Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 17-32

The Breath Segment in Expressive Speech

Chu Yuan*, and Aijun Li*

Abstract

This paper, based on a selected one hour of expressive speech, is a pilot study on how to use breath segments to get more natural and expressive speech. It mainly deals with the status of when the breath segments occur and how the acoustic features are affected by the speaker's emotional states in terms of valence and activation. Statistical analysis is made to investigate the relationship between the length and intensity of the breath segments and the two state parameters. Finally, a perceptual experiment is conducted by employing the analysis results to synthesized speech, the results of which demonstrate that breath segment insertion can help improve the expressiveness and naturalness of the synthesized speech.

Keywords: Breath Segment, Expressive Speech, Emotion, Valence, Activation

1. Introduction

In the current speech synthesis and recognition systems, some characteristics of spontaneous speech are treated as noise, such as disfluent utterances, repeated sounds, filled pauses, salient breaths and coughs. In corpus collection for speech synthesis and recognition systems, the speaking style of the speakers is always strictly controlled and the speaker is usually required to give a "canonical pronunciation" to decrease the speaking noise as much as possible. However, in recent study, researchers have begun to pay more attention to the non-verbal information in natural speech, especially the paralinguistic and physiological information. They have focused on how to use these types of information to improve the naturalness and expressiveness of emotion and attitude in synthesized speech, so that the speaker's intention can be better understood during verbal communication.

In 1989, Cahn compiled a simple feeling editor based on the phonetic characteristics of emotion [Cahn 1990]. Vroomen, Collier and Mozziconacci examined the duration and intonation of emotional speech and proposed that emotions can be expressed accurately by manipulating pitch and duration based on rules. This conclusion showed that, in emotional

* Institute of Linguistics, Chinese Academy of Social Sciences, No. 5 Jianguomennei Dajie, Beijing, 100732 China
E-mail: Yuanchu8341@gmail.com; liaj@cass.org.cn

[Received August 12, 2006; Revised October 22, 2006; Accepted November 1, 2006]

speech, duration and intonation can be employed to observe the speakers' attitude [Vroomen *et al*. 1993]. In 1998, Campbell found that if one compares the same content in different forms, for example, a news item in its read form, its formal spoken or broadcast form, and its informal conversational form, differences are obvious not only in lexis, word-order, chunking, and prominence relations, but also in the mood of the speaker and in the tone of the voice [Campbell 1998].

In 2000, the International Workshop on Speech and Emotion of ISCA (held in Ireland) invited, for the first time, researchers who were devoted to the study of emotion and speech. Before this conference, many researchers had begun to investigate the voice quality, prosodic features, and acoustic features of emotional speech. Alku and Vilkman designed an experiment to illustrate that the phonation types could be separated from each other effectively when the quantification was based on the parameters extracted from the instant of the maximal glottal opening and the minimal peak of the flow derivative [Alku *et al*. 1996]. Heuft, Portele, and Rauth carried out a more sophisticated test in order to determine the influence of the prosodic parameters in the perception of a speaker's emotional state in three different testing procedures. Their studies proved that the recognition rates were lower than those in the preliminary test, although the differences between the recognition rates of natural vs. synthetic speech were comparable in both tests. The outcome of the saw tooth test showed that the amount of information about the speaker's emotional state transported by F0, energy, and overall duration was rather small. However, the relations between the acoustic, prosodic parameters, and the emotional content of speech could be determined [Heuft *et al.* 1996]. Iida recorded a corpus of one speaker which included three kinds of emotion: anger, happiness and sadness. When synthesizing emotional speech, they picked up the corresponding emotional segments from the emotion corpus. The emotion speech, synthesized in this way, achieved a correct recognition rate 50% ~80% higher than through previous means [Iida *et al*. 2000]. Campbell focused on how to express a modal word in spontaneous speech with various emotions and attitudes [Campbell 2004].

Some researchers have also studied the non-verbal information in emotional speech. Trouvain attempted to analyze the terminological variety from a phonetic perspective. He proposed that the overview of various types of laughter indicated that further concepts of description were needed. In a pilot study on a small corpus of spontaneous laughter, the usefulness of the concepts and terms in practice was examined [Trouvain 2003].

