Duration Study for the Bell Laboratories Mandarin Text-to-Speech System

Chilin Shih
Benjamin Ao

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

1 Introduction

This paper reports the design and results of a study of Mandarin Chinese segmental durations. The project is part of a Mandarin text-to-speech system, and our primary goal is to model the duration pattern of natural speech to improve the naturalness of the text-to-speech system. An ideal duration study for a text-to-speech system should investigate all Mandarin speech sounds in all contexts, and capture all the factors that affect duration. Such a goal, of course, is impossible to reach. Therefore the initial step of our task is to decide how to scale down the scope of our study so the task will be manageable without losing crucial information.

Previous duration studies on Mandarin took the form of controlled experiments, where a limited number of contextual factors were examined in a fixed sentence frame [Fen85, Ren85]. Controlled experiments provide excellent contrast from minimal pairs, which are useful in identifying factors that lead to durational variations. A practical concern with this approach is that it is not possible to cover all factors and their interactions. Furthermore, experiments are better suited to answer pre-defined, well-focused questions, while offering very little about factors that are not included in the experimental design. Speech databases [Kla73, Ume77, CH82] are more versatile and better suited for exploratory studies. However, small databases are limited in scope while the construction of large databases is

extremely time-consuming. Moreover, the lack of controlled environments makes it more difficult to ascertain the effects of factors.

To circumvent the problems of both experimental studies and speech databases, we follow the methodology proposed by van Santen. A greedy algorithm [vS93] is used to select text from an on-line corpus to minimize the size of the database without sacrificing the richness of contextual factors. Detailed coding allows us to compare near minimal pairs in the database. Statistical methods [vS92a, vS94] that take advantage of the intrinsic scale of various categories of sound allow us to estimate the duration of a speech sound in a context that is not present in our database. The combination of these choices makes it possible to construct a durational model for the text-to-speech system with satisfactory performance within a relatively short time frame. Due to the nature of a text-to-speech system, we can further control the size of the database by limiting our investigation to one speaker, and primarily to those factors that can be predicted from text.

2 Database

The source of our database is the ROCLING Chinese text corpus, which consists of over 9 million characters of newspaper text collected during a six month period from October 1979 to March 1980 in Taiwan. Aside from news articles, there are also mixed genre texts in the corpus, such as essays, short stories, and kung-fu fiction.

We first extracted from the ROCLING corpus 15620 sentences and short paragraphs that were between 25 to 50 characters long, each sentence (or paragraph) containing several phrases. The character strings were segmented into words and transcribed into phonetic representation using an automatic segmenter [SSGC94]. We represent Mandarin sounds with a system that is very similar to *pinyin*, the official transliteration system used in China. However, when a pinyin symbol is ambiguous or uses two letters, we assign a unique, one letter symbol. Table 1.1 gives the correspondence between pinyin and our notation where there is a difference. If a pinyin symbol is ambiguous, we provide a disambiguating environment in parentheses.

All Mandarin stops and affricates are voiceless, for example, *b, d, g* represent voiceless unaspirated stops. *S, C, Z* represent the retroflex fricative, retroflex unaspirated affricate, and retroflex aspirated affricate respectively. *J* is a retroflex vowel, which only occurs with retroflex consonants. *Q* is a central apical vowel, which only occurs with the dental affricates *z, c* and the dental fricative *s*. *U* is a high front rounded vowel. *R* is a heavily retroflexed vowel with the unique property that it must be the only sound in a syllable; it does not co-occur with any initial or coda consonants. *F* is an allophone of *a*, which is fronted and raised in the context of a following alveolar nasal

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	(i)n	ng
Our Symbol	W	y	w	Y	S	C	Z	N	G

TABLE 1.1. Conversion Chart of Symbols

N. E is our symbol for schwa. Even though Mandarin de-stressed vowels are often reduced to schwa, this vowel, unlike the schwa in English, can be fully stressed (i.e., carrying full tone). Diphthongs are treated as single units. Our symbols *A, I, O, W* represent pinyin *ei, ai, ou, ao* respectively. Syllable final consonants (i.e., codas) in Mandarin are very restriced. Only alveolar nasal *N* and velar nasal *G* are allowed in that position.

Every segment in the on-line text was coded with a set of factor values before the search began. Based on previous reports on Mandarin duration [Fen85, Ren85] and literatures on other languages [Noo72, Leh72, Kla73, Oll73, HU74, Por81, CH82, AHK87, CH88, vS92a, WSOP92, FM93] we choose the following factors as the focus of our investigation.

