PERCEPTUAL STUDY OF INTERSYLLABIC FORMANT TRANSITIONS IN SYNTHESIZED V1-V2 IN STANDARD CHINESE

Li, Aijun
Institute of Linguistics
Chinese Academy of Social Sciences
5 JianNeiDaJie, 100732, Beijing, P.R.China
Email: linmc@sun.ihep.ac.cn Tel: 086-010-65237408

ABSTRACT

Transition between vowels is related to speech continuity[8]. Research shows that the formant intensity between syllables varies in Standard Chinese (SC)[5]. We can classify the intensity of intersyllabic formant juncture/transition into three categories from strong to weak by using the consonant of the second syllable. Different categories take various roles in synthesizing speech: the more intense the formant transition is, the more important role it will take in the synthesis. This paper reports the results of perceptual experiments on intersyllabic formant transitions of one of the categories when the second syllable is a zero-initial syllable (i.e. begins with a vowel).

1. INTRODUCTION

Unlike English the number of possible syllables in SC is about 1,200. Most of them are combinations of the sememe and the tone. So many of us considered Chinese as a kind of syllabic language with syllables isolated without no coarticulatory effect between syllables. Therefore, in some old synthesis systems the speech is produced by simply chaining the isolated tonal syllables into a sequence. Because the quality was poor, more and more of us have realized that this point of view is inaccurate. In fact, the intersyllabic coarticulation exists both on the suprasegmental level and on the segmental level, as in other languages. In some recent systems the suprasegmental coarticulations such as tonal coarticulation and duration distribution pattern have been considered, but little has been done on the segmental level. In concatenative or parametric synthesis, it is important to pay careful attention to the segmental transitions which are not only in the syllable but also between syllables.

Suppose CV(N) is a Chinese syllable where C and V(N) stand for the initial and final, respectively, (N)for nasal coda and C1V1(N1) and C2V2(N2) are adjacent syllabic pairs. We can classify the intensity of intersyllabic formant juncture/transition into three categories from strong to weak by using C2: (a) C2 is a zero-initial (i.e. C2=0). (b) C2 is a voiced consonant/initial (e.g.. /m,n,l,r/). (c) C2 is a voiceless consonant/initial (e.g. /p,p',f,x/).

Some studies on transitions[1,6] have shown that suprasegmental coarticulation is more significant than that of the segment in synthesizing SC. Does this mean coarticulation on the segmental level can be ignored? Experiments in [1,6] suggest so by unreliable and incomplete data. There were three limitations in [6]: first the intersyllabic formant coarticulation was not observed classically; second only one word /zuo1yi1/ was used in the perceptual test.; third the transitional duration used was 40ms which was too short to be compared with actual data. Another perceptual experiment of the formant transition[1] was done only on category (c) and the result showed that the intersyllabic formant transition had little effect on the naturalness of the synthesized speech. So, no experiment on category (a) or (b) has been done, which requires a further research to be made as far as the segmental coarticulation is concerned.

In this paper the perceptual effect on intersyllabic formant transitions of the synthesized speech of category (a) is investigated by three perceptual experiments.

2. FORMANT TRANSITIONS OF SYLLABLE PAIR (C1)V1-V2(N2)

There are 21 initials and 38 finals in SC. Usually a syllable is made up of two parts, initial(C)+final(VN). V refers to a monophthong (/a,i,u,o,.../) or a multiphthong (/ai,ao,an,in,iao,iou, ei, uei .../).