In the light of the above overview of emotion speech research, this paper mainly discusses the function of the non-verbal information in natural speech, specifically the common non-verbal information which includes breath, laugh, filled pause, long silence, and cry. The breath segment is taken as an example to observe how the acoustic characteristics are related to prosodic structure, expressive valence, and activation through statistic analysis of reading and spontaneous speech. The concluded rules are then applied to a perceptual experiment to see how it works.

2. Materials

2.1 Breath Segments

This paper studies breath segments which appear in both read and spontaneous speech, as shown in Figures 1 and 2, annotated between two dotted lines in the read and spontaneous speech, respectively.

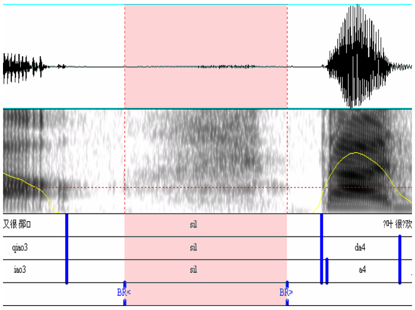

Figure 1. Breath segment in reading speech

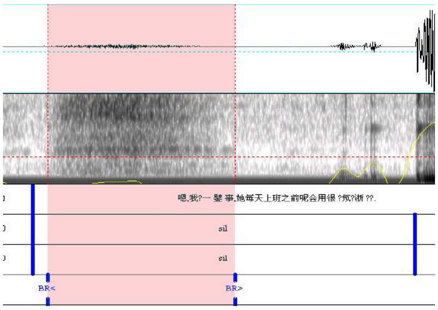

Figure 2. Breath segment in spontaneous speech

The breath shown here is not the normal unconscious physiological exhalation or inspiration process but the deliberate breath for expressing a kind of emotion. Therefore, the following breath segment carries the emotional or attitudinal information of the utterance. Moreover, the acoustic features, such as the length and intensity of the breath segment, may be correlated to the emotional state in terms of valence and activation. Further, the small blanks preceding and following the breath segment which are caused by the physiological need of a breath segment may be inserted when the synthesis of emotional speech is conducted.

The breath has two functions: fulfilling the physiological requirement of the intake of air and the expression of emotion or attitude. The authors determine the activation and valence degrees for each recitation of each phrase and use the information to label the breath segment before this phrase.

2.2 The Corpus and Annotation

The corpus used in this paper is called CASS-EXP which includes read and spontaneous speech. The first part contains some stories read by actors and actresses in emotional and neutral states while the second part includes TV and radio programs along with spontaneous speech: monologues and dialogues.

SAMPA-C [Li 2002] and C-ToBI [Chen *et al.* 2000] are adopted to label segmental and prosodic information. Furthermore, the starting and ending points of breath segments in terms of valence and activation degrees are labeled as well.

The authors labeled the emotion characteristics of the breath segments based on two factors: valence and activation. The theoretical foundation of valence is the concept of a separation of positive or negative emotion. The function of activation is the enabled degree of energy which is in contact with the emotion condition. The activation and valence of one breath segment here refer to the activation and valence of the following intonational phrase.

Emotional valence is categorized into three levels: positive (1), neutral (0) and negative (-1). The activation has three categories as well: excited (1), steady (0) and low (0). When both the emotional valence and activation of a certain breath segment are marked as 0, the breath segment is considered to be a neutral physiological segment without carrying any expressive information.

Three boundary levels (break index) 1, 2, 3 are annotated which stand for prosodic word, minor prosodic phrase, and major prosodic phrase (intonational phrase), respectively. The authors intend to examine whether the breath segment occurs in a normal stop or in an unexpected position. The normal stop refers to the breath at a prosodic phrase boundary, and the unexpected or abnormal position is the breath at a prosodic word boundary or within a prosodic word.

3. Breath Segments in Read Speech

From CASS-EXP, the authors select fifteen fragments from a read story which have different emotional states and attitudes. The valence and activity of nine fragments were labeled.

3.1 Occurring Number and Position of the Breath Segments

Based on what has been labeled, the number of breath segments is calculated for neutral and expressive speech. It was found that the number of breath segments in expressive speech is 50% higher than in that of neutral read speech in the same text. In these nine fragments, the number of breath segments in expressive speech is 334, and only one of them appears in an abnormal stop; the number in neutral speech is 225, of which all appear in normal boundaries, as shown in figure 3.