1. Identity of the current segment (46)
2. Identity of the current tone (6)
3. Identity of the previous segment (10)
4. Identity of the previous tone (6)
5. Identity of the next segment (10)
6. Identity of the next tone (6)
7. Degree of discourse prominence (3)
8. Number of preceding syllables in the word (3)
9. Number of following syllables in the word (3)
10. Number of preceding syllables in the phrase (3)
11. Number of following syllables in the phrase (3)
12. Number of preceding syllables in the utterance (2)
13. Number of following syllables in the utterance (2)
14. Syllable type (9)

Factor 1 has 46 values that correspond to 46 segments, including 15 vowels, 4 diphthongs, 3 glides, 21 consonants, and 3 coda consonants. Factors 2, 4 and 6 each has 6 values that correspond to the 4 full tones, the neutral tone (0) and a sandhi tone (5). Factors 3 and 5 each groups sounds into 10 categories. Factor 7 has three values: normal reading, some prominence, and strong prominence. Factors 8 through 11 have three values each, 0, 1 and 2, where 0 means that the segment in question lies at the boundary, 1 means that it is one syllable away, and 2 means that it is 2 or more syllables away from the boundary. Factors 12 and 13 have two values each, 0 and 1, where 0 means that the segment lies at the boundary, 1 means that

it is 1 or more syllable away from the boundary. Factor 14 has 9 values, corresponding to 9 syllable types.

Factor 7 on discourse prominence cannot be calculated from text information alone; this factor was not included in the input coding for text selection. The values of this factor were obtained later on by transcribing the recorded database.

During the text selection phase, phrasing was coded solely on the basis of punctuation. After the text was selected and the database recorded, phrasing was re-coded to correspond to pauses. Each paragraph-sized unit chosen from the corpus was considered an utterance, and each utterance in our database contains at least 2 phrases. There are 3845 phrases in total, so on average each utterance contains 9 phrases.

There is no factor coding word stress directly, because word stress in Chinese is not as clearly defined acoustically or perceptually as in a stress language such as English. De-stressing, however, is clear and is traditionally described as a process of tonal reduction. So in effect the factor on current tone also coded two levels of stress: full-toned syllables (1-5) are stressed, while neutral-tone syllables (0) are de-stressed.

The factor values of each segment are grouped into factor-triplets, a unit we judged to be an acceptable compromise between controlling the number of possible factor combinations and preserving interesting factor interaction. Each factor-triplet consists of the current segment, the current tone and one of the other 11 factors (factor 7 excluded). Each segment in the on-line text is now represented by 11 factor-triplets. These factor-triplets represent types of interaction that we are interested in and will be deliberately searching for. The following example illustrates how a sentence in the on-line text is transcribed and coded into factor-triplets. The first factor triplet *x_2_b** means that this is a segment *x*, occurring in a tone 2 syllable, and is preceded by silence.

```
Sample Text:

刑事組幹員認為幕後可能有販毒集團，乃喬裝購毒品，串通被補的
廖清裕打電話給林，李兩人，約好交貨時間地點.

Word Segmentation and Transcription:

x2iGS4J z3u g4FNY2eN r4ENw2A m4uh4O k3En2EGy3O f4FNd2u
j2itw2FN } n3I qy2WZw1aG g4Od2up3iN } Cw4FNt1oG b4Ab3udOE
ly4Wq1iG4U d3ady4eNhw4a g3A l2iN } l3i ly3aGr2EN } Y1eh3W
jy1Whw4o S2Jjy1eN d4idy3eN }

Factor Triplets:

x_2_b* x_2_B* x_2_f3 x_2_F1 x_2_w0 x_2_x1 x_2_p0 ...
```

```
i_2_b2 i_2_B* i_2_f9 i_2_F1 i_2_w0 i_2_x1 i_2_p0 ...
G_2_b6 G_2_B* G_2_f1 G_2_F1 G_2_w0 G_2_x1 G_2_p0 ...
S_4_b9 S_4_B1 S_4_f3 S_4_F1 S_4_w1 S_4_x0 S_4_p1 ...
J_4_b2 J_4_B1 J_4_f0 J_4_F1 J_4_w1 J_4_x0 J_4_p1 ...
...
```

There were a total of 1,385,451 segments, or 556,353 syllables, 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 use a greedy algorithm [vS92b] to search through the 15,620 sentences. During each search a sentence is selected if it contains the most factor-triplet types that had not yet been covered. In other words, every sentence is chosen for some unique factor-triplets contained therein, at least at the time it is chosen. Redundant sentences in the sense of factor coverage are effectively eliminated, therefore drastically reducing the size of the recorded database without sacrificing factor coverage. In our case, the search terminated after 427 sentences were chosen when 100% of the input factor-triplets were covered. These sentences are long, each comprising of several phrases and are for all practical purpose similar to short paragraphs. [1] The 427 chosen sentences/paragraphs contain 38,881 segments, 19,150 syllables, and each of the 8,233 factor-triplets occurs at least once.