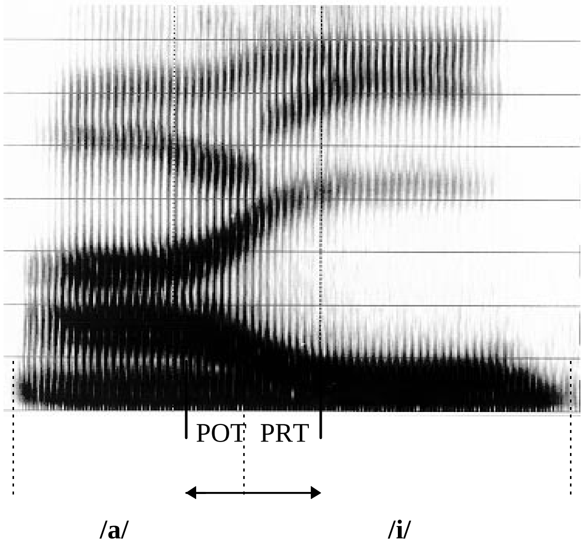
We can see from the spectrograms (Figures 1(a), 1(b)) that intersyllabic transitions of V1-V2 consist of two parts: a post-transition segment in V1 (POT) and a pre-transition segment in V2 (PRT). When coarticulatory place of the final vowel in V1 is the same as that of the beginning vowel in V2, the formant transition curves of F1-F3 are almost horizontal (figure 1(c)). Therefore, the coarticulatory effect can be ignored in this case. Otherwise, very significant transition segments exist as shown in 1(a) and 1(b).

The starting point frequency of the transition approximately equals to the V1's formant target frequency, and the end point frequency equals to V2's onset frequency. The transition curve is a smooth line as for F2 shown in Fig.1(a), or a polyline as for F3 in Fig1(a). The rate of the intersyllabic formant transition changes with different V1V2 syllable pairs[5].

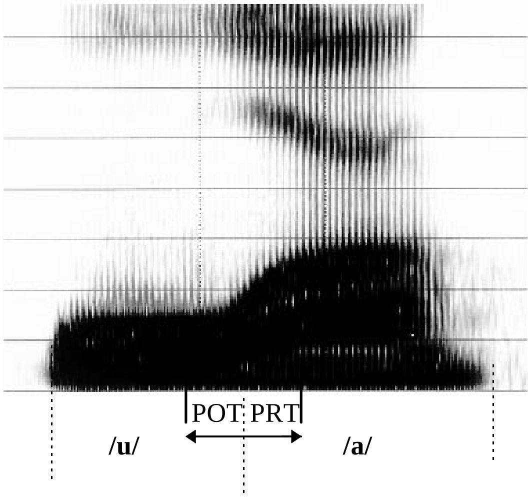
The duration of the transition segment is varied with different V1V2 pairs and different speaking rates. The duration of the transition segment is about 16-30% of the total duration of V1V2. The maximum duration of the intersyllabic transition segment is 180ms at normal speech rate. When we speak faster, the total length of V1V2 is reduced more than that of the transition segment which is about 20-35% of the total duration of V1V2.
The duration ratio to pre-transition in V2 or post-transition in V1 also varies with different V1V2 [5]. For example, 3:2 for syllable pair /i/-/a/, 2:3 for syllable pair/a/-/i/.

3. METHOD AND RESULTS

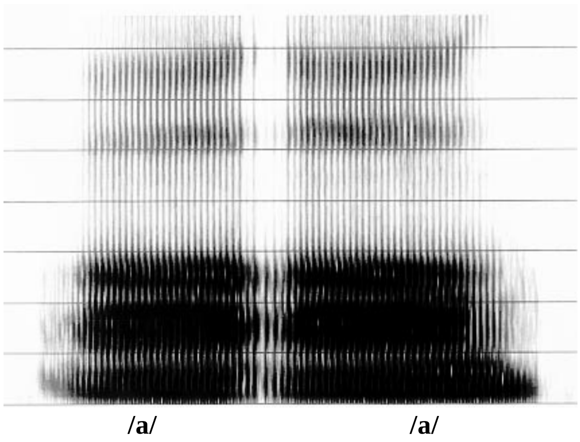
We consider the effect of intersyllabic formant transition on the perception of speech continuity in addition to speech intelligibility. To reduce the effect of the component of "top-down" linguistic processing of the subjects[3] in the experiment, we only use /a/,/i/ and /u/ as V1 and V2 with typical coarticulation to build up 9 meaningless syllabic pairs. After eliminating the pairs in which V1 and V2 have the same vowel types, we have six pairs left: /a/-/i/,/a/-/u/,/i/-/a/,/i/-/u/ ,/u/-/a/ and /u/-/i/.



