The Breath Segment in Expressive Speech

Chu YUAN, Aijun LI

Institute of Linguistics, Chinese Academy of Social Sciences
Yuanchu8341@gmail.com; liaj@cass.org.cn

Abstract. In this paper, we choose about one hour expressive speech and make a pilot study on how to use the breath segments to get more natural and expressive speech. We focus on exploring when the breath segments occur and how their acoustic features are affected by the speaker's emotional states: valence and activation. The statistic analysis has been carried out to find the relationship between the length and intensity of the breath segments and the two state parameters. Finally, a perceptual experiment is done by employing the analysis results to the synthesized discourses, the results imply that breath segment insertion can help to improve the expressiveness and naturalness of the synthesized speech.

Keywords: breath segment, expressive speech, emotion, valence, activation

1 Introduction

In the present speech synthesis and speech recognition systems, some characteristics of the spontaneous speech have been ignored. In corpus recording，annotation, or speech synthesis and recognition systems, the speaking style of the speakers is always strictly controlled, letting them give a "canonical pronunciation" and decrease speaking noise as much as possible, such as physiological breath, cough... But recently, researchers begin to pay more attention to the paralinguistic information and physiological information in natural speech, which we call non-symbolic information. They are focusing on how to use this information to improve the naturalness and expressiveness (emotion and attitude) of the synthesized speech.

In 1989, Cahn compiled a simple feeling editor according to the emotional phonetic characteristics [1]. In 2000, the International Workshop on Speech and Emotion of ISCA held in Ireland, for the first time, it gathered all the researchers who are devoted to the studies of emotion and speech. Lida recorded three kinds of emotion speech to build corpora, including anger, happy and sad, spoken by the same speaker .When synthesizing speech with some kind of emotion, they pick up the corresponding emotion segments from the emotion corpus. The synthesized emotion speech by this way has 50% to 80% of correct recognition rate than before [2]. Nick Campbell made a research on how a modal word be spoken in various ways to express different various emotions or attitudes in spontaneous [3]. Jürgen Trouvain tries to analyze the terminological variety from a phonetic perspective. He thinks that a short overview on various types of laughter indicates that further concepts for description are needed. In a pilot study with a small corpus of spontaneous laughter the usefulness of the concepts and terms in practice is examined [4].

This paper explores the function of the non-symbolic information in natural speech. Common non-symbolic information includes breath, laugh, filled pause, long silence, cry, etc. This paper takes breath segment for example to observe how their acoustics characteristic is related to prosodic structure, expressive valence and activation by statistic analysis on read and spontaneous speech. The concluded rules are then applied to a pilot perceptual experiment to see how it works.

2 Materials

2.1 Breath segment

The object that we will study in this paper is the breath segment which appears in read speech and spontaneous speech, as shown in figures 1 and 2, annotated between two dotted lines in the read speech and spontaneous speech respectively.

Before and after each breath segment there is small blank caused by the physiological need of breath segment. When inserting breath segment in synthesized speech we must notice this blank gap.

The breath has two functions: physiological requirement and expression of emotion or attitude. We label the activation the intonation of one intonation phrase and use the information to label the breath segment before this phrase.

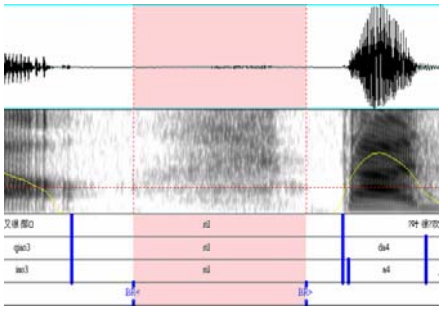


Fig. 1. Breath segment in read speech

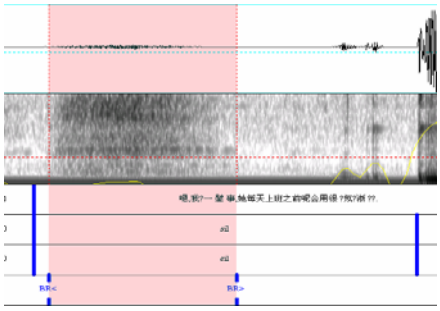


Fig. 2.Breath segment in spontaneous speech

2.2 Corpora and annotation

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

We used SAMPA- C [5] and C-ToBI [6] to make segmental and prosodic annotation. Besides, the starting and ending points of breath segments, the valence and activation degrees are labeled as well.