Figure 1 compares the performance of the greedy algorithm to random selection of text. The effectiveness of a greedy algorithm is apparent. While 427 sentences selected by the greedy algorithm cover 100% of the factor-triplets that are present in the input, 427 randomly selected sentences covers only 74%. If we accept 74% coverage, 42 sentences selected by the greedy algorithm will be sufficient. As more sentences are accumulated, most of the frequently occurring factor-triplets were covered and it becomes increasingly difficult to find a new one. The last 129 sentences chosen by the greedy algorithm each added just one new factor. In comparison, the slow increase is still much better than random selection, where many sentences merely repeats the freqent factors that have already been covered. From the 427th to the 1000th randomly selected sentences, there were only 3% increase in the coverage of factor-triplet.

After manual correction of transcription and word segmentation errors, the selected sentences were recorded by a male native Beijing Mandarin speaker in a sound-proof room, using a Brüel and Kjær microphone 2231. The transcription was edited once again to match the recorded speech. Phrasing and prominence levels were also transcribed to match the reading. The recorded speech was then manually segmented using Waves (Entropic

[1] Three of the 427 sentences turn out to be incomplete and do not make sense in isolation and were taken out. The recorded database therefore contains 424 sentences. Furthermore, a few awkward phrases were edited to facilitate fluent reading.

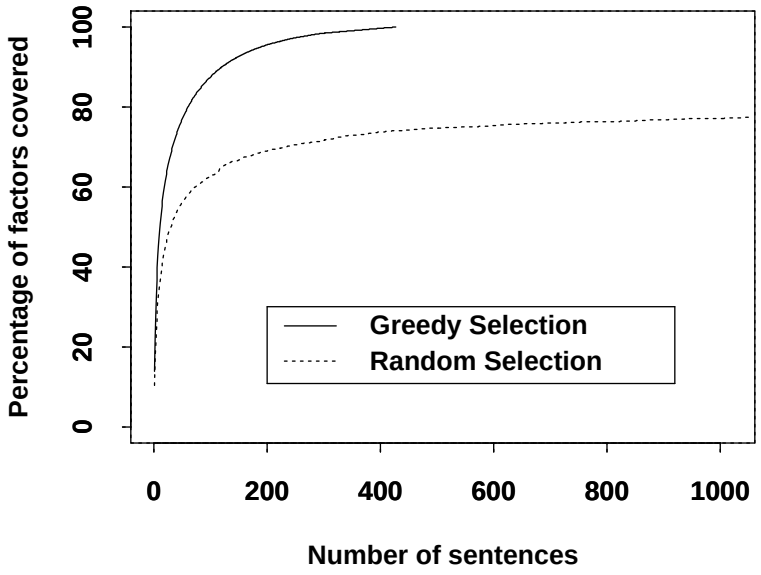

FIGURE 1. Comparison Between Random and Greedy Sentence Selection

Inc.) on an SGI Indigo workstation.

The segmentation of the speech data followed a set of rigid rules [FGO93, OGC93]. We used both the spectrogram and the waveform to determine segment boundaries, and listened to the speech to confirm the placement of boundaries. Typically, segment boundaries were placed when a sudden change in the formant structure was visible. When such a location could not be found, as in the middle of adjacent vowels, the segment boundary was placed at the energy minimum in the transitional region. When no acoustic cues can be found, as it sometimes happens between two identical vowels, the boundary was placed at the midpoint between the two vowels. The boundaries of obstruent consonants were usually easy to identify. Closure and release portion of all plosives were measured separately. The closure portion of an utterance- or phrase-initial plosive, which coincides with silence, was always marked as having no duration of its own and the data was discarded in later analyses. The burst-aspiration duration was measured from the onset of the burst to the onset of the vowel.

After segmentation, the factor values and duration of each segment were coded again into a matrix form suitable for statistical analyses. Excluding pauses and phrase initial closure duration, the final database we used for statistical analyses had 46,265 lines, each line containing the factor values and the duration of a segment. We divided the database into six groups for the purpose of analysis, because each group of sounds may respond to a given factor in different ways: 1. vowels, 2. fricatives, 3. burst and

aspiration of plosive, 4. closure of plosive, 5. sonorants, and 6. syllable coda consonants.

3 Duration Model

In most durational studies, results are analyzed using the raw means of duration measurements. This is fine as long as the experiments are carefully controlled and the factors involved are balanced. The raw mean could be quite misleading in a natural speech database like ours where the frequency distribution is unbalanced. A segment may occur in some environments more often then in others and bias the result. For example, the retroflex vowel *R* (pinyin *er*) is the longest vowel if we look at the raw mean. However, the mean duration of *R* turns out to be artificially long because *R* can only occur in the syllable structure *V* due to a phonotactic constraint, and *V* is the syllable type that yields the longest vowel duration in Mandarin (see Section 3.1). The absence of short samples from other syllable types is responsible for the long *R* in the raw mean. When we isolate the *V* syllable type and calculate its raw mean, *R* is shorter than low vowels and diphthongs.

