SYLLABLE CLUSTERING AND SPECTRAL DISCONTINUITY IN SYLLABLE-BASED TTS SYSTEMS

Fangxin Chen

IBM China Research Laboratory
e-mail: chenfx@cn.ibm.cn

ABSTRACT

This research examined the spectral discontinuity problem existing in syllable-based Chinese TTS systems. Acoustic and phonetic investigations showed that syllables with approximant, nasal or vowel as onset had tendency in forming *syllable clusters* with their preceding syllables in natural speech due to the strong co-articulation effect. In speech synthesis, *syllable clusters* are the major source for the audible spectral discontinuity. The implication of this finding for improving syllable-based TTS voice quality was discussed.

1. INTRODUCTION

One misconception on spoken Chinese is that Chinese is a mono-syllabic language, and its phone co-articulation at syllable boundaries is weaker as compared with other languages. This misconception could originate from the Chinese orthography. In written Chinese, each Chinese character corresponds to a syllable, and in most cases each syllable has independent lexical meaning. There is also no word boundary in Chinese text. Written Chinese gives the false impression that spoken Chinese works the same way. A careful acoustic analysis of spoken Chinese, however, shows no difference from other languages with respect to phone co-articulation effect at the syllable boundaries.

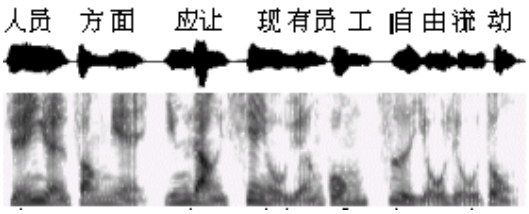


Figure 1. The waveform and spectrogram of the sentence *Renyuan fangmian ying rang xianyou yuangong ziyou liudong* spoken by a female speaker.

It can be observed from above figure that syllables like 人员 *(renyuan)*, 方面 (*fangmian)*, 应让 (*yingrang)*, 现有员 (xianyouyuan), 自由谁 (ziyouliu) are all acoustically clustered togethr due to the strong phone co-articulation effect. Syllable clustering in natural speech generates chunks of speech which look more like inseparable acoustic entities rather than sequences of individual syllables. For easier discussion, this paper introduces the term *syllable cluster as* a sequence of two or more syllables which

- does not have distinctive acoustic boundaries in separating the adjacent syllables;
- has relatively clear formant tracks between the syllable boundary phones;
- has relative high energy level at the syllable boundaris.

A popular approach in Chinese TTS is using syllable as the basic unit for concatenation. Prefering syllable, rather than other smaller speech units for Chinese could possibly for easier handling of the lexical tones. However, syllable approach requires much larger speech corpus. Otherwise there could be more spectral discontinuity problem happening in synthesized speech because of the insufficient coverage of the coarticulation between the syllable boundaries. As observed, the spectral discontinuity noises in a syllable-based system mainly happen in syllable clusters. On the one hand, there is no identifiable syllable boundary in the syllable clustering situation, which could introduce inconsistency in syllable labelling in the speech corpus, either with automatic or manual method. On the other hand, syllable(s) directly cut from syllable clusters are heavily phonetic-context-dependent, and the energy level at the clustered syllable boundary is relatively high, which causes serious spectral discontinuity when they are concatenated with syllables not from the corresponding phonetic context. Following is an illustration of the spectral discontinuity problems observed in three commercially available Chinese TTS systems, which are

named here as SA, SB and SC just for convenience. All those three systems are syllable-based.

Natural speech is used here as baseline for comparision. Figure 2 shows the waveform and spectrogram of the Chinese word 畢業生 taken from a sentence spoken by a female speaker. It can be observed from Figure 2 that the syllables *bi* and *ye* are clustered together and there are observable formant tracks at the boundary of those two syllables, while the syllables *ye* and *sheng* are acoustically separated.