(a)



(b)



(c)

Fig.1 Broadband spectrogram of (a) /a/-/i/, (b) /u/-/a/ , (c) /a/-/a/.

The stimuli were produced by a cascade formant synthesis system developed in our laboratory[7]. A SIFS (syllable-initial-final segment model) is developed in this system, which separates the syllable into seven segments: (1)silence (2)consonant (3)aspirate (4)pre-transition (5)vowel (6)post-transition (7) nasal coda. A final can takes up segments(4)-(7) and an initial can takes up segment (1)-(3). So, V1 takes up segments (5) + [(6)] and V2 only segment (5) in V1V2 (square bracket stands for an optional item).

V1 and V2 are given the same length to represent that they have the same degree of stress in our experiments. Usually, stress in SC can be classified into contrastive stress, normal stress and weak stress. Only normal stress is considered in this paper. Statistically, the second syllable of a normally stressed bisyllabic isolated word is often spoken more heavily than the first one and is much longer in duration, but when the bisyllabic word is put in running speech, the first syllable is longer than the second one. So the degree of stress of the first and the second syllable are different in different situations. However the results do not conflict with each other if the intonation of the isolated word is considered.

ABX method is used in the experiments. Sound A and B are the same except sound A has no intersyllabic formant transition while sound B has formant transition. A 1.5s pause is inserted between A and B, such AB is presented three times to the listeners via headphones with 3.5s pauses. The three AB pairs form one stimuli set. A silence of 5s is inserted between two sets. Each syllabic pair corresponds to two sets of stimuli: ABX set and BAX set.

Experiment 1 and 2 investigate the perceptual effect of formant transition in different speech rate. In the first two experiments we ignored the tonal effect on the perception of the sounds, so all of the syllables are synthesized with first tones. The perceptual effect of formant transition with different tonal patterns is investigated in Experiment 3.

An answer sheet is provided in three experiments to ten subjects (4 females and 6 males) who do not know phonetics or the intention of the experiments. The order of presentation is random. The subjects were asked to mark the sound in each set that is thought to be more continuous in addition to more intelligible.

Experiment 1

The stimuli consist of 12 sets with first tone+first tone pattern for each syllable pair. Both V1 and V2 are 300ms which includes 150ms transitional duration. The speech rate is 3.3 syllables/second. Transition duration of V1 and transition duration of V2 are assigned as 3:2 in /i/-/a/, 2:3 in /a/-/i/ and 1:1 in other cases. The correct recognition rate of formant transition is 72.1%.

Experiment 2

Here we investigate the perception of the formant transition at a faster rate by shortening the whole V1V2 pair. Both V1 and V2 are 225ms which also includes 150ms transitional duration. The speech rate is 4.4 syllables/second. This reflects that the transitional duration is considerably stable at a faster speed. Other parameters used are the same as in Experiment 1. The correct recognition rate of formant transition is 72.9%.

Experiment 3

Mandarin has four contrastive tones. The numerals 1 to 4 are the traditional nomenclature for these tones adopted in this paper. Suppose the Mandarin third tone sandhi rule is applied in a word i.e. a tone 3 changes to tone 2 when another tone 3 follows. So tone 3+tone 3 changes to tone 2+ tone 3. There are fifteen tonal pairs for each VV pairs.

The fifteen pitch patterns and their perceptual characteristics are described in [4]. Here we use these patterns to synthesize /a/-/i/ and /u/-/a/. The other parameters are as in Experiment 1. Each VV corresponds to 15 sets with random order. Table 1 is the statistic results of the correct recognition rate of V1-V2 transition according to the tones of V1 and V2.