To avoid the problems described above, we follow Van Santen and use corrected means, where estimated durations are corrected for the effect of various factors. Segment durations are compared to others that occur in the same coded conditions. The idea is similar to comparing the duration of segments in controlled experiments. We refer the readers to van Santen [vS92a] for mathematical proof and calculation.

When there are too many unevenly distributed gaps in the data matrix, it will not be possible to estimate corrected means. These gaps could be the result of either phonotactic constraints in Mandarin, or accidental gaps in the database. In order to reduce the empty cells, we need to collapse some levels of certain factors. This procedure was carried out very carefully, consulting information from raw mean, number of samples, standard deviation, and sometimes t-test. Only levels that are phonetically similar and affect durational variation in similar ways are collapsed. For example, when we coded our data for statistical analysis, we revived the identity of the preceding and following segments in order to investigate possible effect from individual sounds. After carefully examining the effect of each preceding sound on the following vowel, we collapsed the values of the *preceding segment* for the vowel category from 46 to 11. The resulting values are high vowel, mid vowel, low vowel and diphthong, coda consonant, sonorant, glide, aspirated stop, nonaspirated stop, aspirated affricate, nonaspirated affricate, and fricative.

We perform a number of analyses after combining factor levels, such as computing corrected means by each factor, computing two-way corrected means to investigate the pattern of interaction of any two factors that is of

o	J	E	Q	i	U	u	e
99	109	113	116	120	121	121	128
R	A	O	I	F	W	a	
134	135	138	147	149	155	160	

TABLE 1.2. Corrected Means of Vowels in Msec

interest to us, and build additive and multiplicative models by computing the estimated intrinsic durations of segments, and the coefficient of contextual factors. The multiplicative model in general performs better than the additive model, so in the following we only report the result from the multiplicative model, where $Dur_{i(f2,\ldots,fn)}$ is the predicted duration of a given vowel i with factor levels $f2,\ldots,fn$ for factors $F2,\ldots,Fn$ respectively. $IDur_i$ is the intrinsic duration of the vowel, or $F1$, and $F2_{f2},\ldots,Fn_{fn}$ are the coefficients of the other factor levels.

$Dur_{i(f2,\ldots,fn)} = IDur_i \times F2_{f2} \times F3_{f3} \times \ldots \times Fn_{fn}$

Our results agree with well-known durational phenomena reported in the literature in general terms, but often with refinement on details. For example, [Oll73] found that vowels are lengthened in final position and consonants are lengthened in initial position. In our data, consonant-lengthening is found in all initial positions, and most strongly word-initially, but vowel lengthening is found only in the phrase final position but not in the word final and utterance final positions.

Among the 14 factors that we investigated, the identity of the following tone is the only one that shows nearly no effect on all six classes of sound. The factors that consistently have a strong effect on all classes of sounds are the identity of the current segment and prominence. All the other factors have some effect on some classes but not on others.

3.1 Vowels

Our data shows very clear patterns of intrinsic duration of vowels [Hou61, AHK87, vS92a]. For example, we observed the same scale of vowel duration under various degree of prominence and in different positions of a phrase. The shortest vowel is *o*, followed by the two apical vowels *J* and *Q* and the schwa *E*; all of them are shorter then high vowels *i*, *u*, and *U*. Diphthongs *A*, *O*, *I*, and *W* are longer than high and mid vowel, while the longest vowel is the low vowel *a*. The best estimates of corrected means of vowels for the entire dataset is given in Table 1.2.

Aside from vowel identity, the following are the most important factors that affect vowel duration in the multiplicative model. We rank the importance of the factors by a index number, which is the ratio of the two extreme levels of the factor. For example, the index number 1.82 for the

factor *prominence* is obtained by dividing the highest coefficient 1.29 *(level 2)* by the lowest coefficient 0.71 *(normal)*. Given the multiplicative model, a vowel with prominence level 2 will be 1.82 times longer than a normal vowel.

1. **Syllable type** (1.89): open syllable without glide > open syllable with glide > closed syllable without glide > closed syllable with glide
2. **Prominence** (1.82): level 2 > level 1 > normal
3. **Previous phone** (1.73/1.27): 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

The following factors have some effect on vowel duration:

1. **Identity of tone** (1.49/1.11): full tone > neutral tone; among full tones, 3 > 2 > 4 > 1
2. **Utterance position** (1.39): nonfinal > final
3. **Following phone** (1.33): diphthong > monophthong > plosive and fricative > sonorant
4. **Phrasal position** (1.31): final > nonfinal

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

Two index numbers were given for the factors *previous phone* and *tone*, the first numbers, 1.73 and 1.49, are derived in the usual way. These numbers are high primarily because the two factors in question incorporate complex conditions. The *previous phone* of a vowel can be an initial consonant within the same syllable, or it can be the last phone of the previous syllable. The second index number (1.27) excludes the across-syllable conditions, therefore reflects the effect of the initial consonants on vowels. In the factor *tone*, the high index number is caused by the level *tone 0*, which, as explained earlier, refers to the absence of a full tone and most closely resembles the phenomenon of de-stressing. The second index number (1.11) excludes *tone 0* and reflects the range of the effect of full tones.

