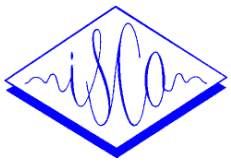

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Mandarin Speech Synthesis by the Unit of Coarticulatory Demi-syllable

Chi-Shi Liu[1,2] Wern-Jun Wang[1]
Shiow-Min Yu[1] Hsiao-Chuan Wang[2]

[1]Telecommunication Laboratories, Ministry of Communication, Chung-Li, Taiwan, R.O.C.
[2]Department of Electrical Engineering, National Tsing Hua University, Taiwan, R.O.C.

Abstract

In this paper, we will discuss the selection of synthesis unit for Mandarin speech synthesis. To choose the synthesis unit which could be easily segmented and concatenated, and produce natural and intelligent speech, we proposed the synthesis unit of Coarticulatory Demi-syllable segmented from di-syllable words. We found that spectral continuity of concatenating demisyllable still kept good. By subjectively listening, testers felt that the Coarticulatory Demi-syllable produced more natural and intelligent synthetic speech than monosyllable, but less than synthesis-by-analysis speech.

1 Introduction

Text to speech synthesis system has been developed many years and it had been built up for some languages[1,2,3,4]. For a good text to speech system it should include language processing, development of sandhi rule, choice of synthesis unit and selection of synthesizer. Among these processing, choice of synthesis unit is one of important factors in text to speech system. It affected the performance of segmental information. Usually the synthesis units include *phoneme, diphone, demisyllable, monosyllable* and others. However, Different language would select different synthesis unit according to their language's characteristic. But, no matter what kind of synthesis unit, the most appropriate unit should be easily segmented, simple rule in concatenating these units and moderate number of units to produce natural synthetic speech. Under these criterions, any segmental methods and phonetic knowledge are jointed to select an appropriate unit.

Before we proposed *monosyllable* as our synthesis unit [5], this unit is the simplest unit because it was easily gotten from any speaker by letting him read Mandarin 408 monosyllables. Although this unit is easily segmented and a not bad synthetic sound could be generated, it would generate unpredicted plosive sound in the unit's boundary due to spectral discontinuity between units. Another unit usually used in unrestricted text-to-speech synthesis is *phoneme*. The main advantage of this unit is its smaller size and less complexity of generating allophone than the unit, *monosyllable*. However, we would spend more time and need more phonological knowledge to generate good *phonemes* and *allophones*. All of these are not easy and usually are reseacher's dependence. To overcome above problem, we proposed the *coarticulatory demi-syllable*(CDS) as our new synthesis unit. This unit was segmented from disyllable words which cover all possible coarticulation between monosyllables. After the segmentation from stable spectral region, four demi-syllables were generated. Only the second demi-syllable, called post-demisyllable, or the third demi-syllable, called pre-demisyllable, was left to use. About three thousand demi-syllables were used as our synthesis unit after applying reduction method in originally segmented demi-syllable. The main advantage of CDS was segmentation and concatenation of these CDS would be easier and adjustment of phonemes and allophones were also avoided. In next section we introduced the method of generating CDS and explained the meaning of CDS. In section 3 we used dynamic time warping and weighted distortion measure to automatically joint some CDSs which had similar spectrum and chosed a representable CDS among these CDSs as a final synthesis unit. We would describe the method of concatenating these synthesis units in section 4. In section 5 the synthetic speech from direct analysis, the unit of monosyllable and the unit of CDS were generated and subjectively evaluated. Conclusion and future work were covered in section 4.

2 Generation of Coarticulatory Demisyllable

A natural and intelligent synthetic speech usually need a good synthesis unit. This unit must (1) be easily and clearly segmented from artificial database, (2) the rules of concatenating these units must be simple, and (3) the number of unit must be moderate. In our Chinese language, it includes 21 consonants(initial sounds) and 39 diphthongs (final sounds). These 39 diphthongs are generated from combination of 3 middle phonemes (/i/, /u/, /iu/), 9 principal phonemes (/i/, /u/, /iu/, /a/, /o/, /e/, /eh/, /er/, /z/) and 4 ending phonemes (/i/, /iu/, /n/, /ang/). Only about 1,300 basic monosyllables in Mandarin and any Chinese characters were pronounced from these basic monosyllables. In Chinese text to speech system some researchers have tried to find a good synthesis unit[6,7]. They proposed *monosyllable* and *phoneme* as synthesis units. Both of them have disadvantage as we said in Section I, either complex concatenation rule or too large size of synthesis unit. To overcome these disadvantage, we proposed coarticulatory demi-syllable as our new synthesis unit. All of demisyllables were segmented from di-syllable words and these disyllable words have been incorporated into coarticulation effect. We summarized the procedure of generating these units and stated as follows :

