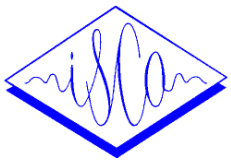

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

DURATION STUDY FOR THE AT&T MANDARIN TEXT-TO-SPEECH SYSTEM

Chilin Shih and Benjamin Ao
AT&T Bell Laboratories
600 Mountain Avenue, Murray Hill, NJ, USA, 07974–0636
{cls,ao-b}@research.att.com

Abstract

We present in this paper the methodology and results of a duration study designed for the Mandarin Chinese Text-to-speech system of AT&T Bell Laboratories. A greedy algorithm is used to select text from on-line corpora to maximize the coverage of factors that are important to duration study. Duration model and interesting results will be discussed.

Introduction

Previous Mandarin duration studies typically took the form of controlled experiments, where a limited number of speech sounds or contextual factors were examined in sentence frames (Feng, 1985; Ren, 1985). While experiments were effective in providing answers to specific questions, the contextual variation was not rich enough to capture durational variation in natural speech. Speech database covering as many speech sounds and contexts as possible will be better suited for the construction of a duration model for a natural sounding text-to-speech system. The drawback, of course, is that the preparation of large database is time-consuming. We tried to have the best of both worlds: creating a versatile database that has good coverage of Mandarin sounds and factors affecting duration, and is manageable in terms of size. The availability of on-line text corpora, word parsing program, greedy algorithm for text selection, and data analysis tools at AT&T Bell Laboratories made the task possible.

Database

We extracted 15620 sentences 25 to 50 characters long from the 9 million character ROCLING Chinese text corpus, parsed the character strings into words using an automatic segmenter (Sproat et al., 1994), and transcribed the sentences into phonetic representation. We basically followed the Romanization system *pinyin* to represent Mandarin sounds, but when a pinyin symbol is ambiguous or consists of two letters, we assigned a unique, one letter symbol. Table 1 gives the correspondence between pinyin and our notation where there is a difference.

Every segment in the sentences was then coded with a set of factor values. Based on previous reports on Mandarin duration (Feng, 1985; Ren, 1985) and literatures on other languages (Allen et al., 1987; Crystal and House, 1988; van Santen, 1992) we chose the following factors as the focus of our investigation. 1. Identity of the segment; 2. Identity of the tone; 3. Identity of the previous segment; 4. Identity of the previous tone; 5. Identity of the next segment; 6. Identity of the next tone; 7. Number of preceding syllables in the word; 8. Number of following syllables in the word; 9. Number of preceding syllables in the phrase; 10. Number of following syllables in the phrase; 11. Number of preceding syllables in the utterance; 12. Number of following syllables in the utterance; 13. Syllable type. One more factor, prominence level, was added later by transcribing the recorded speech.

The factor values of each segment were grouped into factor-triplets, each consisting of the current segment, the current tone and one of the other 11 factors. As a result, each segment is represented by 11 factor-triplets. There were 1,385,451 segments in the input text, with 8,233 unique types of factor-triplet. To ensure that as many types of factor-triplet were covered with the smallest number of sentences, we used a greedy algorithm (Cormen et al., 1990; van Santen, 1993) to search through the coded sentences. During each search the sentence with the most unseen factor-triplet types was selected. 424 sentences were chosen covering 100% of the factor-triplets types. The resulting database contains 38,881 segments, or 19150 syllables.

Figure 1 compares the performance of the greedy algorithm to random text selection. The effectiveness of the greedy algorithm is apparent. While 424 sentences selected by the greedy algorithm cover 100% of the input factor-triplets, 424 randomly selected sentences cover only 74%. If 74% of coverage is acceptable, 42 sentences selected by the greedy algorithm would be sufficient.