Table 1. The correct recognition rate of formant transition of V1-V2 according to the tones of V1 and V2.

tone vowel	1	2	3	4
V1	**77.5%**	**82.5%**	**47%**	**87.5%**
V2	**67.5%**	**67.5%**	**84%**	**83.5%**

5. DISCUSSION

Eight of the subjects consider that Experiment 2 is easier to judge than Experiment 1 and the other two noticed no differences. Obviously from the results in Experiments 1 and 2, we can see that the intersyllabic formant transition makes a significant effect on the speech continuity, especially for slightly faster speech rate. Therefore, the intersyllabic formant transition should not be ignored in the synthesis. Formant transition is somewhat related to tones. In some cases when V1 has a second tone or a fourth tone or V2 has a third tone or a fourth tone syllable, the formant transition has a greater effect on the continuity of speech.

The results can be extended to continuous speech only if it contains subsegment as (left context)V1-V2(right context) without considering the left or the right context. We can also achieve several synthesis rules where formant transition should be added in synthesizing such a subsegment:

```
if a subsegment (left context)V1-V2 (right context) is to
    be synthesized    //category (a)
{
  if coarticulatory places of V1 and V2 are not identical
     if  (tone of V1 is 1, 2 or 4) or ( tone of V2 is 3 or 4)

          add the intersyllabic formant
          transitions         between V1 and V2;
}
else
  if  C2 is a voiced consonant    // category (b)
     {...}
 else
     formant transitions between V1 and V2
     are  ignored;                   //category (c)
```

5. CONCLUSION

This paper presents three perceptual experiments on V1V2 syllable pairs which show intersyllabic formant transition makes a significant effect on the speech continuity. Combining the results in [1] with the result in this paper, we can say different categories take different roles in synthesizing speech. The more intense the formant transition is, the more important role it will take in the synthesis.

In experiment 2, If we speeded up the speech rate more quickly, we might have gotten a more reliable and a more overall result.

Up to now the intersyllabic formant transition has not been fully investigated such as for ( )N1-( )V2( ) and category (b). Although the acoustic property of the intersyllabic formant transition is very complicated in N1-( )V2( ), the occurrence frequency of this syllabic pattern is very high indeed according to statististics from the intersyllabic triphones in 2-4 syllable word corpus. Some preliminary results on the intersyllabic formant transitions have been gotten recently.

In summary, the principles of classification of the intersyllabic formant transitions are as follows: (1) Connective intensity of the intersyllabic formants. (2) Continuity of the intersyllabic F0. (3) Duration distribution of the transition segment. 4. Perceptual characteristics of the intersyllabic formant transitions, which can affirm the rationality of the categories.

Acknowledgment: I have to say thanks to Professor Lin Maocan who encouraged and guided me to complete the work. Also, I'm grateful for Dr. Marilyn Y.Chen who helps me in editing the paper.

6. REFERENCES

[1] Chu Min, Si Hongyan, Tian Xuqing and Ly Shinan, Perceptual Study on Intersyllabic Transitions on Standard Chinese, Proceedings of the 4th National Phonetic Conference, Beijing, , pp.100-101, 1996.

[2] Feng Long, “Duration of the Initial, Final and Tone in Running Speech of Beijing Dialect”, Working Papers of Beijingese in Experimental Phonetics, Edited by Lin Tao and Wang Lijia, Beijing University Press, pp.131-195, 1985.

[3] John Clark & Colin Yallop, An introduction to Phonetics and Phonology, Blackwell Publishers Ltd,USA, 1995.

[4] Lin Maocan, On the Intersyllabic F0 Transition in Standard Chinese and its Perception, Chinese Journal of Social Sciences, Vol. 17, No.4, 1996.

[5] Report of Phonetic Research, Institute of Linguistics, Chinese Academy of Social Sciences, 1992-1993,1994-1995.

[6] Tang Difei, Ly Shinan, Zhou Tongchun and Wang Renhua, The Preliminary Study on Coarticulation Rule in Chinese Speech Synthesis, Proceedings of the 6th National Conference on SICS'93, pp.75-78.

[7] Yang, Shunan & Xu Yi, An Acoustic-Phonetic Oriented System for Synthsizing Chinese, Speech Communication, Vol.7, No.3, ,pp.317-325,1988.

[8] Y. R. Chao, Yuyan Wenti (Language Problems), Shang Wu Press, Beijing, China, 1980.