(1) Record disyllable words. Its structure should be in the set

$$\{M = M_1 + M_2 \mid M_1, M_2 \in \{C_1V_1, C_1V_1V_2, V_1, V_1V_2\}\}$$

M repersents monosyllable word, C is consonant and V is vowel. All of C and V are the smallest synthesis unit, phoneme. To consider coarticulation, these recorded words should include all possible coarticulation between syllables. About four thousand words were recorded.

(2) After recording these disyllable words, we used the feature of delta cepstrum[8] to segment the pre-demisyllables and post-demisyllables. The boundary of pre-demisyllable is the place of maximal spectral change in the left side and the place of minimal spectral change, the most stable region, in the right side. But, for the post-demisyllable the condition of boundary is just reversed to pre-demisyllable. Fig. 1 was shown the boundary of pre-demisyllable /ga-/ and post-demisyllable /-au/ from a disyllable word /ia-ga/ and disyllable word /pau-ta/, respectively.

(3) After step 2 we used these segmented coarticulatory demisyllables as synthesis unit. The number of pre-demisyllables was about $4,488(11 * 408)$ and the number of post-demisyllables was about $1,248(39 * 32)$. This number was too large. To reduce this size, we used the level of spectral similarity to merge these demisyllables. This method would be further stated in next section.

To summarize the characteristics of our mentioned units, in Table I we listed the advantage and disadvantage of the units, coarticulatory demisyllable, coarticulatory phoneme, coarticulatory monosyllable, monosyllable, phoneme used for Mandarin speech synthesis. As seen from Table I, the size of database for CDS is lager than monosyllable and phoneme, but less than coarticulatory monosyllable and coarticulatory phoneme. The complexity of concatenation for CDS is lowest among these synthesis units. The phonological knowledge and influence of reseacher's dependence would be reduced dramatically. Based on above analysis and consideration, we naturally adopt the CDS as our new synthesis unit.

3 Reduction of synthesis unit

In the section 2, we mentioned our new synthesis unit, coarticulatory demisyllable. The total number of synthesis units used in section 2 is about five and half thousands. This number was too large. To reduce this number, we used the level of weighted spectral similarity to merge the similar spectral CDSs and choose the CDS having the minimal distance between CDSs in that class as a final CDS. To do this, we first classified the same pre-demisyllable having the different ending phoneme in their previous sound into one class and also the same post-demisyllable having the different initial phoneme in their next sound into another class. For convenience, we called these classes as *demisyllable classes*(S^{ds}). After classification, we got 11 classes, each class having 222 elements, for pre-demisyllable and 32 classes , each class having 39 elements, for post-demisyllable. To emphasize the spectral continuity, we used weighted spectral similar distance as our criterion to choose representable elements for each class. The weighted spectral similar distance we used was :

$$\begin{aligned} d(S^l, S^m) &= \sum_{j=1}^{n}\{wt(j) * [\sum_{i=1}^{p} we_1(i)(S_c^l(i,j) - S_c^m(i,j))^2 + \\ &\quad \sum_{i=1}^{p} we_2(i)(S_{dc}^l(i,j) - S_{dc}^m(i,j))^2]\} \end{aligned} \quad (1)$$

$S^l(i,j)$ and $S^m(i,j)$ are the feature vector of order i in frame j for demisyllable l and m in the same demisyllable class, respectively. "dc" represents delta-cepstrum, and "we" is the weighting value in the intra frame, "wt" is the weighting value in the inter frame, p is the dimension of feature vector and n is the frame length of each demisyllable. To emphasize spectral continuity in boundary, we put more weight in first few frames for pre-demisyllable and more weight in last few frames for post-demisyllable.

By this weighted spectral similar distance, we merged the similarly spectral classes into the same class and chose anyone class among these classes as a representable one.

