Effect of Prosodic Structure on Segmental Variants

*Yiqing Zu, Hong Zheng**

Motorola China Research Center, China
Yiqing.Zu@motorola.com

*Shanghai Jiaotong University, Department of Electronic Engineering, China
zx_h0291@263.net

Abstract

There is a large amount of segmental variants in a natural speech corpus. It is very important to label those variants correctly for a corpus based TTS system. We successfully applied automatic triphone segmentation to a large speech corpus with syllable segmentation and prosodic annotation. In this paper, we also report (1) recognition error analysis based on prosodic structure, and (2) the relationship between coarticulation phenomena and prosodic position.

1. Introduction

Speech Corpus is the basis for a text-to-speech (TTS) system, especially for a corpus based speech synthesis. In a natural speech corpus, there are abundant segmental variants. To use the corpus efficiently, we need to label phonetic details for those variants as much as possible. Annotation for a TTS speech corpus includes speech unit segmentation and prosodic annotation. In most TTS systems, segmental annotation is based on segmentation of syllables or units smaller than syllable, such as initial and final. Prosodic annotation includes break index and stress label, which are always labeled by perceptual tests. In common sense, syllable segmentation and suitable prosodic annotation will provide important information for concatenating a Mandarin TTS system. With the prosodic information embedded in each speech unit, natural sounding concatenated speech can be obtained by speech unit selection just on phonetic symbolic level [1,2].

Generally speaking, it is reasonable to take syllable as the basic speech unit for a Chinese TTS system, for there exists relatively obvious boundary in both text and speech waveform. In most Chinese concatenative TTS, syllable is selected as the basic speech unit.

Due to the influence of phonetic context, the syllables with the same citation form may have different acoustic variants. Within a prosodic phrase, adjacent syllables connect each other closely and even overlap. In this case, co-articulation variance happens. There are context limitations on using syllables for a concatenated TTS system. Prosody is very important for improving the naturalness of the synthesized speech. Prosody annotation involves break index (prosodic structure) tier, stress tier and intonation tier as described in ToBI system [3,4] (prosodic labeling system). Among 3 tiers, break index is the easiest to access.

In IBM state-of-the-art, trainable, unit-selection based concatenative speech synthesis system [5,6], phonetic transcription and state alignment of a speech database are all provided by Hidden Markov model (HMMs). To analysis segmental variants in a large speech corpus, this study try to use HMMs and automatically locates the segmental variants based on prosodic structure and tries to reveal the relationship between segmental variant and prosodic position. In this work, nasal coda “n” is found to be a very flexible segment in natural speech. We also try to provide the phonetic context for variants of “n”.

2. Speech Corpus

2.1. Speech corpus construction

We have built a large Mandarin speech corpus, which consists of two trainable speakers’ read speech. This corpus covers all tonal syllables in different sentence position. The reading text was selected from People’s Daily.

The speech signal is recorded by DAT in a studio. The speech waveform was stored in a format of 16 bit, 11.025 KHz sampling rate.

2.2. Speech corpus annotation

This corpus includes manually labeled, syllable segmental and prosodic information. Each syllable in this corpus has the following attributes:

- Phonetic transcription described by PinYin, which reflects actually pronunciation in the corpus, such as tone sandih and tone neutralization
- Co-articulation index
- Tonal co-articulation index
- Break index, including prosodic word break, minor break and major break
- Stress effects

3. Automatic Triphone Labeling

3.1. Speech data and Acoustic feature

The goals of automatic triphone labeling on a speech corpus with manually labeled syllable boundary is finding the errors caused by manual work. HTK tools are used for this task [7]. The great amount of syllables extracted from the speech corpus is used as both training data and testing data.

The acoustic feature set includes 13-order MFCCs (Mel Frequency Cepstral Coefficients), 13-order delta coefficients and 13 acceleration coefficients. Namely, there are totally 39 dimensions. Frame period is 10 ms, and the size of Hamming window is 20 ms.

3.2. Training

A syllable is composed of a series of triphones. There are 1680 phonetic classified triphones in our system, which is obtained by considering the effects of the preceding and following phonetic contexts. A 3-state left-to-right context-dependent HMMs, with continuous observation densities, is used to model the triphones (see Figure 1).

