SENTENCE DESIGN FOR SPEECH SYNTHESIS AND SPEECH RECOGNITION DATABASE BY PHONETIC RULES

Zu, Yiqing
Institute of Linguistics
Chinese Academy of Social Sciences
5 Jian Nei Da Jie, 100732, Beijing, P.R.China
Tel. 08601065237408 E-mail: linmc@sun.ihep.ac.cn

ABSTRACT

This paper describes the processing of 2465 sentences (or utterences) which are collected by phonetical rules from a big corpus--recent years' newspaper, "People's Daily" and etc., as materials of speech recognition and speech synthesis database. In these sentences, both phonetic phenomena and sentence patterns are included. We first consider the phonetic distribution among syllables: inter-syllabic diphones, inter-syllabic triphones and final-initial structure. The syllabic balance ensures the intra-syllabic phenomena such as phonemes, initial/final and consonant/vowel. There are roughly 17 kinds of sentence patterns which appear in our sentence set. We have also created a set of phonetically balanced 2-4 syllable phrases which includes all of the tone structures.

1. INTRODUCTIOIN

Since speech recognition and speech synthesis have moved into large vocabulary continuous speech, higher quality, scientific designed, succinct and valid continuous speech database is needed. The complex phenomena due to variants bring difficulties to speech engineering. We think that at the first stage we should consider contextual variant in speech and the speech database should be mainly limited in the segmental aspect of read speech. The material of speech recognition and speech synthesis database should be phonetically compact with low redundancy [1]. In continuous speech, phonemes always appear in the form of allophones which results in coarticulatory effects. The task of designing a speech database should be concerned with various intra-syllabic and inter-syllabic allophone structures. According to the studies on formants transitions between two syllables in read speech of Mandarin by Chen, X. X.[2] and other researchers[2,3,4,5], with 401 syllables without tone information (commonly assumed for Mandarin), there are 415 inter-syllabic diphones, 3035 merged inter-syllabic triphones and 781 merged final-initial structures.

2. THE PHONETIC KNOWLEDGE GUIDING SPEECH MATERIAL DESIGN

2.1 Variations in continuous speech

Variants in continuous speech is the fact that phonetic elements deviate from orthography. In segmental aspects there are two kinds of variants: (1) contextual: variation of speech units in different context; (2) un-contextual: the variation come from speech rate, speech mood, sentence pattern, different speakers and so on. In super-segmental aspects, the variations in tones, duration, energy, and inter-segmental influence result in variants.

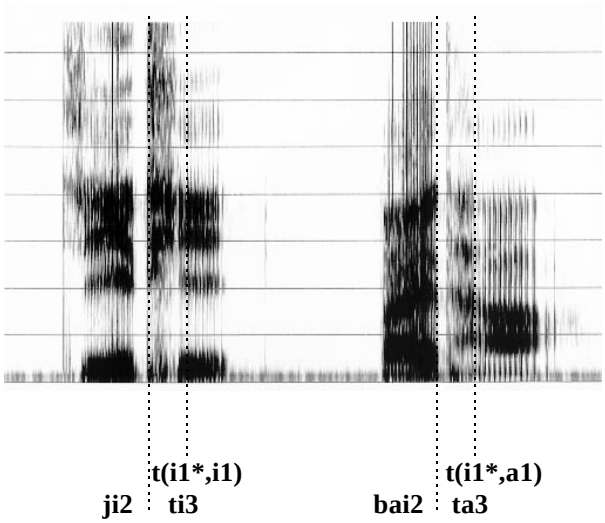
2.2 The basically phones in Mandarin

To study the phonetic phenomena we should provide the basic phonetic elements. Phone is the smallest element of the syllable. There are 37 basically phones:
{a1,a2,a3,b,c,ch,d,e1,e2,e3,er,f,g,h,i1,i2,i3,j,k,l,m,n,ng, o1,o2,p,q,r,s,sh,t,u,x,yv,z,zh,sil }
Where "sil" is silence. The following lists contexts for some vowel phones:

