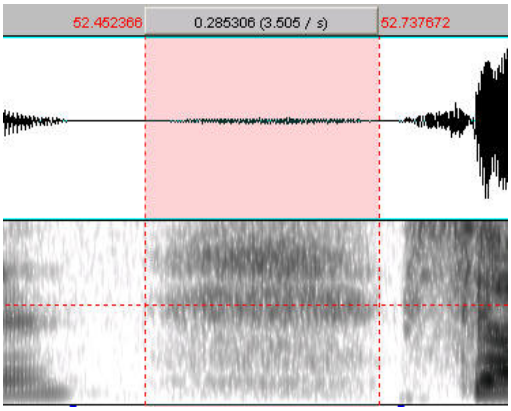

图1，朗读状态

我们这里的主要研究对象，就是这些在语流中出现的喘息段。大家可以看到，喘息段的能量集中区都出现在高频部分，低频的能量很少。而且在喘息段两端都有一小段的空白，这是由于在开始喘息和喘息结束准备发声的时候，从生理上来说发音器官会有一个短暂的准备阶段，因此都会有一小段的停顿。这一特征在我们进行合成时也必须注意，在插入喘息段时，和前后发声段要保持一定的距离。

2.2 语料和标注

本文中选取的语料来自于我们自建的CASS-EXP 自然口语情绪语料库。这个语料库分为两部分，朗读部分和口语部分。我们这次选取的是朗读部分，由专业的演员朗读一些少儿故事，用带有感情的语气和中性语气分别朗读一遍，然后将其切成片断。我们从其中挑选了 1 小时左右的语料及其标注信息，包含有正面情绪和负面情绪以及中性语气的部分。然后从这部分信息中分别统计出喘息的各种状态，重点在考察喘息段的时长、强度和出现位置。最后将统计结果运用到一些样本句子当中并进行听辨实验，并对实验结果进行分析，试图给出一些可观察到的结果。

我们使用SAMPA-C和C-ToBI标注规范对朗读文本和自然口语语料进行了标注，标注内容除了音段和韵律标注外，还标注了包括喘息段在时间轴上的起始位置和结束位置以及时长。由于喘息本身并不携带可供判断的情绪信息，因此我们标注的是紧随每一个喘息段之后的韵律短语的情绪特征，这个情绪特征用来作为对喘息段进行分类的标准而被标记成为某一喘息段的特性。

在对情绪特征进行标注时，我们采用了目前最广为接受的二维模式，其中一维是效价或者愉悦度，其理论基础是正负情绪的分离激活，另外一维是唤醒度或者激活度，指与情感状态相联系的机体能量激活的程度。唤醒的作用是调动机体的机能，为行动做准备。

对愉悦度的标注采用三级标准。正面状态，主要是高兴和友好等情绪或者态度，标记为 1；中性状态，包括惊奇的情绪状态和中性的态度，以及叙述性语言，我们认为这种状态是没有愉悦度表现的，标记为 0；负面状态，包括恐惧、悲伤、愤怒等等情绪及类似的态度，标记为-1。

唤醒度的标注采用三级标准。高昂的，或者是兴奋的，主要是朗读者情绪比较高昂，喘息的动作幅度较大，时间相对比较短，标记为 1；中性的，喘息动作平稳，没有大起大落，标记为 0；低落的，喘息的时间相对较长，有时强度较大，但不如情绪兴奋时猛烈，一般来说喘息的速度比较慢，标记为-1。

当某个喘息的愉悦度和唤醒度的标注结果都为 0 时，我们基本可以断定，这个喘息是自然的生理现象，不携带情绪信息。

另外，我们将语流的韵律标注结果和喘息的结果对应起来，标注关于韵律和喘息对应的信息，以及标注每一个喘息在语流中所处的位置是否为正常停顿。在对喘息位置进行标注时，我们使用了 3 级位置系统，分别是 3（主要韵律短语边界）、2（次要韵律短语边界）和 1（韵律词边界）。

