SYLLABLE DURATION IN MANDARIN

J. Wang

Speech, Hearing, and Language Research Centre
School of English and Linguistics
Macquarie University, Sydney.

ABSTRACT - A Mandarin speech database, aimed at establishing a prosodic model for Mandarin synthesisers, has been established in SHLRC at Macquarie University. The database contains 401 sentences (3168 syllables) read by a female speaker. The duration data was extracted and analysed using *mu+*. The statistical results have shown that the following durational control factors have a strong influence on syllable duration: syllable structure; syllable tone; stress pattern of prosodic words; syllable position in prosodic words; and prosodic word position in sentences. Of these, syllable structure, syllable tone (including unstressed syllables with neutral tone), and sentence-final position are more dominant. In order to quantify the effects of individual control factors, a simple, additive, syllable-based duration model was tentatively modelled.. The duration analysis and its modelling in this paper, to a large extent, reveals the prosodic and rhythmic characteristics of spoken Mandarin since tone, tone sandhi, word stress and rhythmic units were included in the hierarchically structured labels.

INTRODUCTION

Earlier studies have identified over seven factors that are known to affect the acoustic duration of segments or syllables in Mandarin. They are: (1) the nature of segments; (2) syllable structures; (3) syllable tone; (4) stressed/unstressed syllables; (5) the morpho-syntactic structure of polysyllables; (6) the position of segments in words and sentences; (7) intonation; and (8) speech rate (Feng, 1985; Ren, 1986; Shen, 1993). However, it is not possible to gain an overall picture of duration in Mandarin simply by using a limited number of word forms in carrier sentences that are only concerned with one or two of the above factors. Besides, there is insufficient statistical data concerning the significance of certain factors, or the interaction between factors, in the majority of this research. Over the past few years, speech databases and statistical approaches have been widely used to derive text-to-speech duration from natural speech or to develop duration models which illustrate how rhythm functions in various languages. This study involved establishing a Mandarin database of continuous speech, and quantitatively analysing the effects of certain dominant duration control factors which were then modelled using an initial syllable-based duration model.

MATERIALS AND ANALYSIS METHODS

The database for this study consists of 401 sentences (3168 syllables). The majority of these utterances are typical of the day-to-day language used by the author, who is also a native speaker. All sentences were recorded in SHLRC at Macquarie University using DATs, and the speech samples were then digitised onto a Sun computer and manually segmented and labelled using the *Waves* program.

The lowest level in this labelling system is not strictly acoustic-phonetic, since most Mandarin synthesisers choose initials/finals or syllables as synthetic units. Syllables in Mandarin have a structure of (C)V(N). So an initial, in Chinese linguistics, refers to the first consonant in a syllable (it can be a zero-initial if there is no consonant); and a final is the remainder of the syllable excluding the initial. Finals can be either simple vowels, diphthongs or triphthongs, or a combination of a vowel and tautosyllabic nasals like /an/ and /uan/. For initials, such as stops and affricatives, closure and release are labelled separately if there is a clear release mark. With reference to the segmentation criteria in the ANDOSL project, the initial boundary for stops and affricatives, regardless of whether they are aspirated or non-aspirated, is arbitrarily placed approximately 50-60 ms before the beginning of the release phase, although the actual closure time for both stops/affricatives and aspirated/non-aspirated consonants is probably different, with the mean duration being over 65 ms (Feng, 1985). The *Waves* program was used to display the speech wave and spectrogram to aid manual segmentation. Fundamental frequency (F0) was occasionally used to help identify the boundaries of syllables, especially the syllable-across vowel-vowel boundaries, because each syllable in Mandarin