We labeled the emotion characteristics of breath segments in two factors, which are valence and activation. The theories foundation of valence is a separation of positive or negative emotion. The function of activation is the enabled degree of energy which contacts with the emotion condition. The activation and valence of one breath segment refer to the activation and valence of the following intonation phrase. Emotion valence is categorized into 3 levels, positive, neutral and negative. Activation has 3 categories as well: excited, steady and low. When both the emotion valence and the emotion activation of a certain breath segment are marked as 0, we think that this breath segment is a neutral physiological segment and doesn't carry any expressive information.

Three boundary level (Break index) 1, 2, 3 are annotated which stand for prosodic word, minor prosodic phrase and major prosodic phrase (intonational phrase) respectively. We have to decide whether the breath segment occurs in a normal stop or in an unexpected position. The normal stop refers to breathe at prosodic phrase boundary and the unexpected or abnormal position refers to breathe at prosodic word boundary or within prosodic word.

3 Breath segments in read speech

From CASS-EXP we selected 9 fragments from a read story which have different emotion states and attitudes.

3.1 Occurring number and position of the breath segments

Based on what we have labeled, the number of breath segments is calculated for neutral and expressive speech. We find that the number of breath segment in expressive speech is 50% higher than that in neutral read speech regarding to the same text. In these nine fragments, the number of breath segments in expressive speech is 200 and only one appears in an abnormal stop, 133 in neutral speech and all appear in normal boundaries.

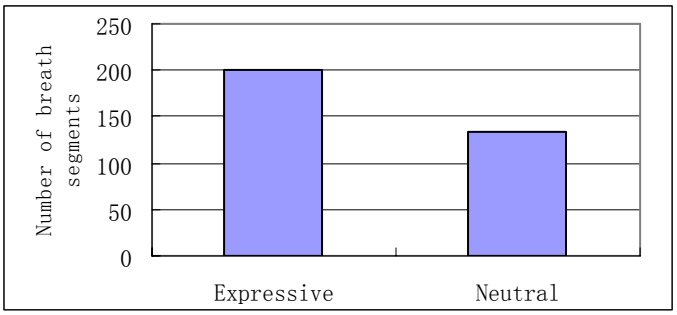


Fig. 3.Number of breath segments in expressive and neutral read speech.

In read fragments, most of breath segments occur at boundary 3(intonation phrase boundary).The number of breath segments in boundary 1(prosodic word boundary) is the least, as shown in Fig 4. Table

1 shows that the boundary distribution of breath segments appearing in expressive speech and neutral speech has no difference. In expressive speech and neutral speech, the number of breath segments in boundary 3 is the least and the number of breath segments in boundary 3 is the most.

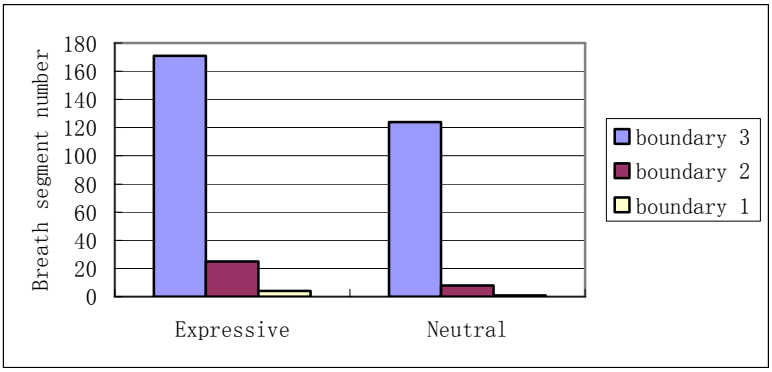


Fig. 4. The number of breath segments at the different boundaries.