Since the duration of elements in each class was different, we must adjust this time variation. The approach used for time alignment is Dynamic Time Warping (DTW) algorithm [8] which was shortly written as :

For a point (n,m) in the grid the minimum accumulated distance $D_a(n,m)$ from the start point (1,1) can be recursively defined as :

$$\begin{aligned} D_a(n,m) &= d(T(n), R(m)) + \\ &\quad min[D_a(n-1,m)g(n-1,m), \\ &\quad D_a(n-1,m-1), D_a(n-1,m-2] \end{aligned} \quad (2)$$

where

$$g(n-1,m) = \begin{cases} 1 & \text{if } w(n-1) \neq w(n-2) \\ \infty & \text{if } w(n-1) = w(n-2) \end{cases}$$

$w(n)$ was the warping path and distance $d(T,R)$ was defined as equation (1).
By this reduction algorithm, we reduced the number of synthesis unit from about five and half thousands to three and half thousands (2,442 for pre-demisyllable, 819 for post-demisyllable). It saved about 2/5 memory space. To check the correctness for such reduction method, we showed the spectrum of the subset after this reduction to see their similar level. Fig.2 is the spectrum of /-au/ from disyllable /pau-geng/and /pau-ka/ which were from the same reduced demisyllable class. From this figure we saw the similarity of their spectrum was very high and prove this reduction algorithm useful.

4 Concatenation of Synthesis Unit

After generating the synthesis unit and doing the reduction of synthesis units, we should concatenate these synthesis units to a unrestricted text to speech. On generating synthesis unit, in fact we assumed that forceful concatenation of pre-demisyllable and post-demisyllable, in which their stable spectral region had the same phoneme but from different monosyllable, should be little spectral discontinuity. If this condition was satisfied, we could consider the following concatenation of allophone. Fig.3 was

shown the spectrum / shieh-ia/ by concatenating pre-demisyllable /i-/ from disyllable /chieh-i/ and post-demisyllable /-ieh/ from disyllable word /shieh-ian/. We didn't find the spectral discontinuity in the concatenating region. This phenomenon let us confirm that this new synthesis unit is feasible.

Before concatenating these demisyllables into the whole sentence, we initially classified the set of these reduced synthesis units into two subsets($S_i, i = 1,2$), one for pre-demisyllable and the other for post-demisyllable. These subsets were again classified into 11 sub-subsets ($S_{1j}, j = 1,...,11$) for pre-demisyllable and 21 sub-subsets($S_{2j}, j = 1,..,21$) for post-demisyllable by the reduction rule of section 3. In each sub-subset, we again classified to get the terminal node(S_{ijk}) , 222 for pre-demisyllable path and 39 for post-demisyllable path. In each sub-subset, we had one and only one representable element(S_{ijk}). By this way, we constructed a coarticulatory tree of four layers for synthesis unit shown in Fig.4. On concatenating these units, we initially analyzed the input text string to know pre-phoneme and post-phoneme for each morpheme. Then, by the existing coarticulatory tree we searched the right demisyllable from root of tree and decided pre-demisyllable or post-demisyllable in the second layer. The path to go into layer 3 and layer 4 was according to pre-phoneme and post-phoneme of neighboring morpheme. After this top-down search we found the right pre-demisyllable and post-demisyllable to concatenate this coarticulatory monosyllable. Any other sound of morphemes would be searched in the same way. In Fig.4 we showed how to use this coarticulatory tree to find the right monosyllable /shieh/ in the word /gau-shieh-ia/.

5 Experiment and Result

To evaluate the performance of CDS synthesis unit used in unrestricted text to speech system, we would synthesize five phonetically balanced sentences by the synthesis unit of CDS and monosyllable. Here, we used monosyllable as compared unit just because it is easily segmented and already used by us. We also generated the synthetic speech from direct analysis as toll quality of these synthetic speech. To describe conveniently, we called the synthesis from any kinds of synthesis unit as synthesis-by-rule(SBR) and the synthesis from directly analyzing as synthesis-by-analysis(SBA). Because the LSP(Line Spectrum Pair) synthesizer has better interpolation and quantization than other linear predictive synthesizer. we decided to use this synthesizer to synthesize all of these artificial speech.