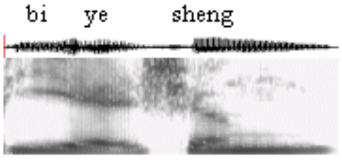
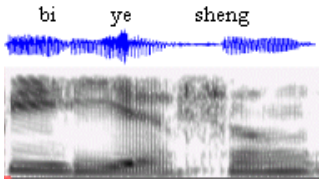


Figure 2. The waveform and spectrogram of the word 畢業生(biyisheng) taken from a sentence spoken by a female speaker.

Figure 3 shows the word 畢業生 taken from the same sentence synthesized by the SA system. There was audible noise existing in that word.



Figue.3. Waveform and spectrogram of the word 畢業生(biyisheng) taken from a sentence synthesized by SA system.

The spectrogram shows that the formant tracks(especially the first formant) are discontinued at the boundary of *bi* and *ye*. The syllable *bi* was obviously not followed by any sonorant phone in the original speech data, which could be told from the natural enery tailing at the offset of the syllable, while *ye* was directly cut from a syllable cluster in the data. The significantly reduced energy at the offset of *bi* and the abrupt rising energy at the onset of the syllable *ye* introduced the audible spectral discontinuity noise.

Figure 4 shows the word 畢業生 taken from the sentence synthesized by the SB system. SB treated all syllable tokens as independent speech units. From the spectrogram we can see the clear spectral gaps at each syllable boundaries. Perceptually the naturalness of speech in syllable cluster situation was lost.

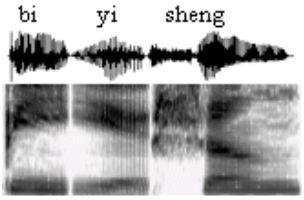
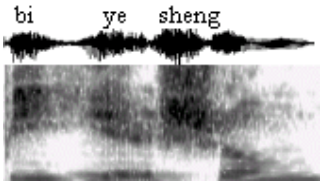


Figure 4. The waveform and spectrogram of the word 畢業生(bi4yi4sheng1) synthesized by SB system.

SC system is an interesting system to observe. For the syllable cluster *biye* in the word 畢業生, it synthesized perfectly (see Figure 5). Obviously, SC speech corpus contained the word
畢業生 and its un-uniform speech unit search algorithm allowed the selection of the whole word directly from the speech corpus.



Figue 5. Waveform and spectrogram of the word 畢業生(biyisheng) synthesized by SC system.

However, the non-uniform search algorithm is only helpful when the exact target syllable cluster exists in the speech corpus. A syllable cluster could contain more than two syllables, as long as the syllables in the sequeence are all strongly co-articulated. If the corpus does not contain the exact same syllable cluster as the target one (either more or less syllables included in the target syllable cluster), the synthesized speech could still have the spectral discontinuity problem. Figure 6 shows the phrase 语音合成 taken from a synthesized sentence by the same SC system, which had audiable click noise between the syllable boundary of *yin* and *he*.

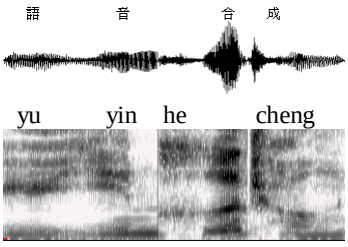


Figure 6. the waveform and spectrogram of the phrase 语音合成 (yuyinhecheng) by SC system.

It can be observed from the waveform that the syllables *yu* and *yin* were taken from the orignal syllable clusters in the speech data. But the abrupt energy level change at the ending of *yin* indicates that

that syllable was also strongly co-articulated with its following syllable in the original corpus. The partially selected syllable cluster created audible spectral distortion at the syllable boudary of yin.

The solution for reducing the spectral discontinuity induced by syllable clusters for the syllable-based concatenative TTS system, then, is to find out:

- What is the articulatory/phonetic nature of the syllable cluster?
- How to maximally include the syllable clusters in the speech corpus design?
- How to design the unit-search algorithm for better synthesis of syllable clusters with the existing speech data?