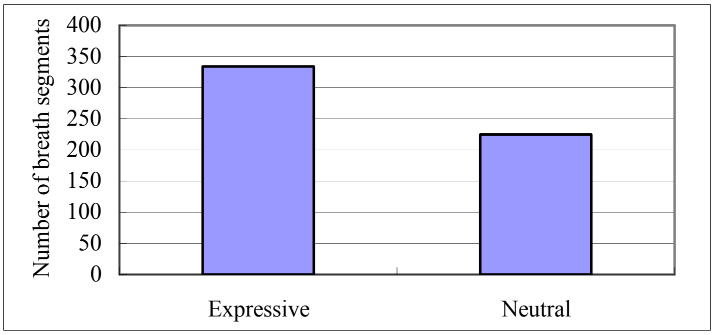

Figure 3. Number of breath segments in expressive and neutral read speech.

In fragments of read form, most of the breath segments occur at boundary 3 (intonational phrase boundary). The number of the breath segments at boundary 1 (prosodic word boundary) is the smallest, as shown in Figure 4. Table 1 demonstrates that the boundary distribution of breath segments appearing in expressive speech and neutral speech exhibits no difference. In expressive and neutral speech, the number of breath segments at boundary 1 is the smallest, and the number of breath segments at boundary 3 is the largest.

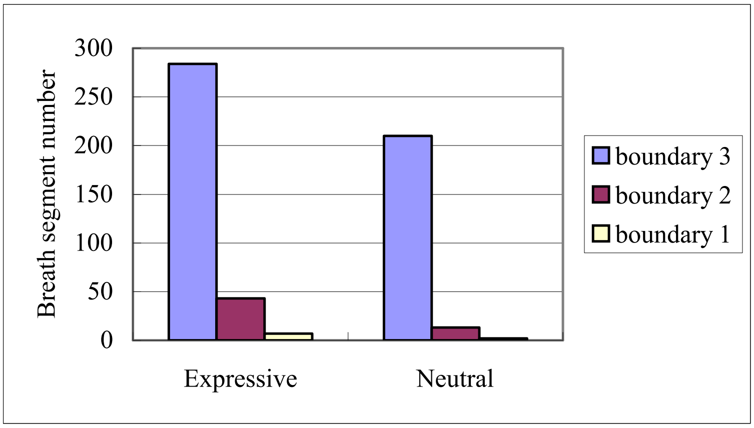

Figure 4. The number of breath segments at the different boundaries.

Table 1. Number and percent of breath segments of emotion and neutral read speech at the different boundaries.

Boundary	Number of breath segments in expressive speech	Percent	Number of breath segments in neutral speech	Percent
3	284	85.2%	210	93.3%
2	43	12.8%	13	5.8%
1	7	2%	2	0.9%

In general, breath segments in read speech, either expressive or neutral, usually appear between two prosodic phrases, especially between two intonational phrases. From the perspective of syntactic analysis, most of the breath segments appear between two intonational phrases or two intonational phrase groups.

We measured the duration of the silence which was between the breath segment and the prosodic phrase following this breath segment. The mean duration of the silence in different valence and activity is shown in Table 2.

Table 2 .The mean duration of the silence in different valence and activity

	Valence			Activity		
	-1	0	1	-1	0	1
Emotional	64ms	54ms	40ms	78ms	52ms	28ms
Neutral		48ms			49ms	

From this table we can know that in neutral speech the duration of the silence which was between the breath segment simply and the prosodic phrase following this breath segment is about 50ms. In emotional speech the durations are different because of the different valence and activity.

3.2 Duration of Breath Segments in Read Speech

In these nine fragments whose valence and activity have been labeled, the number of breath segments in expressive speech is 200, and only one of them appears in an abnormal stop; the number in neutral speech is 133, of which all appear in normal boundaries.

