Effects of Speaking Rate on Initial and Final Duration Structure in Mandarin Chinese

Wen-Hsing Lai
Department of Computer and Communication Engineering
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
e-mail: lwh@nkfust.edu.tw

***Abstract*—An Expectation-Maximization (EM) modeling and a speech corpus with fast, median, and slow speaking rate are applied to explore the effect of speaking rate on segmental duration structure of *Initial*, *Final*, and syllable in Mandarin Chinese. Experimental results showed that the variance of duration was greatly reduced after eliminating effects from additive factors by EM algorithm. By excluding the interference of acoustical factors, the relationship between syllable duration and the structure of *Initial* and *Final* durations for different speaking rate is observed. The result shows that for same syllable duration, the ratio of *Final* to *Initial* becomes larger when the speaking rate becomes faster. Besides, the ratio, generally, becomes larger as the syllable becomes longer. However, for extremely short syllable about less than 100 ms in fast speed, the ratio becomes large, and in syllable duration longer than about 350 ms in median and slow speed, the ratio becomes almost a constant.**

Keywords- speaking rate; duration; Mandarin Chinese

I. INTRODUCTION

Speaking rate is one of the most important factors in spontaneous speech systems, because variability in speaking rate may be often observed in spontaneous speech than in read speech. Studies have shown that the acoustic properties corresponding to phonetic segments of speech are influenced by variability of speaking rate. For example, spectral patterns will be changed and formant positions will be shifted [1]-[4] and the intelligibility and comprehension will be influenced [5]-[8]. Furthermore, changes in speech rate have effects on prosody, like the overall level and range of fundamental frequency (*F*0) and durations. While changing speaking rates, it also causes variations in prosodic phrasing, such as prosodic boundaries. For example, the research of Tseng find that duration adjustment is made systematically at each prosody level during speech production and examining speech rate in relation to prosody units is a significant first step to understanding temporal organization of speech flow [9].

Because of its importance, speaking rate has been put into considerations in many speech application, for instance, speech recognition [10]-[16], emotion classification [17], and Text-To-Speech system [18]. Previous studies revealed that the mismatch between speaking rates of training and test data of speech recognition system will degrade the system performance, therefore, some researches focus on solving the problem of performance degradation caused by speaking rate variability [10]-[16]. Further, different emotional dispositions of a person are strongly expressed in his/her speaking rate [17], therefore, speaking rate also has significant influence in emotion classification. Hence, speaking rate estimation [19] has become an important job. In addition, speaking rate-controlled prosody is also critical for Text-To-Speech system [18].

Though there are some studies about the effect of speaking rate for vowel duration [20] and prosody units [9], there are few studies about the initial and final structure change under various speaking rate. Therefore, our objective of this paper is to find out the change of initial and final structure under various speaking rate to further understand the temporal organization of speech flow, so it could be applied to speech related application.

However, it is difficult to get an obvious pattern from observing the original duration directly or to incorporate qualitative findings into a quantitative model, and there has been rather few prosodic model devoted to investigating detailed effects of speech rate modification on the realization of individual pitch accents, duration, intonation, and prosodic structures. Hence, an Expectation-Maximization (EM) modeling [21] and a speech corpus with fast, median, and slow speaking rate are applied to explore the effect of speaking rate on segmental duration structure in Mandarin Chinese in this article. The achievement will be useful for improving the quality of speech synthesis and the recognition rate of speech recognition.

The paper is organized as follows. In Section II, the EM analysis algorithm, including the factors which have impact on durations, the syllable duration modeling, and the extension to *Initial* and *Final* Duration Modeling, is shown. Section III describes the experimental results. Conclusions are given in the last section.

II. EM ANALYSIS ALGORITHM

A. Factors

In naturally spoken Mandarin Chinese, duration varies considerably depending on various linguistic and nonlinguistic factors [22]. Mandarin Chinese is a tonal and syllable-based language. Each character is pronounced as a

 ISBN: 978-1-61208-342-1

syllable, the basic pronunciation unit. There exist only about 1300 phonetically distinguishable syllables comprising all legal combinations of approximately 411 base-syllables and five tones. Mandarin base-syllables have very regular phonetic structure. Each base-syllable is of the form (C)(C)V(N), where C is a consonant, V is a vowel, and N is a nasal consonant (the symbols between parentheses signal optionality). So, base-syllables comprise one to four phonemes. Generally speaking, syllable duration increases as the number of constituent phonemes increases. Syllables with single vowels are shortest. Syllables with stop initials or no initials, and without nasal endings are pronounced shorter. Syllables with fricative initials and with nasal endings are longer. Therefore, duration is seriously influenced by the phonetic structure of base-syllables, and base-syllable is listed as one of the impact factors.

Tonality of a syllable is characterized by its pitch contour shape, loudness and duration. For example, syllables with Tone 5 (Neutral Tone) are always pronounced much shorter. Therefore, tonality is also considered as a factor which has impact on duration.

To prevent the speed variation in different utterance sentence in recording process for specific speed rate, utterance is also included as a factor for normalization.