carries a lexical tone which has a particular F0 contour. Above the acoustic-initial/final level is the syllable level, whereby initials and finals are grouped into syllables. Tone is the next level above the syllable level, since the tone system in Mandarin is syllable-based. Apart from labelling lexical tone, tone sandhi of 3rd tones (a 3rd tone becomes a 2nd tone when it is followed by another 3rd tone) and neutral tones carried by completely unstressed syllables are also labelled at this level, since there is an intrinsic relationship between tone, tone sandhi and unstressed syllables. Syllables are then grouped into feet and superfeet (prosodic words). The concept of foot and superfoot, and their basic formation rules, was adapted from prosodic approaches to tone sandhi (Shih, 1986; Cheng, 1990) and used to label prosodic words. The actual labelling of prosodic words often required aural differentiation when the rules could not be properly applied. Grouping syllables into prosodic words is fundamental for approaching Mandarin rhythm using continuous speech, and requires a knowledge of lexical words, morpho-syntax, sentence syntax, etc. The top level is the utterance level, since the prosodic phrase is not labelled at this stage for the sake of simplification.

Duration data was extracted and analysed using the *mu+* system, developed at Macquarie University for corpus based speech research (Harrington *et al* 1993). The duration of syllables, with certain structures and tones, in specific positions of prosodic words, with particular stress patterns, etc., can be easily extracted and calculated using this system.

ANALYSIS OF DOMINANT DURATION-CONTROL FACTORS

In this section, the influence of certain dominant duration-control factors is discussed. As described in the introduction, a number of factors affect the duration of segments and syllables in Mandarin, and some may have a stronger influence than others. The following discussion focuses on the effect of certain language-specific prosodic features, such as syllable-based tone and unstressed syllables in prosodic words, in determining dominate duration-control factors.

Durational compensation between initials and finals

In earlier studies, it was found that the acoustic duration of a final, to certain degree, is affected by its proceeding initial. This phenomenon was termed durational compensation between initials and finals (Feng, 1985; Ren, 1986). This compensation is strongly associated to the manner of initial consonants, especially aspiration and fricatives. Therefore, a final has the shortest acoustic duration when it is preceded by an aspirated affricative and second shortest by an aspirated stop, while the longest final occurs before a voiced initial. The compensation of acoustic duration most likely reflects the temporal overlap arising from the dynamics of gesturer interaction during production (Fowler & Saltzman, 1993) but is probably also related to keeping syllable duration constant for certain reasons as well (Feng, 1985). Table 1. gives the mean duration of initials and subsequent finals in stressed syllables, regardless of their classes.

Table 1. Mean duration (in ms) and standard deviations of each initial and final group in stress syllables.

		ZI	VI	UA	UA	FR	AS	AA
Initials	Mean		51	58	79	90	107	86
	Sd		17	17	23	38	32	35
	N		168	426	446	650	308	197
SVF	Mean	127	114	102	94	90	80	86
	Sd	31	32	37	24	38	26	36
	N	64	117	155	117	171	144	107
CVF	Mean	151	134	130	128	121	114	118
	Sd	30	30	28	30	26	31	28
	N	74	115	138	194	215	58	21
VNF	Mean	187	156	147	140	137	134	131
	Sd	42	34	28	29	29	30	35
	N	41	130	98	109	224	103	65

Most of the duration differences between the two groups are significant. Duration of finals in each group, despite having different inherent duration, have a similar degree of shortening when following initials in the same group, which is reflected by a highly negative correlation between the duration of initials and each final group (the correlation coefficients between initials, simple vowels, complex vowels and nasal finals are -0.973, -0.983 and -0.966 respectively; $p < 0.0001$), as well as a highly positive correlation among final groups (the correlation coefficient between simple vowels and