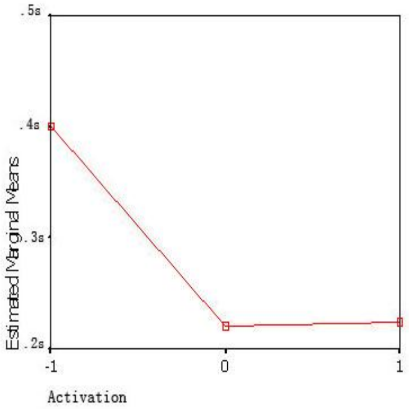

Figure 5. Breath segment mean duration and activation

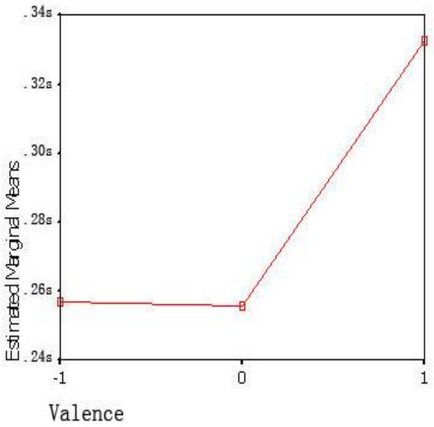

Figure 6. Breath segment mean duration and valence

The durations of breath segments are measured and put into a multi-variance analysis using SPSS. Breath segment means are shown in figure 5 and figure 6. In the analysis of the relationship between the valence degree and the duration of the breath segment, it was found that there is no significant correlation between the three categories of emotion valence and the duration of the breath segment (P=0.063>0.05).

However, activation has significant influence on the breath duration (P=0.000<0.05). The result of the analysis indicates that when the activation is 0 or 1, the discriminative degree of duration is not very high; when the activation is -1, the degree is different from that in other two activation states.

Table 3. Tests of between-subjects: valence and activation effects to the duration and intensity of breath segment.

Source	Dependent Variable	F	Sig.
valence	intensity	.544	.581
	duration	2.801	.063
activation	intensity	10.313	.000
	duration	9.344	.000
valence* activation	intensity	.371	.829
	duration	2.092	.083

Table 3 displays the effect triggered by valence and activation on intensity and duration. The valence has no effect on the breath duration and there is no interactive effect of valence and activation on intensity and little on duration (P=0.083). This result proves that, although the speakers express a certain kind of emotion, the physiological response does not differ from that of neutral speech. Nevertheless, because we do not know that the compute method in SPSS is the same as the person's mental perception mechanism or not. In this kind of case, we think that the effect triggered by valence and activation has influence of breath segments.

In addition to the duration of breath segments, the authors computed the intervals between two breath segments and their distribution. Among the 319 intervals there were 304 intervals shorter than 10 seconds. The other 15 intervals which include error reading were longer than 10 seconds. So this confirms that, when a text is read at normal speed, the time between two breath segments is shorter than 10 seconds.

3.3 Intensity of Breath Segments

Another important characteristic is the intensity of the breath segments. Tables 4 and 5 are the statistical results on intensity grouped by valence and activation.

Table 4. Breath segment intensity grouped by valence

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

Table 5. Breath segment intensity grouped by activation

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

Afterwards, the authors observed the relationship between the intensity of every breath segment and the intensity of the following intonational phrase. Through the examination of the data obtained from SPSS analysis which be shown in table 6, it was found that activation has a significant effect on the intensity ratio of the following intonational phrase in the breath segment; in addition, the effect of valence and the interactive effect of valence and activation are significant as well.

Table 6. Tests of between-subjects effects which is valence and activation effects to the IR

Source	Sig.
Activation	.022
Valence	.913
Activation * Valence	.609

Table 7 provides the means and ranges of intensity ratios of the following intonational phrase to the present breath segment (IR) in three categories of activations. The intensity ratio is the lowest when the activation is 0.

Table 7. The means and ranges of intensity ratios in three categories of activation

Activation	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

3.4 Rules of Inserting Breath Segments to Read Speech

One can obtain rules of breath segment insertion based on the previous analysis of synthesized speech. The breath segment corpus can be set up first for the selected speaker. When the speech is being synthesized, the fitted breath segments can be selected and inserted into the expected positions. The insertion rules are summarized as follows:

A. At every major prosodic phrase boundary, a breath segment can be inserted or produced. The durations of these breath segments are about 0.5 second or longer.

B. Intervals between two breath segments are no longer than 10 seconds, *i.e.* one sentence group length in text is shorter than 10 seconds.

C. Within one intonational group, the number of the breath segments is uncertain, generally, there are one or two breath segments before longer intonational phrase and the breath duration ranges from 0.1 to 0.3 second.