2. METHODS

The main question investigated in this paper was: what is the articulatory/phonetic nature of syllable cluster in spoken Chinese? The focus was on the phone category effect at the syllable boundaries. As known, Chinese syllable structure is relatively simple if lexical tone is not taken into consideration. There are only 23 vowels(include diphthongs) as well as three voiced consonants [n, G, r] which could be in the syllable final position. For the syllable onset, there are 22 consonants, which could be grouped into seven categories: unaspirated stops, aspirated stops, unaspirated affricates, aspirated affricates, fricatives, approximants, nasals. Since vowels could also be in the onset position, the total nubmer of possible phone category combinations at syllable boundaries in spoken Chinese can be grouped into 208.

The speech data used for investigation was a speech recording corpus of 2000 sentences spoken by a female professional speaker in normal speed (4~5 syllables/sec). The recording was conducted in a sound-proof room and sampled at 22k. The 208 possible phone combinations at the syllable boundary were searched from that corpus, and the waveforms and the spectrograms for each case were examined according to the acoustic criteria set for the syllable cluster. Because there was no combination of retroflex and vowel found at the syllable boundary in the data, the analysis for that case was omitted. The finding based on the acoustic analyses of 207 phone category combinations was that: the following phone category combinations at the syllable boundary have strong tendency to cluster the two adjacent syllables together:

- Vowel + Vowel (V+V):
- Vowel + Approximant (V+A);
- Vowel + Nasal (V+N)
- Nasal + Vowel (N+V);
- Nasal + Approximant (N+A);
- Nasal + Nasal (N+N);
- Retroflex + Approximant (R+A);
- Retroflex + Nasal(R+N).

Figure 7 are the examples of observed syllable clusters from natural spoken Chinese in the speech corpus for each of above listed phone category combinations.

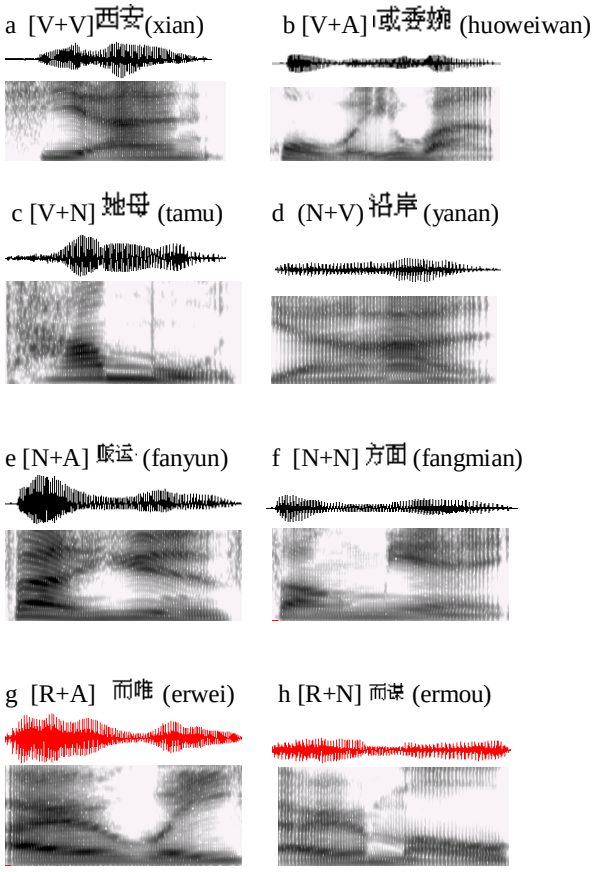


Figure 7. Syllable clusters of different combintion of phone categories.

3. DISCUSSION

The investigation shows that syllable cluster is articulatorily or phonetically conditioned. All the phone combinations which trigged syllable clustering are sonrant sounds, such as vowels, nasals and approximants. Acoustically, sonorant phones have relatively regular formant structure and higher aoustic energy. Articulatorily, sonorant phones are all voiced. The vocal folds keep vibration during the phonation of a sequence of sonorant phones and the articulators in the oral cavity moves from

the current phone target to the following one, which generates strong co-articulation effect.