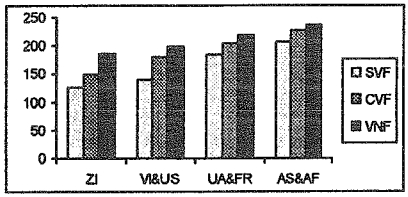
complex vowels is 0.975; simple vowels and nasal finals, 0.950; and complex vowel finals and nasal finals 0.966, $p < 0.01$) This provides strong evidence in support of temporal overlap, rather than of keeping syllable duration constant. For the sake of convenience, the following abbreviations are used in this paper: SVF = simple vowel finals; CVF = complex vowel finals; VNF = vowel-nasal finals; ZI = zero initials; VI = voiced initials; US = non-aspirated stops; UA = non-aspirated affricatives; FR = fricatives; AS = aspirated stops; AA = aspirated affricatives). Since the articulatory overlapping of initials and finals leads to a large variation in the acoustic duration of finals, it is better, not only for the sake of convenience, but also due to the linguistic position of syllables in Mandarin, to use the syllable as the synthetic unit in Mandarin syntheses. The following syllable-based duration model stems from this consideration.

Syllable structures

Apart from indicating the compensatory shortening of final duration with initial duration lengthening, Table 1. also shows that syllable duration varies according to the nature of initials and finals, although this compensation decreases the variation to a certain extent. There are over 400 different syllables in Mandarin, disregarding their lexical tone, and all of them have a similar structure (C)V(N). Usually these syllables can be classified into small groups according to the manner of initial consonants and the nature of finals (similar to the 21 groups in Table 1). Syllables with aspirated affricative initials and nasal finals have the longest syllable duration, while syllables with simple vowel finals and zero initials have the shortest duration. Hence, there is considerable variation in duration between the different syllable structures. Due to the fact that this database is not large enough, and that using a smaller number of groups is more practical in duration modelling for speech synthesisers, the number of groups have been reduced to 12 in Table 2 according to their durational similarities.

Table 2. 12 Groups of syllables and mean duration.
Fig 1. (left) Mean duration of each syllable group

	ZI	VI & US	UA & FR	AS & AF
SVF	126	140	184	207
CVF	150	180	205	227
VNF	186	198	219	237



Tone

Syllable-based lexical tones in Mandarin are crucial for determining the prosodic features of Mandarin speech. Tone features are not only reflected in F0 but also in duration. Moreover, syllables with different tones have significantly different durations, identified as tone duration. The difference in syllable-based tone durations has been studied before using isolated monosyllables (Bai, 1934) or disyllables in a carrier sentence (Feng, 1985), and the relative lengths of tone duration are of particular interest to researchers, probably because it could be seen as one of the key features of tones. However, there have not been any consistent results on it to date, which is probably due to the context-dependent characteristics of duration. In this database, tone durations are calculated by taking the mean duration of syllables with the same lexical tone, assuming that other factors have the same effects on syllables in each tone group. Table 3 shows each tone duration (where H and F refer to 1st and 4th tones respectively; R includes the 2nd tone and 3rd tone sandhi; L contains 3rd tone and half 3rd tone). The differences in duration between most of the two tone groups are highly significant, such as H versus R ($t = 3.85$, $p < 0.001$) and L versus F ($t = 8.419$, $p < 0.0001$), except for the difference between R and F ($t = 2.27$, $p < 0.05$). The relative length order is very different to either Bai's or Feng's results, which could lead to the following questions: Does tone duration exist? If so, is it speaker dependent, context dependent, or neither? In his study, Feng found that relative length was dependent on the syllable position in sentences. For example, the 2nd tone is the longest in the middle of the sentence, with other tone durations being virtually the same, while in sentence-final position, the 3rd tone is the longest, with the 4th tone being the shortest and the 1st tone being slightly longer than the 2nd tone. For testing the potential interaction between tone and syllable position in sentences, ANOVA was performed on each tone groups from two sentential positions: (1) in sentence start/medial prosodic words; (2) in sentence final prosodic words. No evidence of interaction was shown, which suggests that relative tone duration is independent to position, while

both tone and position have significant effects on syllable duration ($F_{tone}(3) = 43.1$, $p < 0.01$; $F_{position}(1) = 1\,67$, $p < 0.01$). Figure 2 gives the duration of different tones at two different sentential positions, with virtual parallel lines indicating no interaction between the two tone factors and sentential position.