三 喘息音段出现的位置分析

根据我们的标注结果，在有文本作为依据且文本内容相同的情况下，具有愉悦度表现的朗读比中性状态的朗读出现的喘息现象要多出 50%左右。在这 9 段文本中，带有情绪的朗读中出现喘息 200 次，出现在正常停顿处的喘息次数为 199 次，占总次数的 99.5%。中性朗读中出现喘息为 133 次，全部出现在正常停顿处，如图 2 所示。

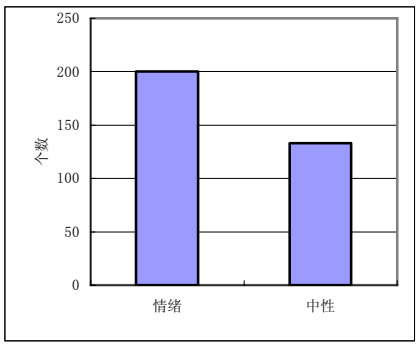

图 2 情绪状态和中性状态下的喘息次数比较

在朗读状态下，不管是带有情绪的朗读还是中性的朗读，喘息出现在 3 边界，也就是主要韵律短语边界上的数量都是最多的。出现在 2 边界，也就是次要韵律短语边界的数量其次，而出现在 1 边界也就是韵律词边界位置的数量是最少的，如图 3 所示。

在带有情绪的朗读和中性的朗读状态下，除了喘息段出现的次数有所区别外，其出现的位置分布并没有很大的不同。其中，带有情绪的朗读状态下，出现在主要韵律短语边界的喘息有 171 次，出现在次要韵律短语边界的喘息有 25 次，出现在韵律词边界的喘息有 4 次。在中性状态下，出现在主要韵律短语边界的喘息有 122 次。出现在次要韵律短语边界上的喘息有 8 次，出现在韵律词边界上的喘息有 1 次。其所占的百分比如表 1 所示。

总的来说，在朗读语篇中，不管是带有情绪的朗读，还是中性的朗读，喘息大都会出现在两个韵律短语之间，从文本分析上看，也就是出现在两个句子之间。而且在朗读语料中，喘息段的分布频率和朗读者的状态无关，也就是说，不管朗读者是带有情绪的朗读还是中性朗读，其喘息段的分布都是类似的，即 3 边界最多，2 边界其次，1 边界最少。

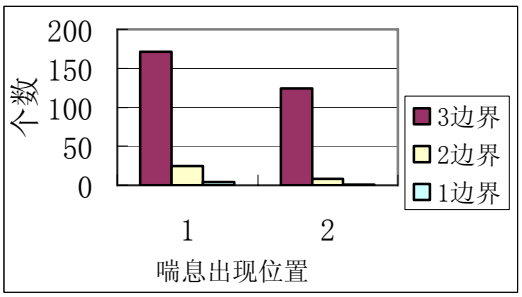

图 3，其中 1 为带情绪的朗读，2 为中性朗读

表 1 各种情绪状态次数及所占百分比

	情绪状态喘息次数	百分比	中性状态喘息次数	百分比
3边界	171	85.5%	124	93.2%
2边界	25	12.5%	8	6%
1边界	4	2%	1	0.8%

四 喘息音段时长的相关量分析

我们将喘息的时长和强度信息以及愉悦度和唤醒度数据导入到 SPSS 中进行多元分析可以看到，在对愉悦度和喘息段时长的关系分析中，愉悦度的三个状态对喘息段时长没有区分度，而且平均时长变化也非常的小(P=0.063>0.05)，见图 4。