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

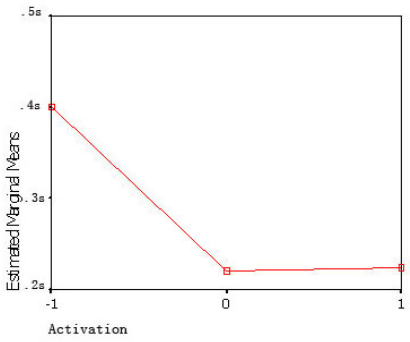
Boundary	Number of breath segments in expressive speech	Percent	Number of breath segments in neutral speech	Percent
3	171	85.5%	124	93.2%
2	25	12.5%	8	6%
1	4	2%	1	0.8%

In a word, breath segments in read speech, either expressive or neutral, usually appear between two prosodic phrases, especially between two intonation phrases. If we analyze the texts syntactically, most of the breath segments appear between two intonation phrases or intonation phrase groups.

3.2 Duration of breath segments in read speech

The durations of breath segments are measured and put into SPSS to do a multi-variance analysis. In the analysis of the relationship between the valence degree and the duration of breath segment, there is no significance between three categories of emotion valence and the duration of breath segment. (P=0.063>0.05).

However, activation has significant influence on the breath duration (P=0000<0.05).The result of the analysis shows that when the activation is 0 or 1, the discriminative degree of durations is not very high. But when the activation is -1 the durations is different from that in other two activation states.



.Fig. 5.Breath segment mean duration and activation

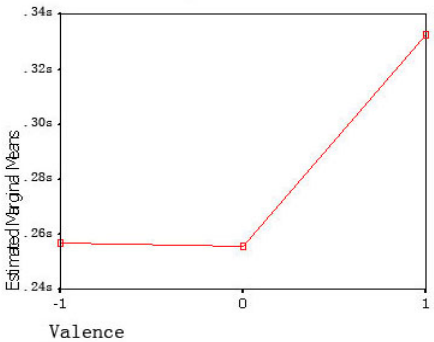


Fig. 6 Breath segment mean duration and valence

Table 2. 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 2 shows the effect of valence and activation to intensity and duration. The Valence has no effect on 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 perform a certain emotion, the physiological response is not different with that in neutral speech.

In addition to the duration of breath segments, we compute the intervals between two breath segments and their distribution. Out of 319 intervals there are 304 intervals shorter than 10 seconds. And the other 15 intervals are longer than 10 seconds, which include error reading. So we can make sure that when reading normally, the time between two breath segments is shorter than 10 seconds.

3.3 Intensity of breath segments

Another important characteristic is the intensity of breath segments. Table 3 and table 4 are the statistic results on intensity grouped by valence and activation.

Table 3. 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 4. 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, we observe the relationship between the intensity of every breath segment and the intensity of the following intonation phrase. Using SPSS analysis we find that activation has significant effect on the intensity ratio of the following intonation phrase to this breath segment, but effect of valence and an interactive effect of valence and activation are significant.

Table.5. Tests of between-subjects effects which is valence and activation effects to the intensity ratio of the following intonation phrase to this breath segment

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

Table 6 gives the means and ranges of intensity ratios of following intonation phrase to the present breath segment in three categories of activations. The intensity ratio is the lowest when activation is 0.

Table.6. The means and ranges of intensity ratios in three categories of activations

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 for inserting breath segments to read speech

We can obtain breath segment inserting rules based on the previous analysis to synthesized speech. A breath segment corpus can be set up first for speakers. When synthesizing speech, we can select the fitted breath segments to insert into the expected position. The insertion rule is 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 Interval between two breath segments is no longer than 10 seconds, i.e. one sentence group length in text is shorter than 10 seconds.

C In one intonation group, the number of the breath segments is uncertain, generally once or twice before rather longer intonation phrase. The breath duration is 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 Breath segment is not the only way to express emotion or attitude in natural read speech. But inserting the breath segments in synthetic speech may makes it more natural and expressive. And the synthesis speech with breathy segment insertion is more acceptable by the subjects.

4 Breath segments in spontaneous speech

We select 9 dialogs from the CASS- EXP corpus. Each dialog is a conversation between an audience and a radio host through an emotional hotline. We consider that the radio host's emotion is the performed emotion and the audience's emotion is natural. In this part we use boundary 4 to label the turn taking boundary.

4.1 Position of breath segment in spontaneous speech

In these nine dialogs, there are 55 breath segments produced by the radio hostess and 17 breath segments are at the abnormal positions i.e., unexpected prosodic boundaries, which accounts for about 32% of the whole breath segments. All the audiences have made 54 visible breaths at normal boundaries and 19 at the abnormal positions，which accounts for about 35.2%.

