A novel set of synthesis units with stable spectral boundaries for HMM-based Mandarin speech synthesis system

Yishan Jiao [1], Xiang Xie, Ming Tu, Xingyu Na

Abstract. Co-articulation is a common phenomenon in human speech, which guarantees the speech sound coherent and natural. Synthesized speech, however, often sounds artificial. This is somewhat because of its inability to imitate co-articulation well. This paper defines a novel set of synthesis units to preserve both intra- and inter-syllable co-articulation. The boundaries of the new unit are located respectively at each essential vowel of two adjacent syllables. It consists of three parts: final-tail of the preceding syllable, initial consonant and final-head of the following syllable so that we call it Nal-Initial-FI (NIF) unit. To locate the boundaries, we adopt the maximum spectral stability criterion. It can find out the most stable point within the essential vowel. In the experiment, we test NIF units on the HMM-based speech synthesis system (HTS) and compare the result to the syllable unit system. The Preference test and the Comparison Category Rating (CCR) test show that the speech synthesized with NIF units has better naturalness than that with syllable units, and the speech quality of both systems is comparable.

Keywords: synthesis unit • co-articulation • maximum spectral stability criterion • Mandarin speech • HTS

1 Introduction

Selecting a proper set of synthesis units is crucial to the speech synthesis system. The length and structure of synthesis units concern the quality and naturalness of synthesized speech, as well as the size of database and the complexity of synthesis system. Optional units are words, syllables, initial-finals, phonemes, etc. In Mandarin text-to-speech (TTS) system, syllables are often used as synthesis units. The reason is that they are the most natural and basic units for pronunciation, and an

[1] Yishan Jiao (✉)
Beijing Institute of Technology, Haidian district, Beijing, China
e-mail: alicechiao13@gmail.com

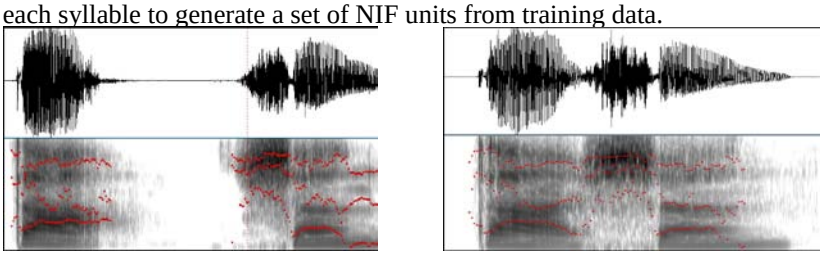
This work is supported by the National Natural Science Foundation of China (Grant No. 61001188, No. 11161140319, No. 90920304 and No. 91120015).

individual syllable is relatively stable in acoustic [1]. However, to keep certain prosodic features, there is usually no pause between syllables in natural and understandable speech. In fact, each pronouncing unit is not isolated and could be affected by its context, which leads to co-articulation.

In Mandarin, there is co-articulation within a syllable called intra-syllable co-articulation. For example, in the syllable "bai", the vowel "a" affects the articulation of the consonant "b". However, in a continuous speech, there is also co-articulation between syllables called inter-syllable co-articulation. Fig. 1 is an example of this kind of co-articulation, which shows how the following fricative "sh" affects the formant tracks of the preceding vowel "a", especially its second formant (F2). Therefore, the traditional syllable units which can only preserve intra-syllable co-articulation are not perfect enough to generate natural speech due to its overlook of the inter-syllable ones.

In concatenative speech synthesis system, there were some techniques proposed many years before which tried to include inter-syllable co-articulation into a synthesis unit, such as diphone [2] and demisyllable [3]. But in recent decades, HMM-based speech synthesis system (HTS) has gradually replaced concatenative approach. However, the basic synthesis units in HTS are often phonemes and syllables, which could only preserve part of co-articulation in continuous speech. Although embedded trainings for context-based models in HTS have attempted to consider inter-syllable co-articulation, the unit itself still cannot include such information. Therefore, it is impossible for HMMs to model inter-syllable co-articulation directly. Based on the above consideration, we define a novel set of synthesis units which can preserve both intra- and inter-syllable co-articulation.