D. When the activation of breath segment is not 0, the intensity of this breath segment is set to 0.6 -0.7 times of the intensity of following prosodic phrase. When the activation of breath segment is 0, the intensity of this breath segment is 0.5 times of the intensity of the following prosodic phrase.

E. Between every breath segment and the prosodic phrase following this breath segment there is a silence.

F. The duration range of different kind of valence and activation is induced from the read speech. The breath segment in the synthesized speech is selected random in the range of corresponding kind.

Although the breath segment is not the only way to express emotion or attitude in read speech, breath segments inserted in the synthetic speech can prompt the naturalness and expressiveness. Also, the synthesis speech with breathy segment insertion is more acceptable to the subjects.

4. Breath Segments in Spontaneous Speech

The authors select nine dialogs from the CASS-EXP corpus. Each dialog is a conversation between an audience and a radio host through a hotline telephone. It is assumed that the radio hostess's emotion is the performed emotion while the audience's is natural. In this part, boundary 4 is used to label the turn taking boundary.

4.1 Positions of Breath Segments in Spontaneous Speech

In these nine dialogs, 55 breath segments produced by the radio hostess and 17 breath segments are at abnormal positions, *i.e.* unexpected prosodic boundaries, which account for about 32% of the total breath segments. The audiences make, altogether, 54 noticeable breaths at normal boundaries and 19 at abnormal ones ，which occupy about 35.2% of the total.

The radio hostess produces 11 physiological breath segments while the audience produces only 6. These 17 segments all appear at major prosodic phrase boundaries. In general, the physiological breaths that appear in spontaneous speech are similar with those in read speech but the frequency of appearance declines greatly.

From Table 8, one can see that the distribution of the physiological breath segments produced by the radio hostess is well-proportioned. The physiological breath segments produced by the audiences appear at boundaries 3 (prosodic phrase) or 4 (turn taking). Thus, the data help prove that when the expressiveness is a performed one, the breath distribution is the same as that in neutral speech. However, for spontaneous speech with natural expression (in Table 9), the breath also appears at boundaries 1 and 2. So, one can confirm that, in natural emotion speech, most of boundaries 1 and 2 are made intentionally. If one synthesizes this kind of speech material, one can consider breaking the original prosodic structures by adding breath segments.

Table 8. The breath segment distribution at prosodic boundaries by the radio hostess

Boundary	Total	Abnormal position	Normal position	Physiological breath
1	6	6	0	2
2	23	10	13	4
3	16	1	15	3
4	10	2	8	2

Table 9. The breath segment distribution at prosodic boundaries by the audiences

Boundary	Total	Abnormal position	Normal position	Physiological breath
1	9	6	3	0
2	9	8	1	0
3	14	2	12	2
4	22	3	19	4

4.2 Duration of Breath Segments in Spontaneous Speech

Figures 7 and 8 show the duration distribution of the breath segments made by the radio hostess according to valence and activation. The bottom and top value are 25% and 75% accumulative frequency, respectively, standing for duration variation range. (Note that in Figure 7, when activation is -1, the token number is relative small).

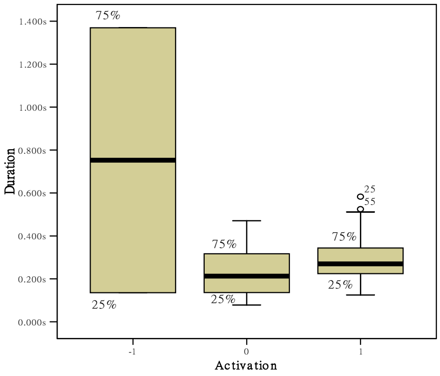

Figure 7. The duration distribution of the breath segments by radio hostess in different activations

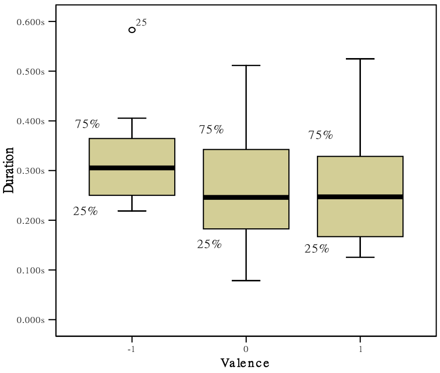

Figure 8. The duration distribution of the breath segments by radio hostess in different valences

Figures 9 and 10 indicate that the duration range of the breath segments produced by the audience is affected by the valence and activation.