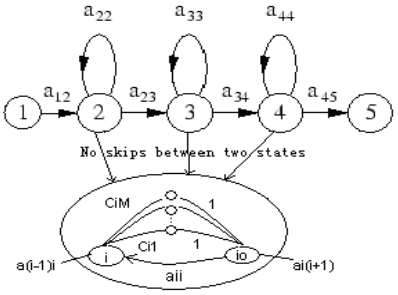

Figure 1 left-to-right HMM

The phonetic classification is based on the left context (see Table 1) and right context (Table 2) of a syllable.

Table 1 Syllable Left Context

Symbol	Articulation place	Phonemes
L-1	a-set	a,e
L-2	i-set	i,v
L-3	u-set	u,o
L-4	Nasal	n,ng

Table 2 Syllable Right Context

Symbol	Articulation place	Phonemes
R-1	Stop, Fricative stop	b,c,ch,d,g,j,k,p,q,t,z,h
R-2	Lip	m,f
R-3	Dental	n,l
R-4	Laynx	H
R-5	a	s,x,sh,r
R-6	a-set	a,e
R-7	i-set	i,v
R-8	u-set	u,o

To describe coarticulation status and prosodic position, each syllable is defined in the form of

$$L_{kj}\ P_i\ R_{lm}$$

Where *L* is left context for a syllable; k stands for left coarticulation context, which is tail phoneme of previous syllable; j is the left prosodic boundary index;

R is right context for a syllable; l stands for right co-articulation index, which is initial phonemes; m is right prosodic break index.

P_i is composed of a series of triphones. In Mandarin, the number of triphones in a syllable varies from 1 to 4. For example, the syllable "chuang" includes four triphones. In phrase "ba chuang kai", syllable "chuang"is composed of four triphones:

a*-ch+u
ch-u+a
u-a+ng
a-ng+k*

Where "*" stands for the phonemes in neighbor syllable, "-" is left context and "+" is right context.

There are five left contexts (see table1) and nine right contexts (see table 2) for a syllable. Silence is also a state in both left and right context.

In training procedure the Baum-Welch re-estimation formulae are used to determine the parameters of HMMs.

Due to the low occurrence of some triphones, a straightforward training method may not be satisfactory. One way to eliminate the influence of insufficient training data on some modeling units is parameter tying. When two or more parameter sets are tied, the same set of parameter values are shared by all the owners of the tied set. Hence, it can reduce the total number of parameters. For triphone, the sets are fixed in training and testing procedures, so-called data-driven Clustering (Furthest neighbor hierarchical cluster algorithm) is used in tying states. In this approach, it is necessary to adjust the threshold of the largest cluster' size. All tying states are sharing means and variance. No transition probability is in sharing mode. M-mixture Gaussian density is also used in training procedure.

3.3 Decoding

In decoding procedure, Viterbi algorithm is used to recognize the testing data. The decoding grammar is constructed according to the corpus transcription. Training sets is regarded as testing sets in recognition process.

4. The Results of Automatic Triphone segmentation

4.1. Primary results

The closed test of recognition in the large amount of syllables extracted from continuous utterances shows that with the 1680 triphone models, the recognition rate is 96.67%, while the syllable rate is 91.03%. The error triphone is that the recognized triphone is different from the triphone defined in transcription.

There is one more prosodic structure in an utterance, which may be associated to syntactic and semantic constraints. The hierarchical prosodic structure [8] in continuous speech is assumed as follows (from large to small): intonation phrase, phonological phrase, prosodic word and foot. The boundaries of the prosodic structure are breaks whose realization can be a pause, pre-lengthening/final lengthening, or pitch movement / F0 reset [9,10].

To find the sources of errors, we have analyzed the result based on prosodic structure. In the speech corpus, we have manually labeled prosodic breaks. There are three break levels:

- Prosodic word boundary (PW)
- Prosodic phrase boundary (PP)
- Intonation phrase boundary (IP)

There is a silence break in PP and IP and always no silence break in PW. So we can classify a syllable into three prosodic positions:

P0: syllable behind a PP boundary
P1: syllable within PP
P2: syllable pre-PP boundary

Then we calculated the recognition rates in three prosodic

classes respectively (see Figure 2). The triphone error rate in the pre-boundary syllables, P2, is lower than in other prosodic position.