The intrinsic scale of vowels from our study is slightly different from the report of [Fen85], where the duration of the mid vowel *o* is similar to that of *e*. The discrepancy in *o* is due to different segmentation strategies: we segmented syllables such as *mo* as having three segments *mwo*, while [Fen85] treats it as having two segments, combining the *w* portion into the duration of *o*, causing the *o* duration to be artificially long. There is another environment where *o* occurs in Mandarin: before a nasal coda and

h	f	S	x	s
98	100	113	119	122

TABLE 1.3. Corrected Means of Fricatives in Msec

without a glide, as in the syllable *gong*. That is an environment where the segmentation issue wouldn't be a problem, but [Fen85] didn't study *o* in this environment. We have a complete set of data on vowels occurring before a nasal coda. The result of that subset of data also confirms that *o* is the shortest vowel.

The fact that the utterance-final vowels have considerable lower coefficients than non-final vowels does not necessarily mean that there is a utterance-final shortening effect in Mandarin. Since the end of an utterance is by definition the end of a phrase, we coded utterance-final vowels as being phrase-final as well. As a result, utterance-final vowels would be lengthened in the model due to their phrase-final status. When in reality there is no utterance-final lengthening effect (see Section 4.2), comparable level of shortening for this position needs to be built into the model to offset the lengthening effect associated with the phrase-final position.

3.2 Fricatives

Among fricatives, *h* and *f* are short, while *s* and *x* are long. Table 1.3 gives the best estimates of corrected means of fricatives.

The important factors affecting fricative duration are:

1. **Following phone** (1.37): high vowel > mid vowel > low vowel
2. **Prominence level** (1.36): with prominence > normal
3. **Position in the word** (1.25) : initial > non-initial
4. **Tone** (1.23): full tone > neutral tone
5. **Syllable type** (1.21) : syllable without glide > syllable with glide

All other factors have index numbers smaller than 1.15. The factors that have nearly zero effect include the following tone, the number of following syllable in the utterance, and the previous phone.

3.3 Burst and Aspiration

Intrinsic burst-aspiration duration is given in Table 1.4. Not surprisingly, manner of articulation is the most important factor determining the length of the burst and aspiration: Unaspirated stops and affricates have shorter

b	d	g	Z	z	j	p	t	k	C	c	q
11	13	21	29	43	46	80	80	86	95	99	113

TABLE 1.4. Corrected Means of Burst-Aspiration Duration in Msec

burst-aspiration duration than aspirated ones, and in either the aspirated or the unaspirated category, stops have shorter bursts/aspiration duration than affricates. Place of articulation has a consistent effect, but the effect is small in comparison to the variations caused by manner of articulation. Among stops, bilabials have shorter burst-aspiration duration than alveolars, which in turn have shorter burst-aspiration duration than velars. Among affricates, the retroflex affricates have shorter burst-aspiration duration than dentals, which in turn are shorter than palatals.

Phone identity, with an index number of 10.06, is undisputably the most important factor controlling the duration of the burst and the following aspiration or frication. The next important factor is *the following phone*, with an index number of 1.84.

1. **Following phone** (1.84) : apical vowel > high vowel > low vowel

Other factors that have some effect on burst-aspiration duration include:

1. **Position in word** (1.23) : initial > non-initial
2. **Tone** (1.20): tone 2 > others
3. **Prominence level** (1.19) : level 2 > level 1 > normal
4. **Syllable type** (1.16): without glide > with glide
5. **Preceding phone** (1.15): high vowel > diphthong > apical vowel > nasal coda > low vowel

The preceding and the following tone, the number of preceding syllables in the phrase, and the number of following syllables in the word, the phrase and the utterance have little effect. It is unclear why tone 2 lengthens the burst-aspiration duration. It could be a matter of personal style.

3.4 Closure

The intrinsic closure duration is given in Table 1.5. The manner of articulation, again, plays a major role: affricates have shorter closure duration than stops, and aspirated ones have shorter duration than unaspirated ones.

Factors affecting the closure duration include:

1. **Position in word** (1.37): initial > non-initial

Ccl	ccl	qcl	zcl	tcl	jcl	Zcl	kcl	dcl	gcl	pcl	bcl
13	13	15	15	16	16	17	18	18	19	20	21

TABLE 1.5. Corrected Means of Closure Duration in Msec

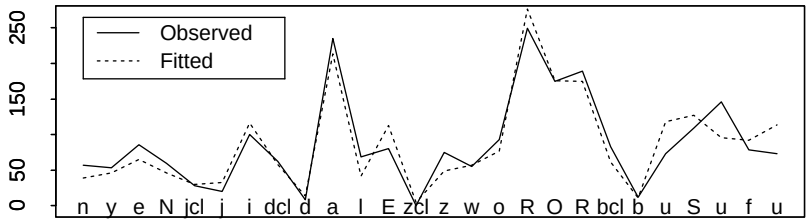

FIGURE 2. Comparison of Observed and Fitted Duration

2. **Tone** (1.31): full tone > reduced tone
3. **Prominence level** (1.17): with prominence > normal
4. **Preceding phone** (1.14): vowel > nasal
5. **Following phone** (1.10): high vowel > low vowel

Preceding and following tones, the position in the utterance, and syllable type have little effect.