a1: ba e1: he, ge
a2: an, ai e2: ei, ye, yue
a3: ao, ang e3: en, eng

i1: bi, yi o1: wo
i2: zi, ci, si o2: ou
i3: zhi, chi, shi

An utterance of continuous speech is structured by a series of syllables, and each syllable is made of a series phones. In continuous speech, as a phenomena of articulation, every phone exists in the form of its allophones. The distinction between continuous speech and isolated word or connected word is that articulation phenomena dose not only exist between syllables but also between phrases[6]. All those syllable are quite different from isolated syllables. Fig. 1 shows the sonagraph of words (ji2 ti3, bai2 ta3, zi3 nv3, ji2 na2) with /t/ and /n/ in different contexts.



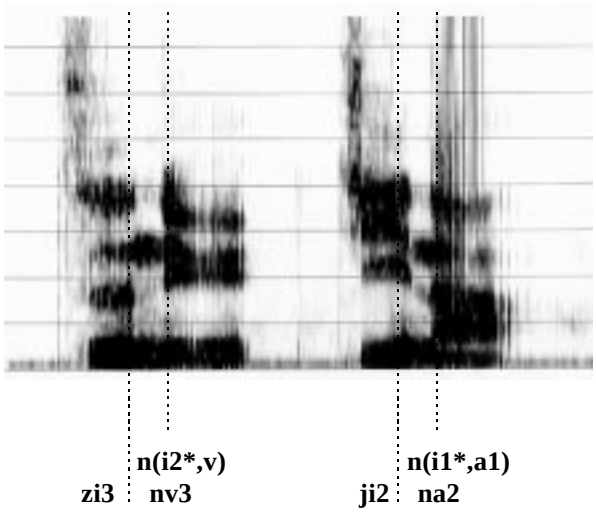


Fig. 1 The sonagraph of "ji2ti3", "bai2ta3", "zi3nv3" and "ji2na2"

32.3 The relationship between several speech units

There are about 400 syllables without tones, which means that there are 16,000 pairs of syllables. This number is too enormous for speech recognition and speech synthesis. To describe transitions between syllables, speech units smaller than the syllable should be chosen. We use diphone, triphone and semi-syllable as speech units. There are more than 400 inter-syllabic diphones, more than 8000 inter-syllabic triphones, and about 1000 final-initial (one kind of semi-syllable). The following example in Table. 1 shows the relationship between them.

Table. 1 The relationship among diphone, triphone and final-initial

word	inter-syllabic diphone	inter-syllabic triphone (left, right)	final-initial
ji2 ti3	i1-t	i1(j,*t), t(i1*,i1)	i1-t
bai2 ta3	i1-t	i(a2,*t), t(i1*,a1)	ai-t/i1-a

where "*" means syllable boundary.

3. THE PHONETIC PHENOMENA FOR SPEECH DATABASE DESIGN

3.1 phonetic balances in segmentation

3.1.1. Syllables without tone

The whole syllable covers all intra-syllabic phenomena such as phone, vowel, consonant, final, initial and all intra-syllabic diphones and triphones. In Mandran syllables "eng, o, ei, yo, lo, hm, tei, nou, kei, rua" are used solely in spoken language. By removing them, there are 401 syllables without tones.

3.1.2. Inter-syllabic diphones

There are 415 inter-syllabic diphones in total.

3.1.3. Inter-syllabic triphones

With 37 basic phones, there are more than 8000 inter-syllabic triphones in Mandarin. Of those, a lot of the triphones are rarely use. The number of triphones is too big for speech recognition and speech synthesis. According to our studies on transitions in pairs of syllables, we can merge them into a small set by articulatory place and articulatory manner with the following rules.