After manual correction of transcription and word segmentation errors, the 424 sentences were recorded by one male Beijing Mandarin speaker in a sound-proofed room with a Brüel and Kjær microphone 2231. The transcription was edited again to match the recorded speech. Phrasing and prominence levels were also transcribed

Pinyin	(sh)i	(d)e	j(u)	(s)i	er	ou	ei	ai	a(n)
Our Symbol	J	E	U	Q	R	O	A	I	F
Pinyin	ao	(d)i(e)	(d)u(o)	yu(e)	sh	ch	zh	N	G
Our Symbol	W	y	w	Y	S	C	Z	(i)n	ng

Table 1: Conversion chart of symbols

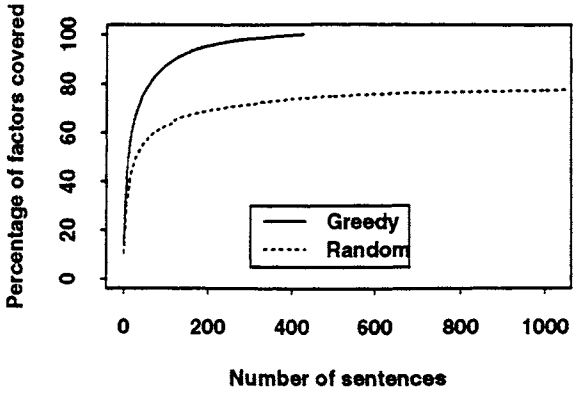

Figure 1: Random and greedy sentence selection

o	J	u	E	U	i	Q	er
111	115	117	118	118	118	122	125
O	e	A	I	F	ao	a	
126	129	131	140	142	144	150	

Table 2: Corrected means of vowels

to match the reading. The recorded speech was then manually segmented, using Waves (Entropic Inc.) on an SGI Indigo workstation, following a set of rigid criteria (French et al., 1993). Finally, the factor values and duration of each segment were coded in a matrix form suitable for statistical analyses.

Modeling

Our data showed very clear patterns of intrinsic duration, which make it possible to estimate the duration of a segment with a predictive model. For example, we observed the same relative scale in vowel duration under various degree of prominence, in different positions of a phrase, and in various subsets of the data. The best estimates of corrected means (in msec) of vowel durations for the entire dataset is given in Table 2. Consistent with previous findings, low vowel *a* and low diphthongs are longer than other vowels. One difference in our study is that the mid vowels are not longer than high vowels. Especially, the mid vowel *o* is consistently the shortest vowel in various subsets of the data.

In the multiplicative model, the most important factors that affect vowel duration are:

1. **Prominence**: level 2 > level 1 > normal
2. **Syllable type**: open syllable without glide > open syllable with glide > closed syllable without glide > closed syllable with glide
3. **Phrasal position**: final > nonfinal

The following factors have some effect on vowel duration:

1. **Identity of tone**: full tone > neutral tone; among full tones, 3 > 2 >4 > 1
2. **Previous phone**: across syllable boundary > within syllable boundary; among across syllable: non-low vowel > nasal coda > low vowel and diphthong; among within syllable: unaspirated plosive and sonorant > fricative and glide > aspirated plosive
3. **Following phone**: diphthong > monophthong > plosive, fricative > sonorant

Previous tone, following tone, and within-word position have very little effect on the duration of vowels.

The fricative scale, *x* (117) > *s* (115) > *S* (111) > *f* (95) > *f* (91), was also observed repeatedly in various subsets of the data. Numbers in parentheses are the corrected means in msec. It is interesting to note that the durations of vowels and fricatives correlate with the intrinsic loudness of the sound within the category.

The important factors affecting fricative duration are:

1. **Prominence level**: with prominence > normal
2. **Following phone**: high vowel > mid vowel > low vowel
3. **Position in utterance, phrase, and word**: initial > non-initial
4. **Tone**: full tone > neutral tone

Factors that have no effect include the previous phone and the following tone.

Mandarin is a tone language. It is natural to ask whether tones have any effect on duration. Previous reports of tonal effect were not consistent, except that the neutral tone (Tone 0), being a reduced syllable, is much shorter than others. In our data, vowel length were affected by tone: *Tone 3* > *Tone 2* > *Tone 4* > *Tone 1*. The differences between each pair are significant, but tones account for only 0.4% of the variation in the data. Tonal effects on syllable duration are in general not significant.