Table 3. Mean of tone duration, standard deviation, and numbers of observations in this study as well as mean duration in Feng' s (only tone duration at sentential final position included) and Bai's research, cited from Feng's paper.

		H (1st)	R (2nd)	L (3rd)	F (4th)
Wang	Mean	212	201	178	195
	Sd	45	52	44	48
	N	695	575	592	828
Feng	Mean	274	320	335	268
Bai	Mean	436	455	483	425

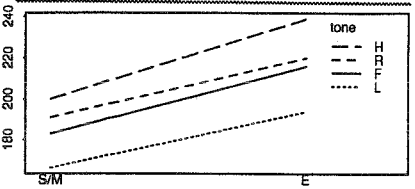


Fig 2. ANOVA analysis on the interaction of tone and position factors (S/M refers syllables in prosodic words at the start and medial position of sentences, END refers syllables in sentential-final prosodic words)

Unstressed syllables

Unstressed syllables which are labelled as N at tone level since unstressed syllables always lose their lexical tones and become neutral tone syllables, are frequently used in spoken Mandarin. Many of these are from lexically unstressed syllables but in continuous speech there are also quite a number which come from unstressed monosyllabic function words, such as 'de', 'le', and 'jiu'. In this database, about 15% of syllables (478 syllables) are unstressed. Considerable research has been done on unstressed syllables which shows that unstressed syllables have a much less syllable duration (25% ~ 50%) than stressed syllables. In this database, the durations of unstressed syllables (mean duration is 136 ms with a standard deviation of 40 ms) are much shorter than those of stressed syllables (mean duration of stressed syllables of all four tones is 196 ms with a standard deviation of 49 ms; $t = 25.48$, $p < 0.0001$). After analysing the structures of unstressed syllables, it was found that most of them have syllable structures containing non-aspirated stops, or voiced initials and simple vowel finals (otherwise most complex vowels would be reduced to simple vowels in unstressed syllables), so the unstressed factor may be confounded by the factor of syllable-structure in comparison with all syllables. To give an accurate comparison between stressed and unstressed syllables, the similar syllable structures VI & US and SVF were chosen and the results again showed that stressed syllables of all tones with a mean duration of 163 ms are much longer than unstressed syllables with a mean duration of 115 ms ($t = 17.03$, $p < 0.0001$).

Stress patterns and syllable position in prosodic words

Apart from the effect of being either stressed or unstressed, syllable duration is also strongly affected by the stress status of subsequent syllables. The earlier studies on pairs of disyllabic words with SS and SU patterns (S refers to stressed syllables and U to unstressed syllables) showed that the first stressed syllables in SS patterns are longer than the first syllables in SU patterns. Since over 93% of prosodic words in this database are either disyllabic or trisyllabic, the following analysis only concerns disyllabic and trisyllabic prosodic words with six stress patterns: SS, SU, SSS, SUS, SSU, SUU, being investigated. The duration data indicated that stressed syllables followed by unstressed syllables are significantly longer than those followed by stressed syllables. For example, duration of first stressed syllables in SU with 201 ms mean duration are longer than those in SS with 193 ms mean duration ($t = -2.26$, $p < 0.05$); the first S in SUS with 196 ms mean duration are much longer than the first S in SSS, of 179 ms mean duration ($t = -2.07$, $p < 0.05$); the second S in SSU with 206 ms mean duration is much longer than that in SSS with 190 ms mean duration ($t = -3.10$, $p < 0.01$). This difference has no relation to the syllable position of prosodic words, with syllable counts not showing a noticeable influence on syllable duration in this database.