The radio hostess produces 11 physiological breath segments. All audiences just have 6 physiological breaths. These 17 physiological breath segments all appear at major prosodic phrase boundaries. In general, the physiological breaths that appear in spontaneous speech are similar as in read speech but the frequency of appearance declines greatly.

From table 7 we see that the distribution of the physiological breath segments produced by the radio hostess is well-proportioned. The physiological breath segments making by all the audiences all appear at boundary 3(prosodic phrase) or boundary 4(turn taking).So the data prove that when the expressiveness is performed one, the breath distribution is the same as that in neutral. But for spontaneous speech with natural expression (in table 8), the breath also appears at boundary 1 and boundary 2. So we can believe that in natural emotion speech most of boundary 1 and boundary 2 is made intentionally. If we synthesize this kind of speech material, we can consider breaking the original prosodic structures for adding breath segments.

Table 7. The breath segment distribution at prosodic boundaries for the radio hostess

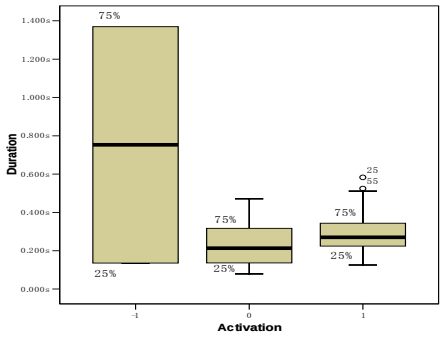
Boundary	Total	Abnormal position	Normal position	Physiological breath
1	6	6	0	2
2	23	10	13	4
3	16	1	15	3
4	10	2	8	2

Table.8. The breath segment distribution at prosodic boundaries for 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 that the duration distribution of the breath segments making by radio hostess according to valence and activation. The box bottom and top lines are 25% to 75% accumulative frequency respectively, standing for duration variation range. In figure 7, when activation is -1 the number of data is too little to be believed.



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

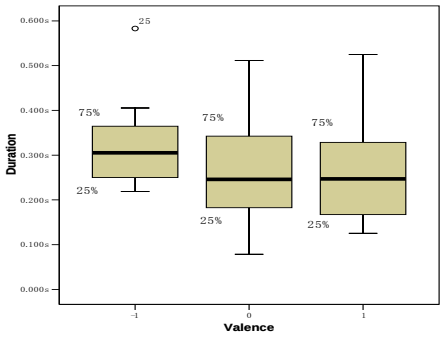
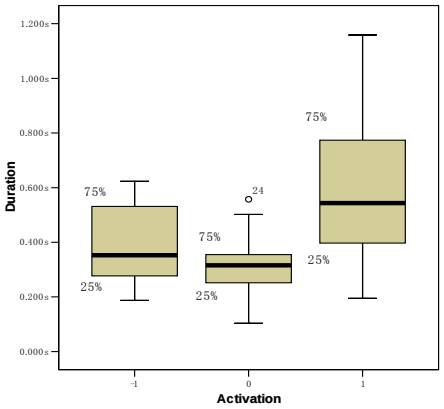


Fig. 8 The duration distribution of the breath segments by radio hostess in different valences

Figure 9 and figure 10 show that the duration range of the breath segments making by audience is affected by valence and activation.

From these four figures, we can get the duration of breath segments when valence and activation is 1,-1 and 0 in spontaneous speech and can use these results in the following perceptual experiment.



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

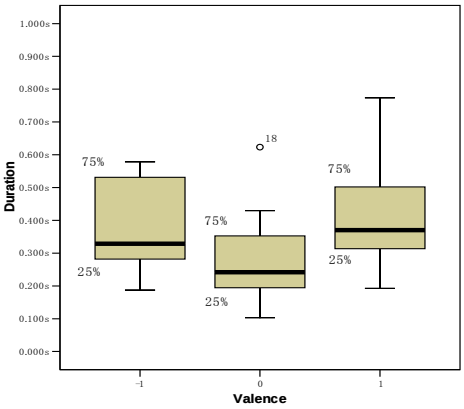


Fig. 10 The duration distribution of the breath segments by audience in different valence

4.3 Rules for inserting breath segments in spontaneous speech

In spontaneous dialog, the insertion rule is obviously more complicated than that for read speech. In spontaneous speech, the function of 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 for breath insertion for synthetic spontaneous speech.