To generate the new units, we need to find out the boundaries which do not affect co-articulation in syllables. In fact, there is a relatively stable period in a Mandarin syllable called essential vowel. Although some parts of the syllable can be affected by co-articulation, there is always a stable pronunciation period in the essential vowel. It can keep the meaning of the syllable consistent in different contexts. In this paper, maximum spectral stability criterion is adopted to find out the most stable spectral boundary in each essential vowel. This criterion was first used in a modified temporal decomposition (TD) algorithm to compute the number and location of event functions. Here we develop it to a four-step procedure and use it for each syllable to generate a set of NIF units from training data.



(a) "da-shan" pronounced seperately (b) "dashan" pronounced continuously

Fig.1 Inter-segmental co-articulation. Red points depict tracks of the first three formants.

The organization of this paper is as follows. Section 2 describes the definition of NIF units and the maximum spectral stability criterion. Experimental setup and the analysis of the results are presented in Section 3. In Section 4 we conclude our work and make further plans.

2 Nal-Initial-Fi units with stable spectral boundaries

2.1 Definition of Nal-Initial-Fi units

Mandarin is a monosyllable language with four tones. Most of Mandarin syllables are composed of an initial and a final. Some of them do not have the initial part, and the final part also has a variety of forms, such as monophthong and diphthong. But every Mandarin syllable must have an essential vowel and can only have one so that we can annotate the tone on it. There are six vowels which can be the essential vowel: “a”, “o”, “e”, “i”, “u”, and “ü”. Even if other parts of the syllable can be affected by co-articulation, there is always a stable period in the essential vowel to keep the meaning of the syllable consistent when it is in different contexts.

Therefore, we set the essential vowel as the breakpoint and segment a continuous speech into a novel set of synthesis units. Each of the units consists of three parts: the final-end of the preceding syllable, the initial consonant and the final-head of the following one. In this definition, final-end is defined as the essential vowel plus the remaining part of the syllable, while the final-head is defined as the final of a syllable up to the essential vowel. For example, in the Mandarin syllable “lián”, the essential vowel is “á”, the initial is “l”, the final-head is “iá”, and the final-end is “án”. Fig. 2 shows the structure of a set of NIF units compared with syllable units. The word “sil” represents the short silent period at both ends of a continuous sentence.

In a common Mandarin database, the number of NIF units without tone is about 1750, which is much larger than 400 toneless syllables. Although it might spend more time to initialize HMMs for NIF units, the numbers of context-based models after clustering are approximately equal (It can be seen from Table 2 in Section 3.2). Thus it does not cause dramatic increase of the computational complexity and storage space.

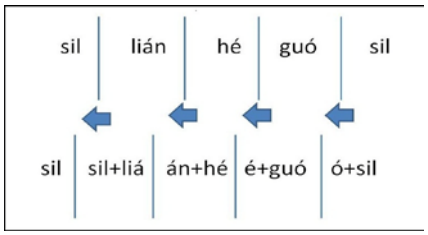


Fig.2 NIF units (the lower) V.S syllable units (the upper). The meaning of this sentence is “the Unitied Nations”

2.2 Maximum spectral stability criterion

To locate the boundaries in the essential vowel exactly, we adopt a method called maximum spectral stability criterion. It was first used in one of the modified temporal decomposition (TD) algorithms to initialize the number and locations of event functions [4]. Since TD is able to determine phonetic events and the maximum spectral stability criterion can ensure the stability of the events, it is supposed to be an effective way to find out the most stable point in the essential vowel.

In this paper, we use this method for syllables in the training speech one by one. To avoid error locating on the pseudo stable points in some initial consonants, we only find out the maximum spectral stability point in the final. This means that the training corpus should have elaborate labels corresponding to each syllable and each phoneme.

The description of this method is as follows. First, the spectral parameter sequence of a syllable is extracted from the original speech. Then the spectral transition rate of the i^{th} spectral parameter at frame n, $LSF_i(n)$, is calculated as the gradient of the best fitting straight line, i.e. regression line, within the time window [n-M, n+M], as given in equation (1). Finally, the mean square of a_i, where i = 1…P, is defined as the spectral transition measure (STM) at frame n, and is given by equation (2).

$$a_i = \left(\sum_{m=-M}^{M} m \cdot LSF_i(n+M)\right) / \left(\sum_{m=-M}^{M} m^2\right), \quad 1 \le i \le P \tag{1}$$

$$STM(n) = \left(\sum_{i=1}^{P} a_i^2\right) / P \tag{2}$$

STM is applicable for any spectral parameter, but the STM of line spectral frequency (LSF) is used in this paper due to its close relations with formant frequency. The local minimum of STM(n) indicates the point with maximum local spectral stability. When we constrain the local area in the final of a syllable, the local minimum of STM(n) corresponds to most stable spectral point in the essential vowel. However, the number of local minima points varies according to the window size M. Thus, in order to find out the only boundary in a syllable, we develop a four-step iterative algorithm to estimate this location:

- **Step 1.** Compute the STM of a specific syllable.
- **Step 2.** Initialize the window size M = 2.
- **Step 3.** Detect local minima of STM within the final part. If there is 1 minimum, return this location, if else, go to Step 4.
- **Step 4.** If there is more than 1 minima, set M = M+1, and return to step 3. If there is no minima or M>8, the location of the essential vowel is determined as the central of the final.