Syllable position in prosodic words and sentences

That sentential position has a strong effect on syllable duration is shown in Fig 2. indicating that there is a sentence-final lengthening effect. By examining stressed syllables in prosodic words with SS and SSS patterns, it was found that sentence-final lengthening affects not only the last syllables of

sentences but also other syllables in sentence-final prosodic words, despite the fact that the degree of lengthening is different. Table 4. gives the mean duration of these syllables in different positions in prosodic words, and in different sentence positions (I, II, and III, refer start, medial and last position in prosodic words, respectively; S, M and E indicate prosodic words in start, middle and final position of sentences, respectively). The *t-test* was used to test the effects of two factors: syllable position in prosodic words; and word position in sentences. It is shown that syllable duration at two positions in SS disyllables have no significant difference at the start and medial sentence positions, but in sentence final position, the second syllable is significantly longer than the first syllables ($t = -5.37$, $p < 0.0001$). For syllables in all word positions, duration varies significantly from the start, medial and end positions of sentences in an order of long, short and longest, and most of the differences between any two sentential positions are significant ($p < 0.05$), especially those between the start, medial and final positions ($p < 0.001$). Similar shortening or lengthening can also be found in SSS prosodic words. Rd (Relative duration) in Table 4 gives the relative duration normalised by the duration in sentence-medial positions. It clearly shows that the final syllables of each of the prosodic words have a greater degree of lengthening than syllables in other prosodic word positions. Further ANOVA analysis showed there is no interaction between word position and sentence position.

Table 4. Mean duration, standard deviation and relative duration of syllables in different prosodic word and sentence positions.

		SS_1	SS_2	SSS_1	SSS_2	SSS_3
	Mean	191	188	175	189	206
	Sd	52	40	47	33	41
	Rd	1.04	1.04	1.01	1.08	1.06
M	Mean	184	180	172	175	194
	Sd	44	38	35	45	40
	Rd	1.00	1.00	1.00	1.00	1.00
E	Mean	210	242	190	202	243
	Sd	43	53	40	40	49
	Rd	1.19	1.34	1.11	1.15	1.25

Other potential factors

In previous studies, two factors were found to affect syllable duration: the number of syllables in words; and the morpho-syntactic structures of polysyllables (Ren, 1986; Wang, 1994). However, both factors in this database were found to have no significant effect on syllable duration.. For example, mean syllable duration is 185 ms, in disyllabic prosodic words, with 44 ms standard deviation; 185 ms, in trisyllabic words, with 44 ms standard deviation; and 188 ms, in quadrisyllabic words, with 43 ms standard deviation. There was no significant difference between them, which means syllable duration in continuous speech does not shorten as the number of syllables in words is increased. Furthermore, the morpho-syntactic structures of polysyllabic words did not seem to have a strong effect on syllable duration, although some subtle differences between structures could still be found (for example, the last syllables in 2 + 1 are longer than those in 1 + 1 + 1 and 1 + 2). These results imply that the influence of certain word-level factors is subordinate to the dominate factors reflecting the rhythmic features of continuous speech.

A TENTATIVE SYLLABLE-BASED DURATION MODEL

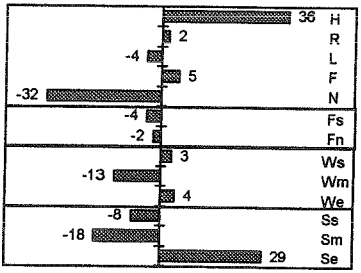
Modelling duration data for speech synthesisers has been developed in many languages and many modelling methods, such as decision tree models, regressive tree models, linear regressive models, additive models as well as three-layer neural network, were employed. In this study, a simple, additive model is used, with reference to the duration model for Japanese (Kaiki, al. 1992; Takeda, 1989). The following duration control factors are included in this model: (1) Syllable-based: syllable structure; syllable tone; and stressed/unstressed nature; (2) Prosodic word-based: stress statue of subsequent syllables within the same prosodic words; syllable position in words; (3) Utterance-based: sentential position. Syllable duration is estimated using the following formula, in which mean duration has to be related to each type of syllable structure.

