A Preliminary Study on Silent Pauses in Mandarin Expressive Speech

Xia Wang[1], *Aijun Li*[2], *Chu Yuan*[2]

[1] Nokia (China) Research Center, China,
[2] Institute of Linguistics, Chinese Academy of Social Sciences, China
xia.s.wang@nokia.com, liaj@cass.org.cn, Yuanchu8341@gmail.com

Abstract

Pause is not only one of the cues of prosodic structure for running speech, but also one of the strategies used by speakers to express their emotions. In this paper, we investigated the silent pauses in read and spontaneous speech, neutral and expressive speech, as well as synthesized speech, trying to find the difference of silent pause patterns in different speech styles. It was observed that all silent pauses in synthesized speech were regular, 2.5% pauses in neutral read speech were irregular, 10% in expressive read speech were irregular and 28% pauses in spontaneous expressive speech were irregular. A comparative study was carried out for the duration and frequency of occurrence of irregular silent pauses in different speech styles, as well as the valence and activation features of their acoustic contexts. A perceptual experiment showed that the removing of irregular silent pauses from spontaneous speech resulted in an obvious decrease of expressiveness, while inserting irregular silent pauses into synthesized speech improves substantially the expressiveness of such synthesized speech.

1. Introduction

The mainstream concatenated Text-To-Speech (TTS) or speech synthesis systems often use neutral monotone speech databases for easier concatenation and try to avoid the richness of paralinguistic phenomena in spontaneous speech because they are too complicated. Now the TTS technology development has reached a level for commercial use and starts to look for new challenges like expressiveness and personalization.

A commercial TTS system could produce well-enough monotone, neutral, fluent human speech, which sounds beautiful for a minute but boring for an hour like in E-book reader applications, because its prosody is generated by a so-called prediction module without many variations. Phoneticians try to solve the problem by studying expressive and spontaneous human speech, and put more efforts on the rich paralinguistic phenomena in spontaneous speech, which was kind of ignored on purpose during the first stage of TTS research. In this paper, we found out that silent pauses play an important role in spontaneous speech and did further experiments to demonstrate how to use them to improve expressiveness of synthesized speech through perceptual experiments.

Pauses could be divided into 3 categories according to its function in communication, i.e. grammatical pause, physiological pause, and emphasis pause. Grammatical pause is easy to predict based on grammatical structures of a language. However in human communication, pause often conveys paralinguistic information, like the emotional status of the speaker, which depends on the habit, cultural background, and mood of the speaker, and often occurs at an unexpected position or for an unexpected duration. Such kinds of pauses are very difficult for current TTS systems to predict or generate.

For example, in “今天是我妈妈的生日” (Today is my mother’s birthday), a predictable pause would be between “今天是” (today is) and “我妈妈的生日”(my mother’s birthday). However, if the speaker wants to express a kind of emotion like sadness because his mother passed away a year before, he may use a longer silent pause and a lower f0 to utter it. In this article, we made a preliminary study on this kind of pause (a kind of ‘irregular pause’ subjectively defined by comparing it to the normal one) and hope that the results could help to improve the expressiveness, spontaneity and naturalness of synthesized speech.

There have been many studies on pauses in recent years. Yuan Zhao and Dan Jurafsky [1] investigated filled pauses (FPs) of Mandarin, including “zhege”, “nage”, “uh” and “mm” based on a large Mandarin telephony conversational database. They observed longer duration of FP demonstratives than that of non-FP demonstratives, and longer duration of “mm” than that of “uh”. Jean-Leon Bouraoui and Nadine Vigouroux [2] conducted a research with respect to different kinds of disfluency phenomena and made a focused study on the "false starts". Katarina Bartkova [3] analyzed the filled and silent pauses by examining two prosodic parameters, duration of pauses and vowels, and F0 slopes, based on a spontaneous speech corpus in French.

In our research, we defined “regular pause” as a pause occurring in neutral fluent speech; while “irregular pause” often occurs in a ‘wrong’ prosodic boundary level, or with unexpected length, conveying some emotional message of the speaker, which is unlikely to happen in neutral speech. The irregular pauses were categorized into irregular silent pauses (ISPs) of pure silence and irregular filled pauses (IFPs) related to hesitating, thinking, false starting, error correction, etc. The occurrence of IFPs is often linked with personal habit and linguistic background of a speaker [6]; therefore we mainly focus on ISPs which are more related to expressiveness in this study.