3.5 Fitted duration

Even though our speaker read the database in a dramatic style, with frequent shift of speaking rate and liberal usage of exaggeration, the performance of our predictive model is very good. The root mean square of the difference between the observed and the fitted values by the multiplicative model is 25 msec. Figure 2 compares the natural segmental duration (in solid line) and the fitted segmental duration (in dash line) of the shortest sentence in our database: *Nian2-ji4 da4 le0, zuo5-er3 ou5-er3 bu4 shu1-fu0*, "Due to (my) old age, my left ear sometimes feels uncomfortable." (SEEAUDIO).

The duration of the first vowel nucleus *e* (spelled as *a* in *nian*) is derived as follows:

$$Dur_e(65msec) = IDur_e(127.94msec)$$

$$\times F2_{tone[2]}(1.075) \times F3_{previous-phone[y]}(0.875)$$
$$\times F4_{previous-tone[null]}(0.996) \times F5_{next-phone[N]}(1.102)$$
$$\times F6_{next-tone[4]}(1.027) \times F7_{stress[0]}(0.709)$$
$$\times F8_{preceding-syllable-in-word[0]}(1.012)$$
$$\times F9_{following-syllable-in-word[1]}(0.975)$$
$$\times F10_{preceding-syllable-in-phrase[0]}(1.016)$$
$$\times F11_{following-syllable-in-phrase[2]}(0.911)$$
$$\times F12_{preceding-syllable-in-utterance[0]}(0.963)$$
$$\times F13_{following-syllable-in-utterance[2]}(1.12)$$
$$\times F14_{syll-type[cgvc]}(0.688)$$

4 Discussion

Our result confirms previous findings on duration in general areas. The duration scale of vowel categories and consonant categories are similar to those from previous Mandarin studies [Fen85, Ren85], even though the database and the methodology are quite different. Since our database is much more extensive, we are able to explore areas that have not been studied before. We discuss two interesting cases below: (incomplete) compensatory effect, and the lack of discourse final lengthening.

4.1 Compensatory Effect

We use two examples to illustrate the compensatory effects: vowels and other segments in a syllable, and vowels and coda consonants. The vowel duration in a syllable 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*. We plot the raw mean duration of vowels by syllable type in the top panel of Figure 3 in ascending order. With the exception of *CVC* and *CGVV*, the differences between all adjacent pairs are significant at the $p < 0.001$ level. The presence of an initial consonant, a glide, or a coda consonant in a syllable shortens the main vowel. Everything being equal, a diphthong (VV) is longer than a simple vowel.

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 raw mean duration of syllable length by type is plotted in the bottom panel of Figure 3 in descending order. The duration of the longest syllable type, CGVC, is 1.5 times the duration of the shortest one, V. The differences among the two-segment group *VC, CV* and *VV* are not significant. Also, the difference between