(1) The triphones related to mono-phone syllables
The mono-syllables are "a, i, u, v, e, er,...". The relating triphones are in the form of:

a(pl*,pr*) -- mono-phone syllable /a/
i(pl*,pr*) -- mono-phone syllable /i/
.....

where pl means left phone, pr means right phone, "*" means syllable boundary, so pl* is the last phone of left syllable, pr is the first phone of right syllable. There are more than 3000 such triphones. All mono-phone syllables have relatively stationery duration and therefore ?(?*,?*) can be taken as two diphones("?" means phone, left phone or right phone), for example:

a(i*,n*) ⟶ i-a, a-n

(2) The triphone focus is in the final of the left syllable (whose right is the syllable boundary)
The form is ?(?,?*). The initial of the right syllable can be classified into several types in terms of articulatory places. The final of the left syllable translates to the same point when the initial of the right syllable prolongs to the same type. Nasal final is an exception.

(3) The triphone focus is in the initial of the right syllable (whose left is the syllable boundary)
The form is ?(?*,?). In the zero-initial case those triphones cannot be merged. All stops and stop fricative have a silence duration in front of them as in isolated words.

(4) Excluded pairs
Syllables which are solely used in spoken language cannot be taken into account. By the above rules there are 3035 simplified triphones. They basically can describe the variations in continuous speech.

3.1.4. Inter-syllabic final-initial structures

Using the same rules, the pairs of 38 final and 32 initials can be merged into 781 final-initial structure. All of the above sets of phonetic units can be used to describe coarticulation in Mandarin systematically.

3.2 The phonetic phenomena in prosody

3.2.1. Sentence patterns

To include the prosody phenomena, we give 17 sentence patterns. (provided by Mr. Li, Zh. Q.)

3.2.2. Tone structures in 2,3,4 syllables structure

There are 4 tones in Mandarin syllables. The number of tone structure of phrases are as follows:
N2(2-syllables) = 16; N3(3-syllables) = 48;
N4(4-syllables) = 2048

4. THE METHOD FOR SELECTING CONTINUOUS SENTENCE AND PHRASES

We use the "People's Daily" and ten other newspapers as the original corpus. Automatically collecting phonetically balanced sentence set from such a large text corpus using algorithm is desired[7,8,9]. Fig. 2 and Fig. 3 are the flow charts of collecting sentences and phrases.