In this paper, we analyzed the duration and occurrence of ISPs and their influence to the adjacent speech units in synthesized speech, read speech in neutral and expressive mode and fully spontaneous speech. The paper is organized as following: Section 2 introduces our research corpus, Section 3 gives an overview of characteristics of pauses in different styles, Section 4 investigates acoustic features of ISPs, Section 5 shows perceptual experiment results of introducing ISPs into

synthesized speech, and the conclusion is presented in Section 6.

2. Speech corpus and annotation

The corpus used in this study includes 3 parts: Part 1 is read speech, containing 9 pieces of stories read by a professional actress, using emotional and neutral modes respectively; totally about an hour's speech; Part 2 is synthesized speech from same text; Part 3 consists of 30 pieces of recorded spontaneous expressive dialogues from TV and radio talk shows or interactive dialogue programs with the length ranging from 50 to 80 seconds.

Manual annotation was done using SAMPA-C [4] and C-ToBI [5] for segmental and prosodic features for all speech data. Additionally, two expressiveness parameters, namely valence and activation (i.e. activity or arousal in other contributions), were also manually labeled, to indicate whether an emotion is positive or negative, and the energy level of such emotions respectively.

The activation and valence of an irregular pause here refers to the activation and valence of the adjacent intonational phrase(s). Valence is categorized into three levels: positive (1), neutral (0) and negative (-1). Activation has three categories as well: excited (1), steady (0) and low (-1). When both of the valence and activation of a certain irregular pause are marked as 0, the irregular pause is considered to be a neutral physiological state that does not carry any expressive information. A labeling example is shown in Figure 1.

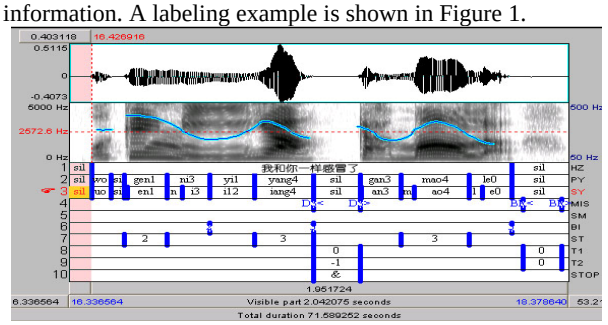

Figure 1: *Annotation Example*

Totally there are ten tiers of annotation in the above example. The 1st tier is the transcription of Chinese characters. The 2nd and 3rd tiers are syllables and initials/finals labeled by Chinese *pinyin* respectively. The 4th tier is the miscellaneous tier for paralinguistic and non-linguistic phenomena. The 5th tier is the turn-taking tier. The 6th tier is the prosodic boundary. The 7th tier is the stress tier. The 8th and 9th tiers are valence and activity states surrounding the silent pause. The last tier is for the irregular silent pauses labeled by "&".

3. Occurrence analysis of silent pauses

In this section, we analyzed the silent pauses in 5 styles of written or spoken materials. Four of them shared the same text material of a fairy tale, including synthesized speech produced by a commercial Mandarin TTS system, pause annotation by 8 transcribers for possible pauses on the text file (no speech), and read speech by an actress in neutral mode and expressive mode respectively. In addition, pure spontaneous speech from TV and radio talk show programs was also analyzed for comparison.

Table 1 gives a summary of silent pauses found in the 5 different types of materials. The meaning of the tokens are: PTTS for pauses in synthesized speech; PTXT for pauses marked in texts by eight transcribers; PNTL and PEXP for pauses in neutral read speech and expressive read speech respectively, and PSPN for pauses in spontaneous expressive speech recorded from TV or radio programs.

The materials based on the same text had a similar number of regular pauses, which is inline with our definition of "regular pause". PTTS and PTXT did not have any irregular silent pauses, which was also reasonable according to our definition. Not surprisingly, 90% of regular pauses in PTTS and PTXT share the same position, and most regular pauses in PTTS, PTXT and PNTL occur at prosodic boundaries.

Among all the silent pauses (both regular and irregular), there were 2.5% irregular silent pauses in neutral read speech, 10% in expressive read speech and 28% in spontaneous expressive speech, which is substantially higher than the other styles, indicating an obvious differentiation of spontaneous expressive speech.

Table 1: *Silent pauses in different speech styles*

Type	#regular pause	#irregular pause	Percentage of irregular pause
PTTS	318	0	0
PTXT	311	0	0
PNTL	317	8	2.5%
PEXP	316	35	10.0%
PSPN	201	78	28.0%

We also found out that when a speaker had a strong emotion towards the topic they was talking about (too excited, too sad, too nervous), or when they made mistakes, irregular silent pauses may occur inside a prosodic word. Otherwise, irregular pauses would still occur at potential prosodic boundaries, but with patterns and acoustic contexts deviated from regular pauses as to be shown in the next section through detailed analysis of the database.