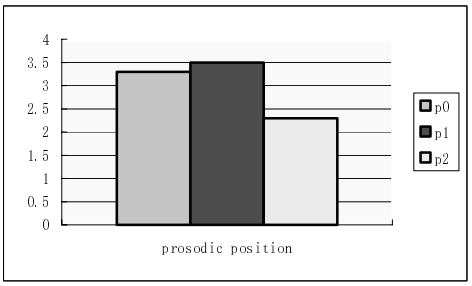

Figure 2: the error rate of syllable in different prosodic position

With the three prosodic position classes, we can review the syllable duration because segmental lengthening is one of pre-boundary cue [11,12,13]. To normalize syllable duration, the z-score duration is defined as following:

$$z_{ij} = (d_i - \mu_j)/\sigma_j \quad (1)$$

Where d_i is the i-th observed syllable duration in that utterance; μ_j and σ_j are the j-th syllable mean and standard variance in the syllable inventory, (there are totally about 1,300 tonal syllables) respectively. z can be positive or negative. The probability of the normalized duration falling between –3 and +3 is 0.998. The duration of shortened syllable is shorter than the mean and will have a negative value of z; and the duration of lengthened syllable is longer than the mean and will have positive value of z. The duration distribution of prosodic position is showed in figure 3.

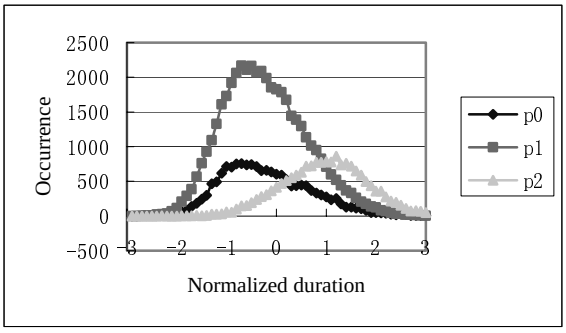

Figure 3: the syllable duration distribution in different prosodic position

Comparing the error rate with the duration distribution, we can conclude that the longer the syllable duration is, the lower error recognition rate becomes. Articulation gesture in a shortened syllable is always under target and even lost. In P1, a syllable connects its left and right neighbor closely and the effect of co-articulation is strong. The reason of higher error rate in P0 than in P2 is that a speaker will try best to keep the initial articulation clearly, so that initial segment is steadier than final, or the final in continuous speech is so flexible that more variants will happen.

4.2. Further evidence

We focused on the syllable with triphone error detected by the recognition procedure and evaluated the errors one by one by perception. We found that some of the "errors" were not errors indeed, because they reflect the actual pronunciation. Those "errors" came from the un-consistency between defined transcription and actual pronunciation. Among the triphones whose actual pronunciation is different from transcription, there are also many recognition errors because the co-articulation effects are so complex that they bring a lot of difficulty to the recognition procedure. If error is redefined as the exact recognition error, the error rate reduces to about 2%. Although the error in pre-boundary position is still lower than other prosodic position, the difference is not so significant.

By analyzing the errors, we can summarize the sources of error:

- Labeling error in source data;
- Lack of training samples;
- Some initials confusion (e.g. "b"/"d");
- Abnormal pronunciation due to co-articulation.

5. Issues in Segment Substitution and Deletion

As a result of recognition error analysis, the main error source is caused by co-articulation, which exhibits as unit substitution and deletion. It is interesting that this kind of phenomena happen frequently in cases of nasal coda, which is counted as more than 60% of the 2% errors.

The perceptive analysis for recognition errors shows that (1) there are less segmental deletions occurring in pre-boundary position than other prosodic positions; (2) among all segments involving errors, nasal coda "n" is the most frequent one. In other word, most of segmental variants with co-articulation effect are related to nasal coda "n". There are three co-articulation statuses, which involve more than 50% of segmental variants (see Table 3). This kind of errors may be due to high articulation rate and is likely to result in segmental substitution or deletion. The main substitution in this corpus is nasal coda "n" becoming "ng" when followed by a back segment, such as "g, k, h, u" [14]. " n" even disappears in some case. Table 3 lists the variants of nasal coda "n".

The syllables with co-articulation effects are not ignorable when a speech corpus is used for concatenative TTS system. In this situation, we need to label them, so that they can only be used together with their left and/or right context, but independently.

Table 3 Co-articulation issues

Co-articulation Issues	**Phonetic Context**	**Prosodic Position**
n - > ng	Followed by g, k, h, u	First or middle position in a prosodic phrase
ian -> ie	With initial m, b, j	First or middle position in a prosodic phrase
men -> me	Tone5, "n" coda in left syllable	First or middle position in a prosodic phrase

6. Discussion and Conclusion

6.1. The nasal coda "n" is a more flexible segment than "ng"

The nasal coda "n" in standard Chinese is more flexible than "ng". It will be lost when fast articulated, especially in nasal context. So far we have not found the similar phenomena in nasal coda "ng" in our speech corpus.