Even though our speaker read the database in a dramatic style, with frequent shift of speaking rate and with

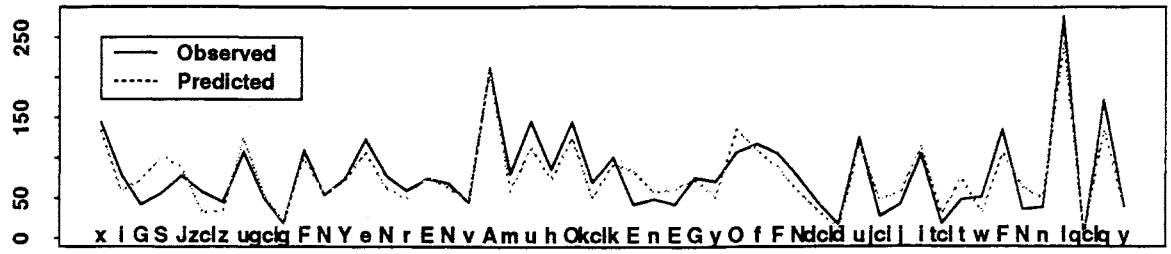

Figure 2: Comparison of observed and predicted duration

liberal usage of exaggeration, the performance of our predictive model is very good. Consistently, a multiplicative model performs better than an additive model. The squared mean difference between the observed values and the predicted values given the multiplicative model is 25 msec. Figure 2 compares the actual segmental duration and the predicted segmental duration of the first 50 segments of the first sentence.

Discussion

One of the most important properties of our database is its richness in the coverage of factor interactions, which allows us to explore the data in many directions, some of them unplanned. We discuss two interesting cases below: (incomplete) compensatory effect, and the lack of utterance-final lengthening.

Compensatory Effect We use two examples to illustrate the compensatory effects: the relation between vowels and syllable types, and the relation between vowels and coda consonants.

Vowel duration is affected by the structure of a syllable. The duration of a vowel in the simplest syllable structure *V* is on average 3.5 times the duration of the same vowel in the most complicated syllable structure *CGVC*. However, the compensatory effect is incomplete. There are still considerable differences in syllable length. The more phonemes there are in a syllable, the longer the syllable duration is. The duration of the longest syllable type, *CGVC*, is 1.5 time the duration of the shortest one, *V*. Vowel and syllable durations by syllable type are plotted in Figure 3. The full length of each bar represents the duration of the syllable of a given syllable type, and the dark shade represents the duration of the vowel. *VV* represents a diphthong. In general, vowel and syllable durations correlate negatively, but in a few places the order is not matched exactly. The vowel in the *CV* structure is shorter than *CVV* and *VC* because the presence of an initial consonant plays a major role in shortening the duration of the vowel. Moreover, everything being equal, *VV* is longer than *V*, as expected.

Vowel and coda consonant also exhibit compensatory effect. Figure 4 shows that the velar nasal coda

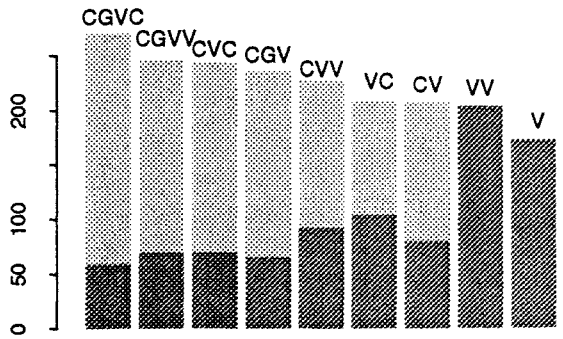

Figure 3: Vowel/syllable length by syllable type

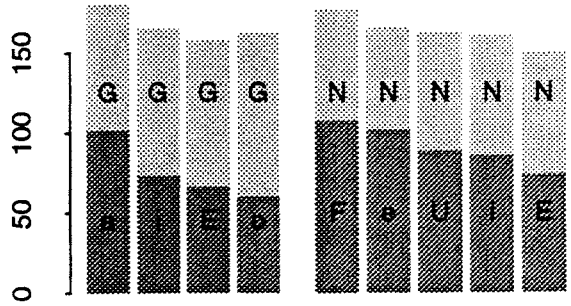

Figure 4: Compensatory effect: vowel and coda

G is consistently longer than the alveolar nasal coda *N* (91 msec vs. 71 msec), and vowels before the velar coda are shorter. The compensatory effect is also observed within each class. Given the same coda, longer vowels tend to be accompanied by a shorter coda. It is also clear that, again, the compensatory effect is not complete. The longest vowel and coda combination comes from the longest vowel *a* and the longer coda *G*; the shortest combination comes from *N* and *E*, the shortest vowel that co-occur with *N*.