All of these di-syllable and monosyllable were filtered by 4kHz low pass filter and sampled at 8 kHz. After sampling, these digitized signal was through pre-emphasis filter ($1\text{-}0.98z^{-1}$) and analyzed by 20 msec Hamming window with 15 msec shifted. The analysis condition for SBA was the same. By these analysis conditions, we got LSP parameters of ten orders. To trully understand and evaluate the performance of these synthesis units used in unrestricted Chinese text to speech, we kept the supersegmental information, "pitch contour", "duration" and "energy contour", which was gotten from directly analyzing test sentences, as supersegmental information of (SBR) and just changed segmental information, spectral variation. Fig.5 was shown the spectrum of the synthetic test speech by concatenating CDS, concatenating monosyllable and directly analyzing test speech for one sentence. We found in the boundary of synthesis unit the spectral continuity for CDS was better than monosyllable. Another experiment was that we asked ten person to listen these synthetic speech. These ten person were not told what sentences were. By a simple AB test, they felt that the quality of SBR by CDS is better than by monosyllable, but slightly worse than the quality of SBA. These two experiments let us believe that this synthesis is moderate for Mandarin speech synthesis, although we needed to increase the number of synthesis units.

6 Conclusion

In this paper we have proposed a new synthesis unit called coarticulatory demisyllable for Mandarin speech synthesis. Its advantage is easy segmentation from disyllable words, unnecessary concatenation rule to produce any word or sentence, and the acceptable size of synthesis unit. To understand and evaluate the performance of this synthesis unit, we also objectively and subjectively evaluated the performance of this synthesis unit used in Mandarin speech synthesis. It showed this unit used in Mandarin speech synthesis was agreeable and the quality was satisfactory.

7 Acknowledgements

The authors would thank Dr. S.C. Lu, Director of Telecommunication Laboratories and Dr. I.C. Jou for their invaluable advice and timely encouragement.

References

[1] D.H.Klatt, "The KLATTalk text-to-speech conversion system, " in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1982, pp.1589-1592.

[2] E. Moulines *et al.*," A real time French text-to-speech system generating high quality synthetic speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1990, pp.309-312.

[3] Hirokazu sato, "Japanese text-to-speech conversion system," review of the Electronical Communication Laboratories, Vol. 32, No.2, 1984

[4] Several tex-to-speech systems about different countries are discussed in the proceeding of ICASSP'82, the IEEE International Conference on ASSP.

[5] C.S. Liu *et al.*, " A Chinese text-to-speech system based on LPC synthesizer," TL Technical Journal in Taiwan, Vol.19, No.3, 1989, pp.269-285.

[6] T.Y. Huang, C.F. Wang and Y.H. Pao, " A Chinese text-to-speech system based on an initial-final model," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing , 1982, pp.1601-1603.

[7] L.S. Lee, C.Y. Tseng and M. Ouh-Young," The synthesis rules in a Chniese text-to-speech system," IEEE Trans. Acoust., Speech, Signal Processing, Vol.37 No.9, Sept., 1989, pp.1309-1319.

[8] F.K.Soong and A.E.Rosenberg,"on the use of instantaneous and transitional spectral information in speaker recognition," IEEE Trans. on Acoust., Speech, Signal Processing, Vol.36, pp.817-879, June 1988.

Table 1 Some Comparisons of different kinds of syhthesis unit

Compared Items \ Synthesis Unit	Coarticulatory Demi-Syllable	Coarticulatory Phoneme	Coarticulatory Monosyllable	Monosyllable	Phoneme
Unit of Segmentation	di-syllable word	tri-phone word	tri-syllable word	monosyllable word	phoneme
Rule for Allophone	no	no	no	complex	complex
Size of Basic Units	11*408+39*32	11*21*9+32 *11*32	408*408*408	408	32
Average Length of Each Unit	105 msec	60 msec	210 msec	240 msec	60 msec

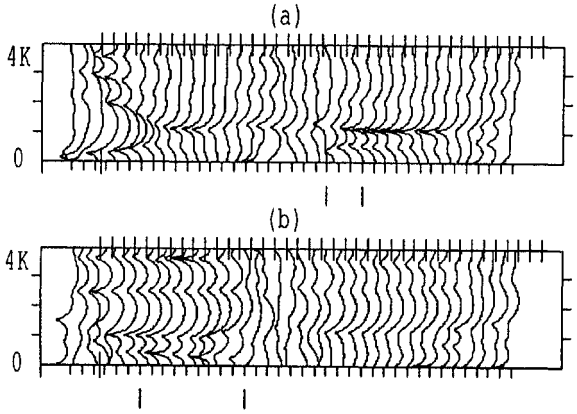

Fig.1 (a) The boundary of pre_demisyllable /ga-/ from disyllable /ia-ga/
(b) The boundary of post-demisyllable /-au/ from disyllable /pau-ta/