4. Acoustic analysis of ISPs

4.1. Duration analysis of irregular silent pauses

One-way ANOVA analysis was made for duration of the irregular silent pauses in all speech data, using SPSS. Because there might be ISPs occurring one after another, the number of ISPs may vary from experiment to experiment in this section. Results showed that both valence and activation had a significant influence on the duration of the ISPs ($F(2,205)=0.00<0.05$, $F(2,205)=0.00<0.05$ respectively).

Table 2: *ISP average duration in different valences*

停顿时长

Student-Newman-Keuls [a,b]

愉悦度	N	Subset for alpha = .05	
		1	2
0	84	.414645	
1	35	.428315	
-1	89		.675016
Sig.		.871	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 58.013.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Table 3: *ISP average duration in different activations*

停顿时长

Student-Newman-Keuls [a,b]

唤醒度	N	Subset for alpha = .05	
		1	2
1	38	.328962	
0	88	.356929	
-1	82		.804724
Sig.		.711	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 60.151.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

We also computed the mean duration under different valences and activations. The results are illustrated in table 2

and table 3, which indicates that the duration of ISPs is obviously longer when valence or activation is -1 than the other two valence or activation states. N means number of ISPs. The average length of ISPs for 3 valence/activation states was counted in seconds.

4.2. F0 analysis of ISP acoustic contexts

In this experiment, the highest and lowest F0 values of the preceding and following syllables of every ISP in all speech data were analyzed. PSf0H, PSf0L, ASf0H and ASF0L stand for the highest F0 of the syllable preceding the ISP, the lowest F0 of the syllable preceding the ISP, the highest F0 of the syllable following the ISP and the lowest F0 of the syllable following the ISP. The values of f0H and f0L of an ISP are defined by comparing f0 values preceeding and following the ISP, as shown in table 4.

Table 4: *Definition of f0H and f0L*

F0	PSf0H> ASf0H	PSf0H< ASf0H	PSf0L> ASf0L	PSf0L< ASf0L	others
f0H	-1	1	-	-	0
f0L	-	-	-1	1	0

We made ANOVA analysis and found that activation had a significant influence on f0H and f0L (F(2, 98)=0.009; F(2, 109)=0.000) but valence did not (F(2, 98)=0.062; F(2, 109)=0.553 for f0H and F0L respectively). In other words, the highest and lowest values of f0, rise or decline, between every two syllables preceding and following the ISP of interest were significantly different among 3 activation states.

We also computed Df0H=PSf0H-ASf0H and Df0L =PSf0L-ASf0L. Maximum, minimum and mean values of Df0H and Df0L by 3 valence states and 3 activation states were listed in tables 5 and 6.

Table 5: *Maximum, minimum and mean values of Df0H and Df0L by 3 valence states*

		N	Mean	Std. Deviati	Std. Erro	95% Confidence Inte Mean		Minimur	Maximum
						Lower Bou	Upper Bou		
Df0H	-1	91	-19.520	72.2266	7.5714	-34.562	-4.4783	-242.26	128.05
	0	104	-35.721	78.0553	7.6539	-50.900	-20.541	-260.99	127.28
	1	36	-23.624	91.5163	15.2527	-54.589	7.3401	-242.85	163.18
	Total	231	-27.453	78.1174	5.1397	-37.580	-17.326	-260.99	163.18
Df0L	-1	91	-24.030	58.2256	6.1037	-36.156	-11.904	-217.91	117.87
	0	104	-38.860	69.5197	6.8170	-52.380	-25.340	-203.08	117.37
	1	36	-28.077	65.6819	10.9470	-50.300	-5.8535	-145.25	167.13
	Total	231	-31.337	64.7752	4.2619	-39.735	-22.940	-217.91	167.13

Table 6: *Maximum, minimum and mean values of Df0H and Df0L by 3 activation states*

		N	Mea	Std.	Std.	95% Confidence Interval fo Mea		Minimu	Maximu
						Lower Bound	Upper Bound		
Df0	-	8	-	60.220	6.493	-	-	-	112.7
	0	8	-	82.602	8.959	-	-	-	163.1
	1	6	-	92.684	11.867	-	5.317	-	128.0
	Tota	23	-	77.952	5.117	-	-	-	163.1
Df0	-	8	-	55.665	6.002	-	-	-	112.8
	0	8	-	70.668	7.665	-	-	-	167.1
	1	6	-	66.575	8.524	-	-	-	117.8
	Tota	23	-	64.637	4.243	-	-	-	167.1

4.3. F0 values affected by expressive parameters

Another ANOVA analysis was conducted on the f0 values of the syllables preceding and following the ISP affected by expressive parameters of valence and activity.