From these four figures, one can get the duration of breath segments when valence and activation are 1,-1 or 0 in spontaneous speech, whose results can be used in the following perceptual experiment.

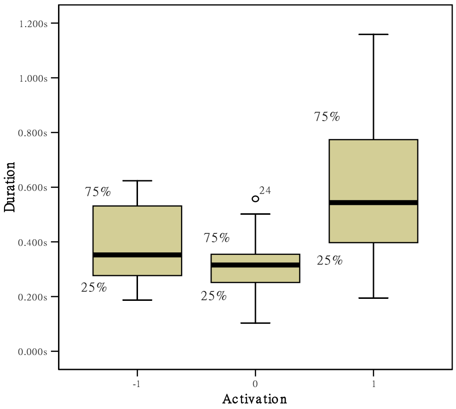

Figure 9. The duration distribution of the breath segments by audience in different activation

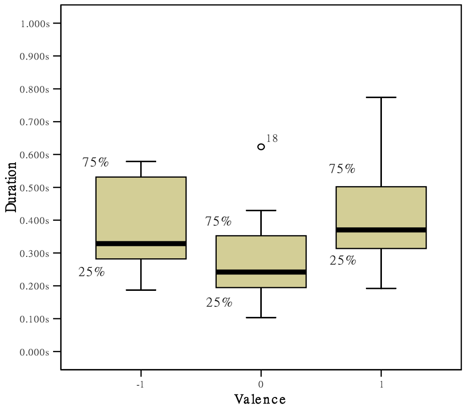

Figure 10. The duration distribution of the breath segments by audience in different valence

4.3 Rules for Inserting Breath Segments in Spontaneous Speech

The insertion rule in spontaneous speech is more complicated than that in read speech. In spontaneous speech, the breath segments will be divided into two types according to their functions: the physiological activity and the expression of emotion or attitude. The following rules can be used in breath insertion for synthesizing spontaneous speech.

A. Physiological breath insertion without emotion is the same as that in read speech as described above. However, in dialogs there is some turn-taking. Sometimes, the breath appearing at the turn taking may overlap with the words spoken by the interlocutor or appear close to the boundary of the turn taking.

B. When the activation is -1, the duration of breath segment is set randomly from 0.2 to 0.6 second. When the activation is 1, the duration of breath segment is set randomly ranging from 0.1 to 0.4 second. When the activation is 0, the duration of breath segment is set randomly from 0.2 to 0.5 second.

C. When the valence is -1, the duration of breath segment is set randomly from 0.1 to 0.4 second. When the valence is 1, the duration of breath segment is set randomly from 0.2 to 0.5 second. When the valence is 0, the duration of breath segment is set randomly from 0.2 to 0.6 second.

D. Between every breath segment and the prosodic phrase following this breath segment there is a silence.

5. Perceptual Experiments

5.1 Stimuli

A pilot perceptual experiment is conducted to test the obtained results. The texts are selected from a read story and spontaneous dialogs. The original synthesized speech is produced by using the synthesizer provided by iFLYTEK. After that, breath segments are inserted into the synthetic speech, based on the previous rules.

Twenty subjects recruited to join the perceptual experiment are asked to judge the differences between the speech materials with and without breath for both the original and the synthesized speech. The perceptual process consists of two steps: first, the subjects are asked to compare the speech from the read story. Then, these subjects are required to perceive the breath effect in the synthesized dialogs.

Speech fragments from a read story (Little Red Hat) are numbered as X-1 (the original speech), X-2 (the original speech minus the breath segments), X-3 (the synthetic speech) and X-4 (the synthetic speech inserted with breath segments). For speech based on the spontaneous speech scripts, the two stimuli are numbered as Y-1 and Y-2, which are synthesized speech and inserted with breath segments.

5.2 Results

In the first experiment, the whole speech or segmented clips are compared. Five clips are segmented for each X. Totally, 20 clips are attained for X1, X2, X3 and X4 by segmenting at the same text boundaries. Subjects are asked to listen to and compare all counterparts with and without breath segments to judge if they are different or not and which is more natural. The subjects are only allowed to listen to the stimuli a maximum of 3 times.