3 Experiments

3.1. Experimental setup

The database we used is ASCCD corpus [5]. It contains a set of syllable-balanced Mandarin speeches uttered by 3 female speakers (F002, F003, and F004). The number of training sentences was 500 for each speaker, and their contents were the same. Every speaker had another 100 sentences for evaluation, which are not included in the training data. The length of each sentence was about 8 seconds in average.

To implement the maximum spectral stability criterion, we should first know the boundary of syllable and the segment point between initial and final. For ASCCD corpus, each sentence has a TextGrid file, which segments the speech into various pronouncing units, such as initial and final, syllable, words and phrases. And every unit has a label corresponding to its content. Therefore, with the aid of TextGrid files, all of the boundaries were computed automatically with the maximum spectral stability criterion. Between each two boundaries was a cross-syllable NIF unit. And it could be easily labeled according to the definition in Section 2.1. Note that although the jointer between two semi-syllables was "+" used in figure 2, in the real label files it could be any character as long as it is different from other symbols in the full-context labels. In our experiments, it was set to "0".

Speech signals used for training were sampled at 16 kHz and windowed by a 25-ms Hamming window with 5-ms shift. The parameter vectors consisted of 25 Mel-cepstral coefficients including the zeroth coefficient, log F0, and their delta and delta-delta coefficients [6]. They were all obtained by using STRAIGHT [7]. 10-state left-to-right context-dependent HMMs without skip paths were used to train acoustic models. Each state had a single Gaussian pdf (probability density function) as the state output probability and another single Gaussian pdf as the state duration pdf.

To carry out tree-based context clustering and make comparison with syllable system, we developed a set of phonological and lexical information for NIF units referring to syllable units. Table 1 shows the contextual factors of NIF and syllable unit. The decision tree-based context clustering technique was separately applied to distributions for F0, spectrum and state duration. In our experiment, minimum description length (MDL) criterion [8] was used to stop tree growth.

Table 1 Contextual factors of NIF unit and syllable unit

	NIF unit	*Syllable unit*
Unit Layer	Current, preceding and following units;	Current, preceding and following syllables;
Tone, Initial & Final	Tone of current unit's previous and following syllables;	Tone of current syllable and its preceding and succeeding syllables;
	Initials and finals of current unit's previous and following syllables;	Initial and final of current syllable and its preceding and succeeding syllables
Prosodic Structure	Number of syllables, words and phrases in utterance;	Number of syllables, words and phrases in utterance;
	Number and position of current unit's following syllable in word, phrase and utterance counting from forward and backward;	Number and position of the current syllable in word, phrase and utterance counting from forward and backward;
	Number and position of current words in phrase and utterance counting from forward and backward;	Number and position of current words in phrase and utterance counting from forward and backward;
	Number and position of current phrases in utterance counting from forward and backward	Number and position of current phrases in utterance counting from forward and backward;

3.2 Results of evaluations

The total number of leaf-nodes clustered for each parameter is shown in Table 2. From the table, we can see that the numbers of leaf-nodes after clustering training are almost the same for the two systems, thus we can compare their performance together.

To evaluate the naturalness of synthesized speech, the Preference test was conducted. 8 subjects were asked to listen to 30 pairs of sentences (3 speakers $\times$ 10 test sentences) chosen from the synthesized speech in random order and give their preference on naturalness. Figure 3 shows the results of the Preference test. It can be seen that the speech synthesized with NIF units has better naturalness than that with syllables.

Table 2 Number of leaf-nodes of constructed decision trees for logF0, spectrum, and duration.