It can be seen from tables 7 and 8 that valence has significant influence on ASf0H ($p<0.05$) and activation has significant influence on PSf0H, PSf0L, ASf0H and ASf0L.($p<0.05$)

The following four tables 9-12 present the grouping of PSf0H, PSf0L, ASf0H and ASf0L by different activation states. When activation is -1, the values of PSf0H, PSf0L, ASf0H and ASF0L are smaller than those of the other two activation values. The values of ASf0H and ASf0L PSf0H are not discriminative when activations = 1 and 0, but PSf0L is grouped into two categories when activations = 1 and 0.

Table 7: *ANOVA analysis on f0 parameters by 3 valence states.*

		Sum of Squares	df	Mean Square	F	Sig.
PSf0H	Between Groups	25836.37	2	12918.183	2.971	.053
	Within Groups	991299.7	228	4347.806		
	Total	1017136	230			
ASf0H	Between Groups	67322.87	2	33661.434	4.434	.013
	Within Groups	1730792	228	7591.195		
	Total	1798115	230			
PSf0L	Between Groups	3191.041	2	1595.521	.384	.682
	Within Groups	948030.5	228	4158.028		
	Total	951221.5	230			
ASf0L	Between Groups	21801.98	2	10900.988	1.793	.169
	Within Groups	1386234	228	6079.972		
	Total	1408036	230			

Table 8:*ANOVA analysis on f0 parameters by 3 activity states.*

		Sum of Squares	df	Mean Square	F	Sig.
PSf0H	Between Groups	148321.6	2	74160.796	19.546	.000
	Within Groups	868851.6	229	3794.112		
	Total	1017173	231			
ASf0H	Between Groups	111397.2	2	55698.606	7.561	.001
	Within Groups	1687036	229	7366.969		
	Total	1798433	231			
PSf0L	Between Groups	129444.8	2	64722.418	18.032	.000
	Within Groups	821964.0	229	3589.362		
	Total	951408.8	231			
ASf0L	Between Groups	88366.89	2	44183.444	7.667	.001
	Within Groups	1319693	229	5762.853		
	Total	1408060	231			

Table 9: *PSf0H grouping in activation*

PSf0H

Student-Newman-Keuls [a,b]

唤醒度	N	Subset for alpha = .05	
		1	2
-1	86	189.9772	
0	85		234.9502
1	61		249.4247
Sig.		1.000	.149

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 75.403.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Table 10: *PSf0L grouping in activation*

PSf0L

Student-Newman-Keuls [a,b]

唤醒度	N	Subset for alpha = .05		
		1	2	3
-1	86	155.2898		
0	85		192.7332	
1	61			213.2215
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 75.403.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Table 11: *ASf0H grouping in activation*

ASf0H

Student-Newman-Keuls [a,b]

唤醒度	N	Subset for alpha = .05	
		1	2
-1	86	221.0669	
0	85		265.3466
1	61		267.8448
Sig.		1.000	.858

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 75.403.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Table 12: *ASf0L grouping in activation*

ASf0L

Student-Newman-Keuls [a,b]

唤醒度	N	Subset for alpha = .05	
		1	2
-1	86	190.1118	
0	85		230.4310
1	61		230.6417
Sig.		1.000	.986

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 75.403.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

5. Perceptual experiments

In this section, we designed two experiments to see how ISPs affect the perception of expressiveness, by removing the ISPs from expressive spontaneous speech, or inserting ISPs into synthesized speech respectively. In each experiment, we recruited five subjects to perceive the differences of each stimuli pairs, and asked them to tell which one is more natural and expressive.

5.1. Experiment I setup: expressive speech minus ISPs

In this experiment, 5 utterances from an expressive spontaneous dialog formed stimulus A. Text transcription of those utterances is listed in Table 13, in which "&" indicated the occurrence of irregular silent pauses and irregular filled pauses. The activation and valence states of each utterance are

shown in table 14. The counterpart stimulus B was obtained from stimulus A by removing irregular pauses.