The results are listed in Table 10, in which 1 stands for the counterparts (with and without breath segments) which are different, 0 means there is no difference between the perceived counterparts. 70% subjects fail to distinguish between X1 and X2. Carefully comparing X3 with X4, subjects can perceive their differences, and feel that X-4 is more natural. When smaller fragments are compared, only 38% (38 out of 100 times) can be perceived with discrepancy. The results on X3 and X4 are slightly higher, reaching 92% (92 out of 100 times). This experiment reveals that when one changes the parameters of breath segments, such as their duration, intensity and position, most of the subjects are able to perceive the differences between the original and the breath insertion speech.

Table 10. The perceptual results of the first experiment based on reading story

Subjects	X-1 and X-2 (in five clips)	X-3 and X-4 (in five clips)
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	2/5	4/5
11	2/5	4/5
12	1/5	5/5
13	2/5	5/5
14	3/5	4/5
15	2/5	5/5
16	2/5	4/5
17	1/5	5/5
18	1/5	4/5
19	2/5	5/5
20	1/5	5/5
Total	38/100	92/100

Table 11. The result on spontaneous dialogues Y1 and Y2

Subjects	Y-1			Y-2		
	breath	naturalness	expressiveness	breath	naturalness	expressiveness
1	1	1	0	1	1	1
2	0	0	0	1	1	0
3	1	1	0	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	0	0
6	1	0	1	1	0	0
7	0	0	0	1	1	0
8	1	0	0	0	0	0
9	0	0	0	0	0	0
10	1	1	1	1	1	1
total	5/10	3/10	2/10	6/10	4/10	2/10

The second experiment is rather simple, compared to the first one. The subjects are asked to judge which group of the two dialogs Y1 and Y2 has breath segments. If the subjects can tell the difference, they have to judge whether or not the breath segments insertion can increase the naturalness and the expressiveness. The result is shown in Table 11. The rates of breath insertion recognition are 50% and 60% for Y1 and Y2 respectively, but only 20% for expressiveness and 30% to 40% for naturalness.

6. Conclusion

This paper, with a statistical analysis made on breath segments in read and spontaneous speech, proposes some preliminary principles for inserting breath segments in synthesized speech. These principles or rules can help one better understand the physiological and expressive features in speech synthesis. Though the authors got relatively limited results in the perceptual experiments, it proves that non-verbal information is not just a simple physiological breath; instead, it is an essential element in transmitting expressiveness or attitude. In this regard, future studies should focus on other frequently encountered paralinguistic and nonlinguistic information, so that further steps may be achieved in understanding breath segments by classifying valence into more categories.

References

Alku, P., and E. Vilkman, “A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers, ” *Folia phoniatrica et logopaedica* Karger ,48(55), 1996, pp. 240-254.

Cahn, J.E., “Generating Expression in Synthesized Speech,” Master' s Thesis, MIT,1989.

Campbell, N., “Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation,” In *Proceeding of 8th International Conference on Spoken Language Processing,* Jeju, Korea, 2004, pp. 881-884.

Campbell, N., “Where is the Information in Speech?” In *Proceedings of the Third ESCA/COCOSDA International Workshop*, 1998, Australia, pp. 17-20.

Chen, X.-X., A.-J. Li, et. al, “Application of SAMPA-C in SC,” In *Proceeding of ICSLP2000*, 2000, Beijing, pp.VI 652-655.

Heuft, B., T. Portele, and M. Rauth, “Emotions in time domain synthesis,” In *Proceeding of 4th International Conference on Spoken Language Processing*, Philadelphia, USA,1996, pp. 1974-1977.

Iida, A., N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, “A Speech Synthesis System with emotion for Assisting Communication”, In *Proceeding of ISCA Workshop on Speech and Emotion*, Northern Ireland , 2000, pp. 167–172.

Li, A.-J., “Chinese Prosody and Prosodic Labeling of Spontaneous Speech” In *Proceedings of International. Workshop on Speech Prosody*, Aix-en-Provence, France, 2002, pp. 39-46.

Trouvain, J, “Segmenting Phonetic Units in Laughter, Conference of the Phonetic Sciences,” In 15th. *International Conference of the Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2793-2796.

Vroomen, J., R. Collier, and S. Mozziconacci, “Duration and intonation in emotional speech,” In *Proceedings of the Third European Conference on Speech,* Berlin, Germany, 1993, pp. 577–580.