Speakers	*Methods*	*LF0*	*Spect.*	*Dur.*
F002	Syllable	2545	926	110
	NIF	2493	990	94
F003	Syllable	2215	878	120
	NIF	2239	966	89
F004	Syllable	2032	1036	103
	NIF	1938	968	89

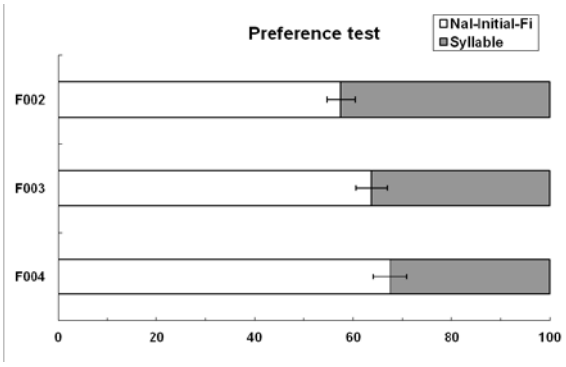


Fig. 3 Preference test between NIF and Syllable system for 3 different speakers with 95% confidence interval.

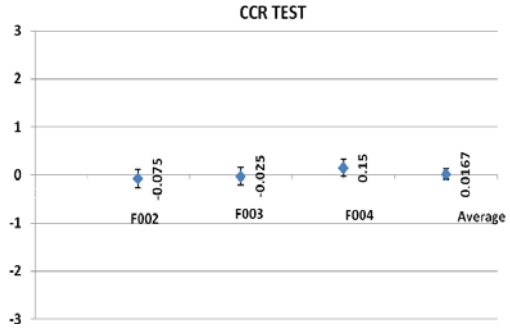


Fig.4 Evaluation result for the quality of speech synthesized with NIF units. Positive value means it is better than syllable unit.

To compare the quality of speech synthesized with these two systems, we performed Comparison Category Rating (CCR) test. The number of speech samples were 60 (30 pairs) picked randomly from the two systems. Another 8 listeners were told to score the quality of the second one compared to the quality of the first. Note that the order of each pair was unknown to the listeners. There were 7 ranks: 3 means Much Better, 2 means Better, 1 means Slightly Better, 0 means

About the Same, -1 means Slightly Worse, -2 means Worse, and -3 means Much Worse. The results for each speaker and the average score have been presented in figure 4. From the figure, we can see that the quality of speech synthesized with NIF units is comparable to that with syllable units.

4. Conclusions

In this paper, we have defined a novel set of synthesis units called NIF. It could preserve both the intra- and inter-syllable co-articulation. To locate the boundaries of each unit, the maximum spectral stability criterion has been adopted to find out the most stable point in the essential vowel. To evaluate the performance of NIF units, we have tested them in HTS and compared the results to the syllable system. The preference test has shown that NIF unit could improve the naturalness of synthesized speech, and the CCR test has shown that the quality of the speech synthesized from the two systems is comparable.

In the future, we will try to develop an unsupervised approach to locate the boundaries of NIF units rather than with the help of detailed TextGird files.

References

1. X. J .Yang, and H. S. Chi, "Digital Speech Signal Processing", Publishing House of Electronics Industry, pp. 301–319, 1995.
2. R. Dixon and H. Maxey, "Terminal analog synthesis of continuous speech using the diphone method of segmenta assembly," IEEE Trans on Audio and Electroacoustics, Vol. AU-16, NO. 1, pp. 40-50, Mar 1968.
3. J. B. lovins, M. J. Macchi and O. Fujimura, "A demisyllable inventory for speech synthesis," in Speech Communication Papers, presented at the 97th Meeting of the Acoustical Society of America, J. J. Wolf and D. H. Klatt (eds.), pp. 519-522, 1979.
4. Nandasena A. C. R, M. Akagi, "Spectral stability based event localizing temporal decomposition," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2: 957-960, 1998.
5. Phonetics Lab, Institute of Linguistics, CASS, ASCCD: Read Discourse Corpus with prosodic, segmental and syntactic annotation, http://ling.cass.cn/yuyin/english/resc6.htm.
6. K. Tokuda, T. Kobayashi, and S. Imai, "Speech Parameter Generation from HMM Using Dynamic Features," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 660–663, May 1995.
7. H. Kawahara, "STRAIGHT, Exploration of the other aspect of VOCOODER: Perceptually isomorphic decomposition of speech sounds," Acoust. Sci & Tech., 27(6), 349-353, 2006.
8. K. Shinoda and T.Watanabe, "MDL-based Context-dependent Subword Modeling for Speech Recognition," J. Acoust. Soc. Japan (E), vol. 21, pp. 79–86, Mar 2000.