The nasal coda "n" substitution is likely to occur in the following conditions:

- The prosodic position is not pre-boundary;
- The following segment is back one.

The nasal coda "n" loosing is likely to occur in the following conditions:

- The syllable has a neutral tone,
- The syllable initial is nasal and the former syllable is also a nasal coda;
- The prosodic position of the lost nasal coda is a middle syllable in a prosodic phrase or the first syllable in an utterance, because in these two cases the durations are shorter than what are in other positions.

This result is derived from one speaker's data. Although we have found similar phenomena in another speaker, we need more speakers' data to verify this conclusion.

6.2. Perception aspect in co-articulation

It is an interesting phenomenon in nasal coda deletion that listeners can exactly percept "n" code in some lost nasal coda syllable in a sentence level. For example, in some case, the Chinese syllable "me" may be listened as "men". But when this syllable is extracted from the sentence, it is listened as "me". In this case, the former syllables always have a nasal coda. In other word, the former nasal coda syllable can influent the next syllable in a perceptual level.

6.3. Break levels for prosodic annotation

As we used before, there are four levels break indexes in prosodic annotation: 0 (no pause); 1 (minor break); 2 (major break); 3 (intonation break).

To deal with the entirely segmental variables, we introduce another break level, "-1", for segmental overlapping. The break index annotation is listed in Table 4.

Table 4 Break Index Suggestion

Break Index	Prosody
-1	Segmental overlap
0	No pause
1	Minor break
2	Major break
3	Intonation break

The "-1" index signals a warning for using it.

6.4. Error control of speech corpus

In building a TTS corpus, a lot of manual work is needed in segmental and prosodic annotation. Manual work will inevitable introduce mistakes. The influence of a small error can be enlarged in a TTS system. For example, if a frequently used syllable carries error, it will repeat in many places of the synthesis speech. Therefore we need to develop a scientific method to locate errors in a large speech corpus.

6.5. Improvement of TTS system

As a result of automatic triphone segmentation on TTS speech corpus, we wrote the phonetic details on annotation files. For instance, the change of nasal coda "n" to "ng" is annotated on the hosting syllable. This annotation will tell the TTS system that the "n" coda identity is kept only when in word context and when it is isolated from the context its phonetic identity changes into "ng". With the annotation of phonetic details on TTS speech corpus the quality of output of TTS system is improved.

7. References

[1] Black, Alan W.; Campbell, Nick, 1995. Optimizing selection of units from speech databases for concatenative synthesis. *Proceedings of Eurospeech 95*, vol. 1, 581-584.

[2] Campbell, Nick, 2000. Processing a Speech Corpus for CHATR Synthesis. ICSLP2000.

[3] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.W.; Price, P.J.; Pierrehumbert, J.; Hirschberg, J.; 1992. TOBI: A standards for labeling English Prosody. *Proc. ICSLP'92*, 867-870.

[4] Beckman, Mary E.; Gayle, Ayers Elam; 1997. Guidelines for ToBI Labeling.

[5] Donovan, Robert Edward et. al., 2001. Current Status of the IBM Trainable Speech Synthesis System. *4th ISCA workshop on Prosody*.

[6] Donovan, Robert Edward, 1996. Trainable Speech Synthesis. *Ph.D. dissertation*.

[7] Yong, Steve et. al., 1999. The HTK Book version2.2.

[8] Selkirk, E., 1990. Phonology and syntax: the relation between sound and structure. Cambridge. MA: MIT Press.

[9] Campbell, Nick, 1993. Automatic Detection of Prosodic Boundaries in Speech. *Speech Communication* 13, 343-354.

[10] Santen, J P. H. Van, 1992. Contextual effects on vowel duration. *Speech communication* 11, 513-546.

[11] Crystal, Thomas H.; House, Arthur S.; 1982. Segmental durations in connected speech signals: Preliminary results. *J. Acoust. Soc. Am.*, 72(3), 705-716.

[12] Crystal, Thomas H.; House, Arthur S.; 1988. Segmental durations in connected speech signals: Current results. *J. Acoust. Soc. Am.*, 83(4), 1553-1573.

[13] Zu, Yiqing, 1999. Segmental Duration and Lengthened Syllables. *Proceedings of ICPhS'99*.

[14] Huang, Jingjing, 2001. Acoustic Analysis on variety of /n/ to /ng/. *The Proceeding of 5th National Conference on Modern Phonetics*, 130-132.