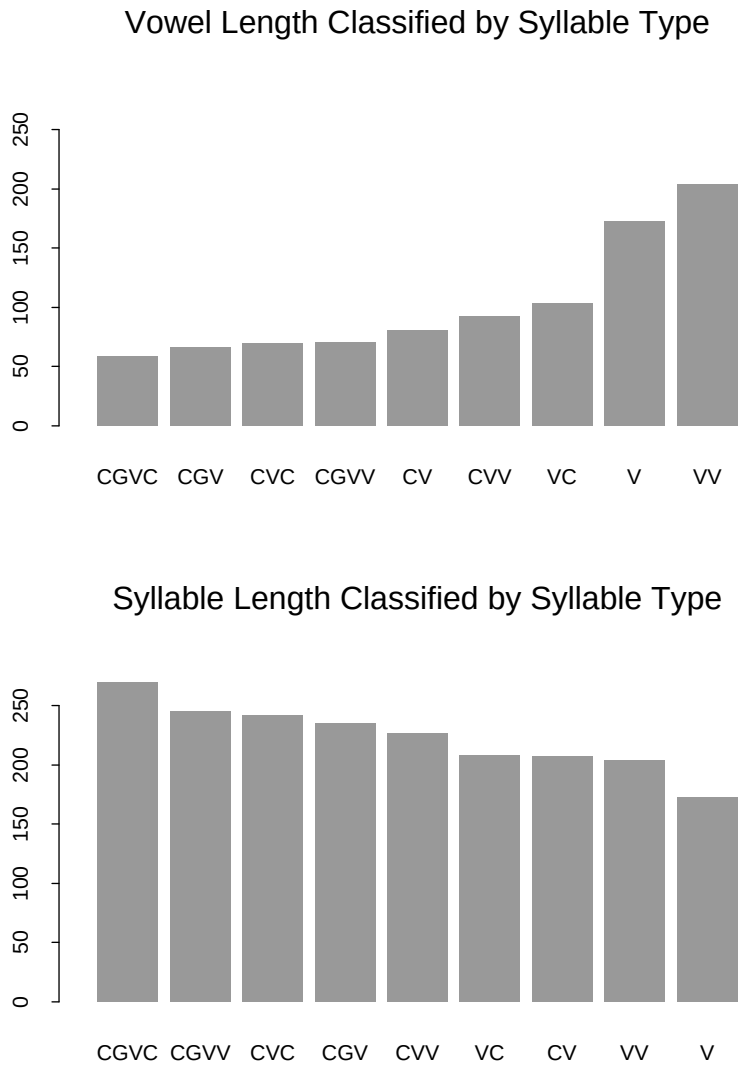

FIGURE 3. Vowel and Syllable Length Classified by Syllable Type

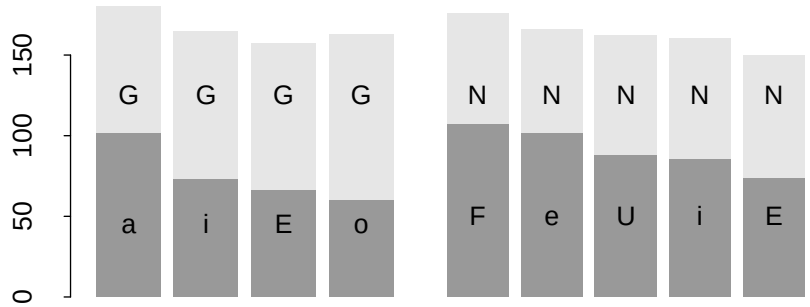

FIGURE 4. Compensatory Effect Between Vowel and Coda

Utt final	Utt penul	Phr final	Phr penul	Others
207	221	254	216	214

TABLE 1.6. Final and Non-final Syllable Duration

CGVV and *CVC* is not significant. The differences between all other pairs are significant at the $p < 0.001$ level.

Vowel and coda consonants also exhibit compensatory effect. See Figure 4. The velar nasal coda *G* (91 msec) is longer than the alveolar nasal coda *N* (71 msec), and the vowel before the velar coda is shorter. The compensatory effect is also observed within each class. Given the same coda, a longer vowel tends to be accompanied by a shorter coda. Again, the compensatory effect is not complete. The longest vowel and coda combination comes from the longest coda *G* and the longest vowel *a*; the shortest combination comes from the shortest coda *N* and *E*, the shortest vowel that co-occurs with *N*.

4.2 Lack of Utterance Final Effect

One clear finding from our study is that there is no utterance-final lengthening effect. Table 1.6 compares the raw mean duration of utterance final syllables with some other conditions. The mean duration of utterance-final syllables is 207 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 position and syllable type. Consistently, utterance-final vowels are comparable to utterance-penultimate,

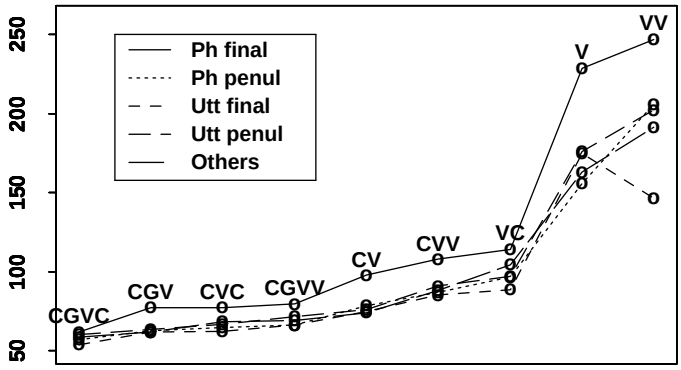

FIGURE 5. Vowel Duration Classified by Position and Syllable Type

phrase-penultimate, and the non-final, non-penultimate vowels (the *other* condition), while phrase-final vowels are longer. There is only one sample of utterance-final VV syllable in our database, so its low value should be taken with a grain of salt. The fact that the phrase-final vowels are longer while the phrase-penultimate vowels are similar to other vowels suggests that the domain of final-lengthening is confined to the last syllable of the phrase. If there were an utterance-final effect, we would expect to see a difference between the duration of the utterance-final and the utterance-penultimate vowels.