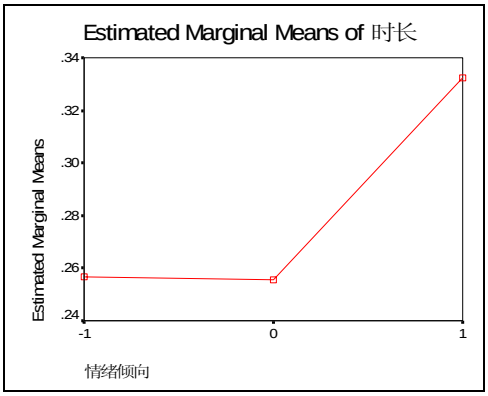

图 4 时长与愉悦度的关系分析

另一方面，唤醒度对喘息段时长有明显的影响(P=0000<0.05)。我们看到 SPSS 的分析结果显示（图 5），当唤醒度为 0 和 1 的时候，时长的区分度不大，而当唤醒度为-1 的时候，和前面两种状态有显著区别。当情绪比较低落时，喘息段时长会与情绪正常或

者比较高昂时区别较大。从表 2 中我们也可以看出，唤醒度对时长的 P 值为 0.000，也就是说唤醒度对时长的影响非常显著。

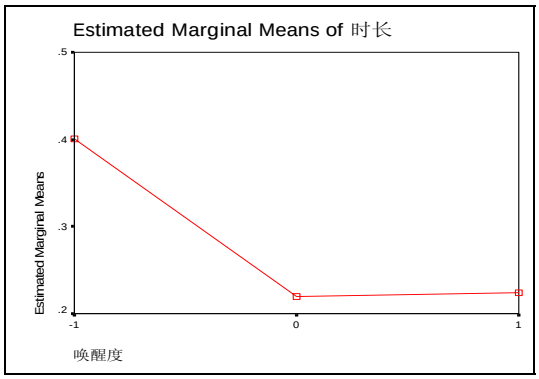

图 5 时长和唤醒度的关系分析

表2 时长对愉悦度和唤醒度的多因素分析

Source	Dependent Variable	Mean Square	F	Sig.
愉悦度	时长	3.534E-02	2.801	.063
唤醒度	时长	.118	9.344	.000

由表 2 我们可以看出，愉悦度对时长的影响不显著，而唤醒度对时长的影响比较显著。

除了喘息时长，我们还计算了每两个喘息之间的间隔时长以及分布情况，得到的结果是，在全部 319 个喘息间隔中，有 304 个喘息间隔时间短于 10 秒，其余 15 个大于 10 秒的喘息间隔，都是包含了口误的间隔，也就是说不是完整的韵律短语。由此我们可以基本确定，当朗读状态正常，即没有口误的时候，两个喘息之间的间隔时间不会超过 10 秒。

五 喘息音段强度的相关量分析

在朗读性语料中，喘息段的另一个重要特征就是强度。我们以喘息段开始和结束的时间为边界，用 praat 中的计算平均强度命令提取出每一段的强度参数，然后将这部分参数导入到 SPSS 中进行愉悦度和唤醒度的多元分析。表 3 和表 4 是我们分别对强度和愉悦度以及唤醒度进行分析得到的结果。

之后，我们还计算了每一个喘息段的强度及紧随该喘息段的韵律短语强度，并用 SPSS 对喘息段强度和韵律短语强度的比值受愉悦度和唤醒度的影响程度进行分析，得到的结果是，唤醒度对后接的韵律短语强度有显著影响，但是愉悦度以及唤醒度和愉悦度的交互作用对后接的韵律短语强度没有显著影响。

表3 强度与愉悦度的关系

愉悦度	N	Subset	
		1	2
0	155	37.8143	
1	29		41.9793
-1	16		43.8315
Sig.		1.000	.202

表4 强度与唤醒度的关系

唤醒度	N	Subset		
		1	2	3
0	120	36.5159		
-1	21		39.5437	
1	59			43.5185
Sig.		1.000	1.000	1.000