A Physiological breath insertion without emotion is the same as that in read speech. But in dialogs there are some turn takings. Sometimes the breath appearing at turn taking may overlap with the words

speaking by the interlocutor or appear close to the boundary of turn taking.
B When activation is -1, the duration of breath segment is set randomly during 0.2 to 0.6 second. When activation is 1, the duration of breath segment is set randomly during 0.1 to 0.4 second. When activation is 0, the duration of breath segment is set randomly during 0.2 to 0.5 second.
C When valence is -1, the duration of breath segment is set randomly during 0.1 to 0.4 second. When valence is 1, the duration of breath segment is set randomly during 0.2 to 0.5 second. When valence is 0, the duration of breath segment is set randomly during 0.2 to 0.6 second.

5 Perceptual experiment

5.1 Stimuli

A pilot perceptual experiment was conducted to test the results which we have got. The texts were selected from a read story and spontaneous dialogs and the original synthesized speech was produced by using the synthesizer provided by iFLYTEK Co.. After that, breath segments were inserted into the synthetic speech according to the previous rules.

Ten subjects were recruited to join our perceptual experiment to listen to the difference between speech with and without breath for both original and the synthesized speech. The perceptual process consists of two parts: at the first part, subjects were asked to compare the speech from the read story. At the second part, we ask the subjects to judge the breath effect in synthesized dialogs.

Speech fragments from a read story (Little Red Hat) are numbered as X-1 (the original speech), X-2 (the original speech by throwing away 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 inserted with breath segments.

5.2 Results

In the first experiment, the whole speech or segmented clips were compared respectively. 5 clips were segmented for each X, totally 20 clips were got for X1, X2, X3 and X4 by segmenting at the same text boundaries. Subjects were asked to listen and compare all counterparts with and without breath segments, to tell if they were different or not and which was more nature. The subjects could listen to the stimuli no more than 3 times.

The results are shown in table 9, 1 stands for the counterpart of the two sounds are different, 0 stands for there is no difference between the perceived counterparts. 60% subjects could not distinguish between X1 and X2. Carefully comparing X3 with X4, subjects can perceive their difference, and feel X-4 is more nature. When smaller fragments were compared, only 40% (20 out of 50 times) can be perceived with discrepancy. The result on X3 and X4 is slightly higher, reaching 94% (47 out of 50 times). This experiment reveals that when we change the parameters of breath segments such as their duration, intensity and occurring position, most of subjects are able to perceive the distinguish between the original and the breath insertion speech.

Table.9. The perceptual result of the first experiment based on read 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	1/5	5/5
Total	20/50	47/50

Table.10. 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 than the first one. We asked the subjects to judge which group of two dialogs Y1 and Y2 had breath segments, if the subject could tell the difference, then they had to judge whether the breath segments insertion could increase the naturalness and the expressiveness. The result is shown in table 10. 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

In this paper, we make a statistical analysis on breath segments in read and spontaneous speech, and draw some preliminary principles for inserting breath segments in synthesized speech. These principles or rules can promote physiological and expressive features in speech synthesis. Even though we got a relatively limited result in perceptual experiment, at least it is proofed that non-symbolic information is not a simple physiological action. Instead, it is an essential element in transmitting expressiveness or attitude.

The next research should focus on other frequently appeared paralinguistic and nonlinguistic information, and go a further step into breath segments by classify valence into more categories.

References

1. JE Cahn: Generating Expression in Synthesized Speech. - Journal of the American Voice I/O Society (1990)
2. A Iida, N Campbell, S Iga, F Higuchi, M Yasumura : A Speech Synthesis System for Assisting Communication, ISCA Workshop on Speech and Emotion (2000).
3. 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)
4. Jürgen Trouvain: Segmenting Phonetic Units in Laughter, Conference of the Phonetic Sciences, Barcelona, Spain (2003)
5. Aijun Li, Chinese Prosody and Prosodic Labeling of Spontaneous Speech, Speech Prosody ,Aix-en-Provence (2002),
6..Xiaoxia Chen, Aijun Li, et. al. Application of SAMPA-C in SC, ICSLP2000, Beijing (2000)