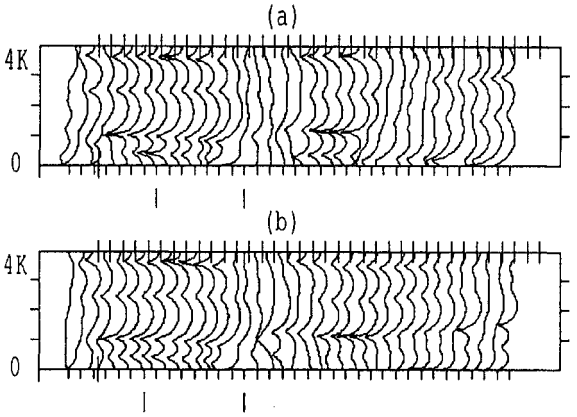

Fig.2 The similar spectrum of post-demisyllable /-au/
(a) from disyllable /pau-geng/
(b) from disyllable /pau-ka/

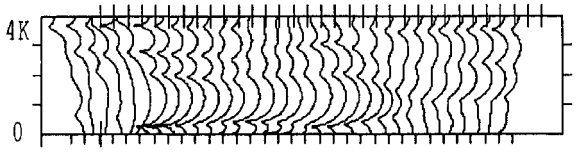

Fig.3 The spectrum of word /shieh-ia/.
The pre-demisyllable /i-/ comes from disyllable /chieh-i/, and post-demisyllable /-ieh/ comes from disyllable /shieh-ian/

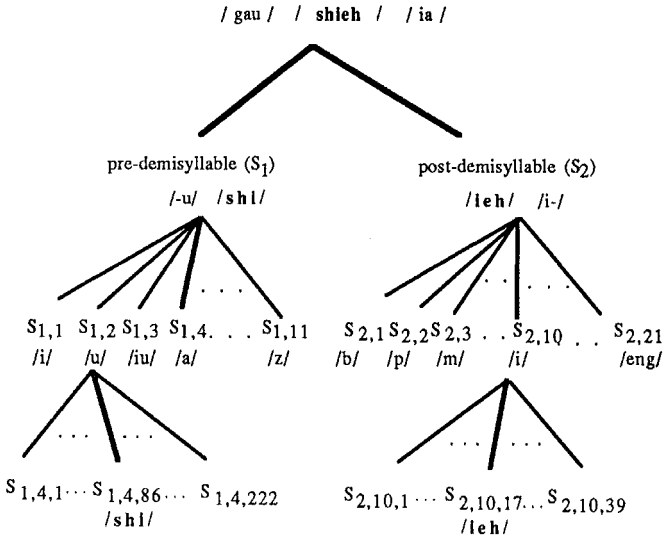

Fig4. The 4-layer coarticulatory tree used for synthesis of a monosyllable /shieh/ in word /gau-shieh-ia/.

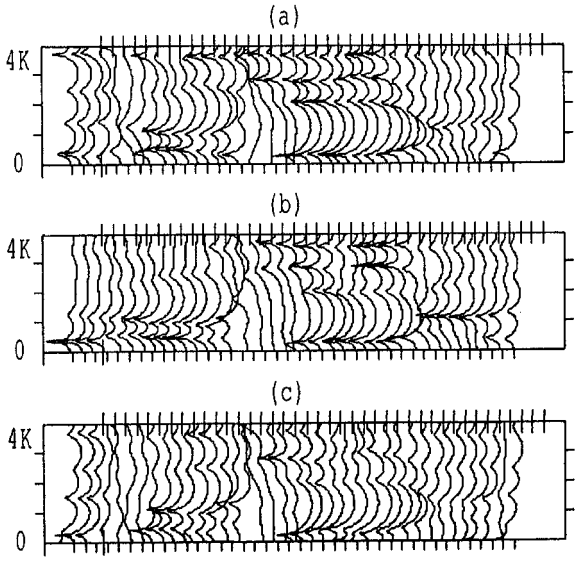

Fig.5 The spectrum of word /gau-shieh-ia/
(a) balanced sentence (b) monosyllable
(c) coarticulatory demisyllable.