表5 强度对愉悦度和唤醒度的多因素分析

Source	Dependent Variable	F	Sig.
愉悦度	强度	.544	.581
唤醒度	强度	10.313	.000

表6 愉悦度唤醒度及双重作用对后接韵律短语强度影响。

Source	Sig.
Corrected Model	.000
Intercept	.000
唤醒度	.022
愉悦度	.913
唤醒度*愉悦度	.609

表 7 给出的是当唤醒度处于三种状态下强度比值的均值范围。从这里我们能看出，当唤醒度为 0 时，其强度比值的均值范围明显低于唤醒度为非 0 时的均值范围。

表7 三种唤醒度状态下的强度比值均值范围

唤醒度	Mean	95% Confidence Interval	
		Upper Bound	Lower Bound
-1.00	0.634	0.682	0.592
0.00	0.558	0.573	0.544
1.00	0.646	0.674	0.619

六 语音合成中喘息段的应用规则

从以上的数据我们可以得出一些关于朗读语料中喘息信息的应用规则。这些针对朗读语料的规则不完全是量化的，因为人的生理活动在很多情况下并不能用量化的数字来模拟，正因为如此，在合成时我们就可以采用随机的方式来设定一些细节。而针对喘息音段我们可以单独建立一个喘息声的语料库，合成时从中选取合适的喘息声插入合成语料中。另外，因为合成朗读语料时会有文本标准，因此可以通过文本来获得一些喘息的位置信息。

A 首先，在每个 3 边界也就是主要韵律短语的边界上，从文本上看则是两个较长的句群之间，一般来说会有一个喘息，喘息时间较长，一般在 0.5 秒左右。

B 两个喘息之间的间隔时间一般不会超过 10 秒。

C 在对喘息段进行选取时，当愉悦度或者唤醒度为 0 时，选取强度较弱的喘息段，当愉悦度或者唤醒度为-1 时，选取强度适中的喘息段，当愉悦度或者唤醒度为 1 时，选取强度最强的喘息段。

D 当唤醒度不为 0 的时候，喘息段的强度是后接韵律段强度的 0.6 倍到 0.7 倍，当唤醒度为 0 时，喘息段的强度是后接韵律段强度的 0.5 倍到 0.6 倍。具体的数值在限定的范围内随机产生。

最后，喘息并不是朗读时表达情绪的唯一手段，因此我们在合成的朗读语料中加入喘息，能让合成语音显得更加自然，但是并不能完全表达某种情绪。喘息要和别的条件配合才能表达出情绪的变化。

七 合成及听辨实验

7.1 合成实验语料

为了检验规则是否有效，我们设计了一个小型的听辨实验。我们首先从之前使用的语料中挑出了两段朗读语料，从统计使用的朗读语料中挑选了一段语料，我们称为 1 号语料，去掉其中的喘息段部分作为 2 号语料。之后将转写的文本使用合成系统合成得到 3 号语料（以下简称合成语料），然后根据我们的结论，将不同长度和强度的喘息插入到这些合成语料当中，得到 4 号语料（以下简称实验语料）。我们的听辨实验受试人一共有 10 名。

7.2 听辨实验

听辨的过程，第一轮将这 4 段语料分别在相同的位置切分成 5 段，一共 20 段。每个受试者听辨被切开分别两两对应的 1 号和 2 号语料以及 3 号和 4 号语料。其中播放次数由受试者自己控制。但一般来说不建议重复播放 3 次以上。

第一轮听辨朗读语料时，受试者普遍觉得1号和2号语料的差别不大，90%的受试者听不出这两者的区别。在仔细比较3号和4号语料之后，有40％的受试者能听出两者的差别，觉得添加喘息之后给人的感觉更加真实。1号和2号语料虽然被放在一起进行比较，听辨的正确率稍低，只有40％（20组次，总数50组次）能被分辨出来。3号和4号语料各个片断的听辨成功率却相当的高，达到了86％（43组次）。结果见表8。我们对听辨者只要求给出“是”或者“否”的答案，“是”表示两者有区别，“否”表示两者无区别。在表9当中，0代表答案为否，1代表答案为是。