> Estimated syllable-duration = mean duration (related to different syllable structure) + offsets for tone (including neutral tones of unstressed syllables), stress status of subsequent syllable, position of syllable in prosodic word, position of prosodic word in sentence.

It needs to be pointed out that, at present, this database is not large enough, nor is it factor-balanced, so the modelling here is mainly used to investigate dominant duration-control factors and provide insights on modelling duration for Mandarin synthesisers. From Table 2 and Fig. 3, we can see that

syllable structure has the greatest influence on syllable duration with the largest deviation range in duration (from -60 ms to 50 ms), with syllable tone (including neutral tone) being the second most dominant factor, and sentence position the third.

Fig 3 (right) gives the duration of syllables with UA & FR initials and SVF finals (with a mean duration of 184 ms) which varies under the influence of different factors. For example, if an 'shi' with forth tone, followed by a stressed syllable, in the medial of prosodic word 'boshisheng' and in the end position of sentence, the duration of 'shi' could be predicted as follows: 184 + 5 – 4 – 13 + 29 = 201 ms



CONCLUSIONS

Like other languages, Mandarin as a tonal language has its language-specific timing features which are related to the following prosodic and rhythmic characteristics: (1) various syllable structures of (C)V(N), with strong compensation between the acoustic duration of initial consonants and finals; (2) syllable-based tones with different tone-related durations; (3) a significant number of short unstressed (neutral tone) syllables, some of which are from monosyllabic functional words in spoken Mandarin; (4) the dominant role of disyllabic and trisyllabic prosodic words as the basic rhythmic units in forming Mandarin rhythm; and (5) durational lengthening in the sentence-final position, which is probably due to the fact that the majority of the information focus in neutral sentences in Mandarin occurs in the sentence final position.

The duration data in this study may be of help in understanding the rhythmic features of Mandarin, and providing an overall insight into duration modelling. However, further research using an expanded database and labelling rhythmic units which are larger than prosodic words and smaller than utterances, as well as considering other factors, such as speech rate and pauses, should be carried out.

REFERENCES

Cheng, M. Y. (1990) *What must phonology know about syntax*? The Phonology-Syntax Connection, ed by Inkelao, S.and Zec, D. (The University of Chicago Press), 19-46.

Feng, L. (1985) *Duration of initials, finals and tones in Beijing dialects (in Chinese)*, Working Papers in Experimental Phonetics (Peking University Press), 131-195.

Fowler, C. A. and Saltzman, E. (1993) *Coordination and coarticulation in speech production*, Language and Speech, 36 (2, 3), 171-195.

Harrington, J., Cassidy, S., Fletcher, J. and McVeigh, A. (1993) *The mu+ system corpus based speech research*, Computer Speech and Language 7, 305-331.

Kaiki, N., Takeda, K. and Sagisaka, Y. (1992) *Linguistic properties in the control of segmental duration for speech synthesis*, in Talking Machines: Theories, Models, and Design, ed by G. Bailey, C. Benoit, and T. R. Sawallis (North Holland, London), 255-263.

Ren. H. M. (1986) *Linguistically Conditional Duration Rules in a Timing Model for Chinese*, UCLA Working Papers in Phonetics , 34-49.

Shen, X. S. (1993*) Relative duration as a perceptual cue to stress in Mandarin*, Language and Speech, 36(4), 415-433.

Shih, Ch.-L. (1986) *The prosodic domain of tone sandhi in Chinese*, PhD dissertation, University of California San Diego.

Takeda, K, Sagisaka, Y. and Kuwabara, H. (1989) *On sentence-level factors governing segmental duration in Japanese*, J. Acoust. Soc. Am. 86(6), 2081-2087.

Wang, J. (1994) *Morpho-syntactic structures and syllable duration in polysyllabic words in Mandarin*, (in Chinese), Zhongguo Yuwen, (in press).