Lack of Utterance Final Effect One surprising finding is that there is no utterance-final lengthening effect. The mean duration of utterance-final syllables is 207

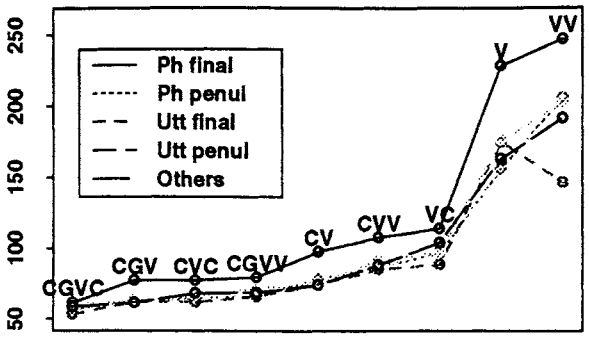

Figure 5: Lack of utterance-final lengthening

msec, which is shorter than the mean of the utterance-penultimate syllables (221 msec). In contrast, the mean duration of phrase-final syllables (utterance-final excluded) is 254 msec, which is considerably longer than the mean duration of phrase-penultimate syllables (216 msec), and the mean duration of all the non-final, non-penultimate syllables (214 msec). We found the same pattern looking at vowel durations. Figure 5 plots the vowel durations broken down by syllable type. Consistently, utterance-final vowels are comparable to, or even shorter, than utterance-penultimate, phrase-penultimate, and the non-final, non-penultimate vowels, while phrase-final vowels are consistently longer.

On the surface, our finding seems to contradict previous reports on final-lengthening effect (Klatt, 1975; Edwards and Beckman, 1988; Berkovits, 1993). However, since there is at the same time considerable amount of phrase-final lengthening in our database, we interpret our data to be consistent with previous findings. Sentences used in previous experimental studies were comparable in size to our phrases. Some Discourse studies (Klatt, 1975; Crystal and House, 1988) actually reported on phrase-final effect because there were very few samples of discourse final syllables. Our sentences are more comparable to short paragraphs, consisting of several phrases and exhibit full discourse structure, with dramatic discourse-final lowering toward the end. We suspect that the lack of discourse-final lengthening is linked to the dramatic drop in f0 and amplitude.

Conclusion

One of the most important differences of this study from previous studies on Chinese duration is the design and the versatility of the database. In general, it is possible to investigate any factor and the interaction of factors that have been coded or can be coded easily, provided that there are sufficient number of data points in the database representing the factor in question.

Some of the major findings from this study include the strong intrinsic scales of many categories of sound. We reported vowel and fricative scales in this paper. We also find incomplete compensatory effects, and the lack of utterance-final lengthening.

ACKNOWLEDGMENTS

We acknowledge ROCLING for providing us with the text database. We also wish to thank Jan van Santen for duration analysis tools and extensive advice.

REFERENCES

J. Allen, S. Hunnicut, and D. H. Klatt. 1987. *From text to speech: The MITtalk system.* Cambridge University Press, Cambridge, UK.

Rochele Berkovits. 1993. Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, 21(4):479–489.

T. H. Cormen, C. E. Leiserson, and R. L. Rivest. 1990. *Introduction to algorithms.* The MIT Press, Cambridge, Massachusetts.

T. H. Crystal and A. S. House. 1988. Segmental durations in connected-speech signals: current results. *JASA*, 83:1553–1573.

J. Edwards and M.E. Beckman. 1988. Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica*, 45(2):156–174.

Long Feng. 1985. Beijinghua yuliu zhong sheng yun diao de shichang (duration of consonants, vowels, and tones in Beijing Mandarin speech). In *Beijinghua Yuyin Shiyanlu (Acoustics Experiments in Beijing Mandarin)*, pages 131–195. Beijing University Press.

R. M. French, A. Greenwood, and J. P. Olive. 1993. Speech segmentation criteria. Technical report, AT&T Bell Laboratories.

D. H. Klatt. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3:129–140.

Hongmo Ren. 1985. Linguistically conditioned duration rules in a timing model for Chinese. In Ian Maddieson, editor, *UCLA Working Papers in Phonetics 62*, pages 34–49. UCLA.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1994. A stochastic finite-state word-segmentation algorithm for Chinese. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 66–73. New Mexico State University.

Jan P. H. van Santen. 1992. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546.

Jan P. H. van Santen. 1993. Perceptual experiments for diagnostic testing of text-to-speech system. *Computer Speech and Language*, 7(1):49–100.