Table 13: *Text transcription of stimulus A and B*

No.	Utterance Transcription
1	如果人&可以为了&一点一点小小的进步&就不懈的去努力，那就是最快乐的事情了。
2	但是我没想到大学里面&已经&人烟稀少至此了。
3	我我我一直都&，本来想&，这个问题我&我真的不好&，不怎么好讲。因为&我觉得&（another speaker:哪方面的事情?）我觉得&，我的心态有时候&很不正常的。
4	所以&在有些方面&也不好&跟他&过多的解释。
5	反正&只要一提起他&我都&会比较难受。

Table 14: *Valence and activation states of stimulus A*

Utterance No.	Valence	Activation
1	1	0
2	0	1
3	-1	-1
4	-1	-1
5	-1	0

Tables 15 showed the perceptual results of 5 subjects, in which "1" means utterance in stimulus A is more expressive than that in stimulus B, i.e. the original is more expressive than the one created by removing ISPs; while '0' stands for utterance in stimuli A is less expressive than that in stimuli B. From the table, we could draw a conclusion easily that the expressiveness went down when irregular pauses were removed.

Table 15: *Perceptual result of Experiment I*

Subject No.	Experiment I (stimulus A vs. B)				
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
Sub. 1	1	1	1	1	1
Sub. 2	1	1	1	0	0
Sub. 3	1	1	1	1	1
Sub. 4	1	1	1	1	1
Sub. 5	1	1	1	1	1

5.2. Experiment II setup: Synthesized speech plus ISPs

In this experiment, 5 utterances synthesized by a TTS system formed Stimulus C. The text was from a famous fairy tale called "Little Red Riding Hood", as shown in Table 16. Its counterpart stimulus D was created by inserting ISPs into stimulus C in positions marked by "&" shown in Table 16. Besides, the acoustic parameters of the syllables before and after were modified according to the valence and activation shown in table 17, based on the results from Section 4.

Table 16: *Text transcription of stimulus C and D*

No.	Utterance Transcription
1	可是&她觉得声音怪怪的。
2	您&您的耳朵变得好大哦。
3	这只大野狼&实在太坏太可恶了。
4	猎人说&可以了。
5	猎人、&小红帽、&老婆婆三人&又叫又笑，真是高兴极了。

Tables 18 showed the perceptual results of 5 subjects, in which "1" means utterance in stimulus D is more expressive than that in stimulus C, i.e. the utterance with inserted ISPs is more expressive than the original synthesized speech; while '0' stands for utterance in stimuli D is less expressive than that in stimuli C. It is very obvious that the expressiveness improved a lot after the introduction of ISPs.

Moreover, we also found out that ISPs were more effective to express a kind of negative or depressed emotions, less effective for positive emotions, which was inline with the analysis in the previous section.

Table 17: *Valence and activation states of stimulus D*

Utterance No.	Valence	Activation
1	-1	-1
2	-1	1
3	-1	0
4	0	-1
5	1	1

Table 18: *Perceptual result of Experiment II*

Subject No.	Experiment II (stimulus D and stimulus C)				
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
Sub. 1	1	1	1	1	1
Sub. 2	1	1	1	1	1
Sub. 3	1	1	1	1	1
Sub. 4	1	1	1	0	0
Sub. 5	0	0	1	0	1

6. Conclusions

This paper reported our preliminary study results on how silent pauses were used to express different emotions by speakers. Perceptual experiments have proved that irregular silent pauses play an important role in representation of expressiveness, under different states of valence and activation. Specifically, ISP is effective for rendering a kind of negative or depressed emotion. Acoustically, it affects the acoustic features such as the F0 of its acoustic context, meaning the syllables before and after. We are happy to see the improvement of expressiveness of synthesized speech by the insertion of ISPs and the modifications of F0 of its acoustic contexts.

It would be interesting to see the characteristics of another category, i.e. irregular filled pauses and the influence of irregular pauses upon the overall intonational phrases.

Acknowledgement: This work was supported by the National 863 high-tech project 2006AA01Z138.

7. References

[1] Yuan Zhao & Dan Jurafsky, "A preliminary study of Mandarin filled pauses", *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop*.10–12 September 2005, Aix-en-Provence, France, pp. 179–182.

[2] Jean-Leon Bouraoui & Nadine Vigouroux ,"Disfluency phenomena in an apprenticeship corpus", *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop*.10–12 September 2005, Aix-en-Provence, France, pp. 33-37.

[3] Katarina Bartkova, "How far can prosodic cues help in word segmentation?"

[4] Li, Aijun., "Chinese Prosody and Prosodic Labeling of Spontaneous Speech" In *Proceedings of Speech Prosody*, Aix-en-Provence, France, 2002, pp. 39-46.

[5] Chen, Xiaoxia., Li, Aijun, et. al, "Application of SAMPA-C in SC," *In Proceeding of ICSLP2000*, 2000, Beijing, pp.VI 652-655.

[6] Yuan Zhao & Dan Jurafsky, "A preliminary study of Mandarin filled pauses", *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop*.10–12 September 2005, Aix-en-Provence, France, pp. 179–182.