表 8 第一轮听辨实验结果

听辨人编号	1 和 2 (各 5 个片断)	3 和 4 (各 5 个片断)
1	2/5	5/5
2	5/5	5/5
3	5/5	5/5
4	2/5	5/5
5	1/5	4/5
6	1/5	5/5
7	0/5	4/5
8	1/5	4/5
9	2/5	5/5
10	1/5	5/5
Total	20/50	43/50

第二轮听辨对话语料时，过程相对简单，只需要受试者判断两组对话中哪一组为加入了喘息之后的合成语料，如果能正确判断，再要求受试者回答喘息对自然度的提高和情绪的表达有无影响。第二轮对话语料的听辨效果比第一轮的朗读语料相对来说要好一点，两段对话的识别率分别达到了 50%和 60%。不过对于喘息表意的听辨效果依然不尽如人意，两段

对话都只有 20%的辨识成功率。在调整了喘息段的强度和后接韵律段的强度的比例关系后，我们获得了不错的听辨结果。由此证明，喘息段的强度会对后接韵律段的强度造成一定的影响，甚至会影响到韵律段中单个韵律词的强度变化。因此在插入韵律段时，如果能同时对后接韵律段的强度进行调整，合成的效果会更加理想。

要说明的是，我们并没有希望能够单纯通过副语言学信息就能全面地表达情绪和态度的倾向，而只是期望能够通过对副语言学信息的研究，提升情绪语音合成的自然度，从而获得更加自然的语音合成效果。

六 结论及展望

本文通过对朗读语料、对话语料以及相对应的合成语料进行统计和比较分析，总结出了一些可以应用到合成朗读语音和对话语音上的普遍规律，这些规律可以为合成语音添加生理特点和情绪特征，虽然将规则运用到实际的合成当中获得的效果比较有限，但至少可以证明，包括喘息在内的副语言学信息在表意方面是确实存在意义的，特别是在口语对白中不只是单纯的生理活动。

另一方面，本文力图为使用副语言学信息提升语音合成自然度的研究方向探索出一条新路。本文提出的思路是搭建一定规模的副语言学信息语料库，并建立一定的规则将副语言学信息加入到合成语料中。该语料库和合成规则将可以适用于任何合成系统。

参考文献

1. JE Cahn: Generating Expression in Synthesized Speech. - Journal of the American Voice I/O Society (1990)
2. J Vroomen, R Collier, S Mozziconacci:Duration and intonation in emotional speech Proceedings of the Third European Conference on Speech(1993)
3. N Campbell:Where is the Information in Speech.Proceedings of the Third ESCA/COCOSDA International Workshop(1998).
4. A Iida, N Campbell, S Iga, F Higuchi, M Yasumura : A Speech Synthesis System for Assisting Communication, ISCA Workshop on Speech and Emotion (2000).
5. Nick Campbell,: Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation, 8th International Conference on Spoken Language Processing (2004)
6. Alku P and Vilkman E:A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. (1996)
7. Heuft, B. Portele, T. Rauth, M: Emotions in time domain synthesis, 4th International Conference on Spoken Language Processing 6(1996)
8. Jürgen Trouvain: Segmenting Phonetic Units in Laughter, Conference of the Phonetic Sciences, Barcelona, Spain (2003)
9 Scott Sophie and Sauter Disa: Non-verbal expressions of emotion - acoustics, valence and cross cultural factors, SPEECH PROSODY 2006(2006)
10. Aijun Li, Chinese Prosody and Prosodic Labeling of Spontaneous Speech, Speech Prosody ,Aix-en-Provence (2002),
11..Xiaoxia Chen, Aijun Li, et. al. Application of SAMPA-C in SC, ICSLP2000, Beijing (2000)
12 Hiroya Fujisaki, Prosody, Information ,and Modeling, with Emphasis on Tonal Features of Speech, Workshop on Spoken Language Processing, India(2003)