For syllable onset with voiceless stops and affricates, the initial closure and ensuing burst disrupts the formant tracking between the adjacent syllables. Consequently, syllables with stop or affricate as onset do not cluster with their preceding syllables. There usually exists clear boundary between a syllable and its following syllable which has stop or affricate as its onset due to the silence introduced by the closure.

Syllables with fricatives as onset have relatively weak co-articulation with their preceding syllables because of the sustained low-energy frication noise, and the syllable boundaries with their preceding syllables are relatively easy to identify.

In spoken Chinese, only vowel, nasal, approximant and retroflex trig strong syllable clustering, and for Chinese syllable, only vowel, nasal and retroflex could be in the syllable final position. This results that the syllable boundary phones at leftside always satisfy the syllable clustering condition. What really matters in syllable clustering for Chinese, then, is the onset of the rightside syllable. Any rightside syllable with vowel, nasal or approximant as onset is very likely to cluster with its preceding syllable.

The syntactic structure in which the syllables are embedded does not play major role in syllable clustering. For example, in natural speech of 三峡文化 (*sanxiawenhua*) by the female (see Figure.8), the syllables clustered together is *xiawen*, rather than *sanxia* and *wenhua*. Syntactically, *sanxia* and *wenhua* are two independent words. The only reason for *xia2* and *wen2* clustering together is because the both phones at the syllable boundary belong to the sonorant phone category.

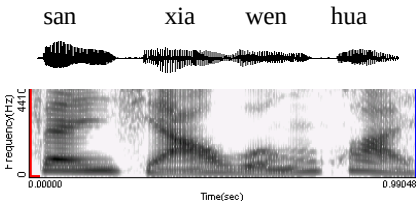


Figure 8. Waveform and spectrogram of the phrase 三峡文化 by a female speaker

Howerver, there are cases in speech data that the syllables did not cluster together though they met with the phone category combination condition. For example, in the phrase 她傲雪斗霜报春归的精神 *ta aoxue doushuang bao chungui de jingshen*, *ta* and *ao* were not clustered together because the speaker had obvious intention to put stress on *ta* when reading the phrase. This suggests that phonetic condition for syllable clustering is a default articulatory mechanism. But the syllable clustering could be prevented by the speaker's intention to emphrase on a particular character (ie. a syllable), or indicate an important syntactic boundary.

Syllable cluster is in nature a phone co-articulation phenomenon. The reason we discussed only on syllable cluster in this paper is that the original intended research was looking for solution to the spectral discontinuity problem in the syllable-based TTS system. In that case, the intra-syllable phone clustering is not an issue. What matters is the phone cluster at the syllable boundaries. For TTS systems using sub-syllable unit for concatenation, it is then necessary to expend the discussion to phone clustering at their sub-syllable unit boundaries. The report on human detection of concatenation discontinuities in diphone-based synthetic speech [Syrdal] is consistant with the finding in this paper. According to that report, postvocalic sonorants introduce more audible concatenation discontinuities than the postvocalic non-sonorants. This is in fact the phone clustering problem as implicitly discussed in this paper.

4. CONCLUSION

Syllable clusters contain strong formant transitions between the syllable boundary phones and they are the major cause for the audible discontinuity problem in syllable-based TTS systems. Syllable clusters are phonetically conditoined and predictable. For spoken Chinese, syllables with approximant, nasal or vowel as their onset tend to cluster with their preceding syllables.
For the syllable-based TTS systems, the finding in this research suggests that :

- We need to get statistics from a comprehensive text corpus for the frequently used syllable clusters and have the speech corpus well designed to cover those frequently used syllable clusters as fully as possible;

- Dynamic unit search algorithm in synthesis stage should be implemented in such that those syllable clusters in speech recording be treated as inseparable speech segments. Particially selecting syllable(s) within a syllable cluster be prevented to avoid spectral discontinuity in the synthesized speech.

5. REFERENCE

A. Syrdal, Phonetic Effects on Listener Detection of Vowel Concatenation, ICASSP 2001.http://www.research.att.com/projects/tts/pubs.html