Our result is most similar to the sentence-final shortening effect of Japanese [Tak89]. Interestingly, they also found a mixture of effects. In conversational speech there are lengthening effects in both the prepausal position (similar to our phrases-final position) and in sentence-final position (similar to our utterance-final position). However, in read speech there are lengthening effects in phrase-final position but shortening effects in sentence-final position. Our database contains read speech only, therefore the finding is entirely consistent with [Tak89].

On the surface, our finding seems to contradict many previous reports on final-lengthening effect [Kla75, EB88, Ber93]. 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. The final-lengthening effects reported in some discourse studies [Kla75, CH88] were actually phrase-final effect; there were very few samples of discourse final syllables in those two studies. 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.

5 Conclusion

One of the most important difference of this study from previous studies on Chinese duration is the design of the database. The major advantage of our methodology is the efficiency of the database. Using a greedy algorithm to select text produced a small database which is rich in factors that are relevant to duration studies. Moreover, our database is not limited to the chosen factors and turns out to be an excellent source for exploratory study. More factors are collected as a by-product of collecting the specified factors, while some others may be created as the result of the reader's rendition of the text, one example is the variation in prominence.

Another encouraging aspect is the degree our result confirms previous findings in well-known areas, suggesting that no discrepancies are introduced by the differences in materials and in statistical methods. Against that background, we are confident in interpolating our result to previously untapped areas.

The major findings from this study include the intrinsic scales of all categories of Mandarin sounds, and the major factors affecting their durations. We reported the scales of vowel, fricative, burst-aspiration duration, and closure duration, and ranked the effects of 14 factors on them. 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. It would be impossible to do this project without his extensive advice and duration analysis tools.

6 References

[AHK87] J. Allen, S. Hunnicut and D. H. Klatt. 1987. *From text to speech: The MITalk system.* Cambridge University Press, Cambridge, UK.

[Ber93] Rochele Berkovits. 1993. Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, 21(4):479–489.

[CG86] R. Carlson and B. Cranström 1986. A search for durational rules in a real-speech data base. *Phonetica*, 43:140–154.

[CH82] T. H. Crystal and A. S. House. 1982. Segmental durations in connected speech signals: preliminary results. *JASA*, 72:705–716.

[CH88] T. H. Crystal and A. S. House. 1988. Segmental durations in connected-speech signals: current results. *JASA*, 83:1553–1573.

[EB88] J. Edwards and M. E. Beckman. 1988. Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica*, 45(2):156–174.

[Fen85] 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.

[FM93] J. Fletcher and A. McVeigh. 1993. Segment and syllable duration in Australian English. *Speech Communication*, 13:355–365.

[FGO93] R. M. French, A. Greenwood, and J. P. Olive. 1993. Speech segmentation criteria. Technical report, AT&T Bell Laboratories.

[HU74] M. S. Harris and N. Umeda. 1974. Effect of speaking mode on temporal factors in speech: vowel duration. *JASA*, 56:1016–1018.

[Hou61] A. S. House. 1961. On vowel duration in English. *JASA*, 33:1174–1178.

[Kla73] D. H. Klatt. 1973. Interaction between two factors that influence vowel duration. *JASA*, 54:1102–1104.

[Kla75] D. H. Klatt. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3:129–140.

[Leh72] I. Lehiste 1972. The timing of utterances and linguistic boundaries. *JASA*, 51(6.2):2018–2024.

[LR73] D. Lindblom and K. Rapp. 1973. Some temporal regularities of spoken Swedish. *Publication of the Institute of Linguistics, University of Stockholm*, 21:1–59.

[Noo72] S. G. Nooteboom. 1972. *Production and perception of vowel duration.* University of Utrecht, Utrecht.

[Oll73] D. K. Oller. 1973. The effect of position in utterance on speech segment duration in English. *JASA*, 54:1235–1247.

[OGC93] J. P. Olive, A. Greenwood, and J. Coleman. 1993. *Acoustics of American English speech: A dynamic approach.* Springer-Verlag, New York.

[Por81] R. F. Port. 1981. Linguistic timing factors in combination. *JASA*, 69:262–274.

[Ren85] 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.

[SSGC94] R. W. Sproat, C. Shih, W. Gale, and N. 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.

[Tak89] K. Takeda, Y. Sagisaka, and H. Kuwabara. 1989. On sentence-level factors governing segmental duration in Japanese. *JASA*, 86:2081–2087.

[Ume77] N. Umeda. 1977. Consonant duration in American English. *JASA*, 61:846–858.

[vS92a] Jan P. H. van Santen. 1992a. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546.

[vS92b] Jan P. H. van Santen. 1992b. Diagnostic perceptual experiments for text-to-speech system evaluation. pages 555–558. Proceedings of ICSLP.

[vS93] Jan P. H. van Santen. 1993. Perceptual experiments for diagnostic testing of text-to-speech system. *Computer Speech and Language*, 7(1):49–100.

[vS94] Jan P. H. van Santen. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8(2):95–128.

[WSOP92] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *JASA*, 91:1707–1717.