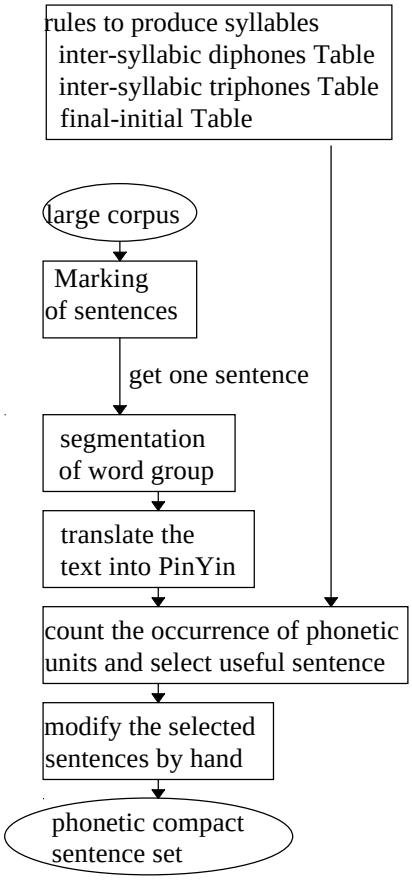


Fig. 2 The flow chart of collecting sentences from the large corpus

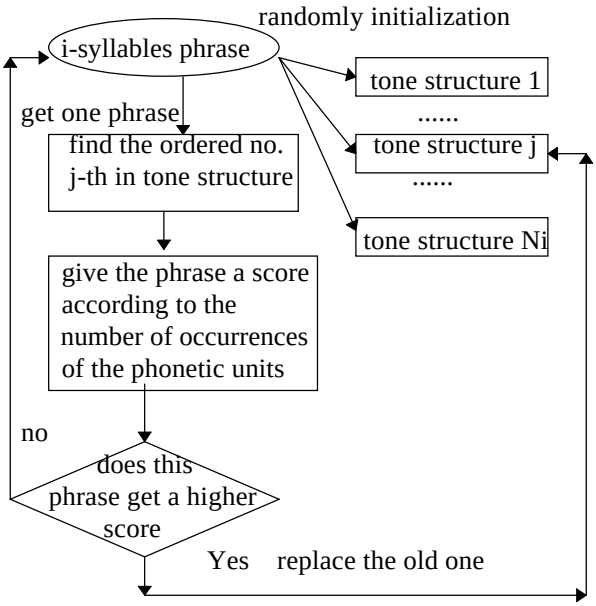


Fig. 3 The flow chart of collecting phrases from large corpus

After getting all of the needed sentences, we check them manually to remove errors from automatic segmentation on sentences and phrases. We can select sentences and phrases for any tasks of speech synthesis and speech recognition.

5. RESULTS

2185 sentences and 380 phrases were selected. The results of the sentence set is shown in table 2. In our results, the syllable discarded is "zhei", which is another pronunciation for "zhe". Twelve unseen triphones are rarely appearances in natural speech. All of the diphone and final-initial structures are present.

Table. 2 The results of selected sentences and phrases

	the number of phonetic units	appearances in the selected set	the probability of the units appearing
unit1	401	400	99.8%
unit2	415	415	100%
unit3	3035	3023	99.6%
unit4	781	781	100%
SP	17	17	100%

unit1: syllable
unit2: inter-syllabic diphone
unit3: inter-syllabic triphone
unit4: final-initial structure
SP: sentence pattern

6.CONCLUSION

6.1. The content of big corpus

By using the "People's Daily", the phonetic information is not sufficient because the style is too formal. This corpus consists of about 75% inter-syllabic triphones and final-initial structrues (syllables and inter-syllabic diphones are fully attained). If we extend the corpus to ten more newspapers, the information increased (including about 86% of the inter-syllabic triphones and final-inial structures). At this time additional corpus cannot contribute further to the phonetic phenomena. To attain fully inter-syllabic triphones and final-initial structures 380 phrases are added by hand.

6.2. Phonetic phenomena

The understanding of phonetic phenomena is based on sonagraph. We think that speech recognition needs more details than speech synthesis. The procedue for building speech database is a research process, it can help us to understand more invariants in continuous speech and find the rules in it.

7. REFERENCES

[1] Zue Victor, Seneff Stephanie, and Glass James (1990), "Speech database development at MIT: Timit and Beyond", Speech Communication Vol.9,No. 4, pp.351-356.

[2] Chen, X. X., "Study on articulation of three articulatory places C2 in 2-syllables structure CVCV in Mandarin", Phonetic Research, 1994-1995, PP. 54-63.

[3] Yan, J. Zh., "A study on formants transitions between nasal-final and zero-initial", Phonetic Research, 1994-1995, PP. 41-53.

[4] Sun, G.H., "Formants transitions in 2-syllables V1-Z in Mandarin", The Proceeding of Third Phonetic Conference, 1996, pp.108-110.

[5] Cao, J. F., "2-syllables table in pairs", Application of Language and Characters, 1/1997, pp.60-68.

[6] Li, A. J., "Pauses among news read speech in Mandarin"£¬The Proceeding of 1997 youth in Acoustic Society, pp.262-266.

[7] Kurematsu Akira, Takeda Kazuya, Sagisaka Yoshinori, Katagiri Shigeru, Kuwabara Hisao * and Shikano Kiyohiro (1990), "ATR JAPANESE speech database as a tool of speech recognition and synthesis", Speech Communication Vol.9,No. 4, pp.357-363.

[8] H.-M. Wang, Y.-C. Chang and L.-S. Lee, "An algorithm for automatically selecting phonetically balanced sentences from a large corpus for training and testing a speech recognition system", Proc. Int. Conf. Computer Proc. Oriental Lang. (Korea),1994, pp.507-510.

[9] Sun, J. S., Wang, Z. Y., Wang, X. & Li, B.(1995) "Constructing a word table for training of continuous speech," In Proceedings of the 2th National Academic Conference on the Latest Development of Computer Intelligent Interface and Intelligent Application (Tsing Hua University Press), pp.116-121.