Aside from the acoustic factors mentioned above, including lexical tone, base-syllable, and utterance, other high-level linguistic components, such as word-level and syntactic-level factors, like the boundary index, position in word, length of word factors used in [22], also seriously influence the duration of an utterance. In the model, the prosodic state, as a substitute for high-level linguistic information, is used to indicate the state in prosodic word or prosodic phrases, for example, to indicate the possible prosodic word or phrase boundaries, or the notions of prominence. Therefore, the prosodic state is used to account for the influence of all high-level linguistic features. There are two advantages of using the prosodic state to replace high-level linguistic features. Firstly, duration information is a prosodic feature, so the variation of the duration should better match the prosodic phrase structure than the syntactic phrase structure. Secondly, as mentioned above, some unsolved problems, such as the ambiguity of word-segmentation and word-chunking in Mandarin Chinese and the difficulty of performing automatic syntactic analysis on unlimited natural texts, can be avoided in the current duration modeling approach. This prevents us from using improper or incomplete high-level linguistic information. By doing so, the modeling of duration can simply consider the effects of prosodic state and acoustical factors, like tone, utterance and base-syllable factors. Due to the fact that the prosodic state is not explicitly given, it has been treated as a hidden variable in the EM algorithm. The number of prosodic states is set as 16 in our modeling. A by-product of the EM algorithm is the determination of the hidden prosodic states of all the units in the training set. This is an additional advantage. From the sequence of prosodic states, some high-level linguistic phenomenon could be observed, like the possible prosodic phrase boundaries.

In sum, four major affecting factors including tone, base-syllable, utterance, and prosodic state are considered.

B. Syllable Duration Modeling

By considering the factors in Section II-A, an additive duration model can be expressed by

$$Z_n = X_n + \gamma_{t_n} + \gamma_{y_n} + \gamma_{j_n} + \gamma_{l_n}, \quad (1)$$

where Z_n and X_n are, respectively, the observed duration and the normalized (residual) duration of the *n*th syllable. X_n is considered as the residual duration after excluding all the impact from factors and is modeled as a normal distribution with mean μ and variance ν. γ_{t_n}, γ_{y_n}, γ_{j_n}, and $\gamma_{\ln}$ are the impact value of the lexical tone, prosodic state, base-syllable, and utterance identification number factor of the *n*th syllable, indicated by t_n, y_n, j_n and l_n.

To illustrate the EM algorithm, an auxiliary function is defined in the expectation step as

$$Q(\bar{\lambda}, \lambda) = \sum_{n=1}^{N} \sum_{y_n=1}^{Y} p(y_n \mid Z_n, \bar{\lambda}) \log p(Z_n, y_n \mid \lambda), \quad (2)$$

where N is the total number of training samples; Y is the total number of prosodic states; $\lambda = \{\mu, v, \gamma_t, \gamma_y, \gamma_j, \gamma_l\}$ is the set of parameters to be estimated, and λ and $\bar{\lambda}$ are, respectively, the updated and old parameter sets. γ_t, γ_y, γ_j, and γ_l represent the impact value of all the lexical tone, prosodic state, base-syllable, and utterance identification number factor. For example, the possible t_n, the lexical tone of the *n*th syllable, is 1 to 5, therefore, γ_t represent γ_1, γ_2, γ_3, γ_4, and γ_5.

$$p(Z_n, y_n \mid \lambda) = N(Z_n; \mu + \sum_{p} \gamma_{p_n}, \nu), \quad (3)$$

where $N(Z; \mu, v)$ is a normal distribution of Z with mean μ and variance ν. $p(y_n \mid Z_n, \bar{\lambda})$ can be represented as

$$p(y_n \mid Z_n, \bar{\lambda}) = \frac{p(Z_n, y_n \mid \bar{\lambda})}{\sum_{y'_n=1}^{Y} p(Z_n, y'_n \mid \bar{\lambda})}. \quad (4)$$

To cure the drawback of the non-uniqueness of the solution because of the use of additive factors, the optimization procedure in the Maximization step (M-step) is modified to a constrained optimization via introducing a global duration constraint. The auxiliary function then changes to

$$Q(\bar{\lambda},\lambda)=\sum_{n=1}^{N}\sum_{y_n=1}^{Y}p(y_n\mid Z_n,\bar{\lambda})\log p(Z_n,y_n\mid\lambda) + \eta(\sum_{n=1}^{N}(\mu+\sum_{p}\gamma_{p_n})-N\mu_z) \quad , \qquad (5)$$

where μ_z is the average of Z_n and η is a Lagrange multiplier [23]. The constrained optimization is finally solved by the Newton-Raphson method [23].

Initializations of the parameters in λ are done by estimating each parameter independently. Then, iterative sequential optimizations of the parameters in λ are performed in the M-step. Iterations are continued until a convergence is reached. The prosodic state can finally be assigned by

$$y_n^*=\arg\max_{y_n} p(y_n\mid Z_n,\lambda)\,. \qquad (6)$$

C. Extension to Initial and Final Duration Modeling

Each Mandarin syllable is composed of an optional consonant *Initial* and a *Final*. The *Final* comprises an optional medial, a vowel nucleus and an optional nasal ending. To exploit the relationship between the syllable duration and its component *Initial* and *Final* durations in different speaking rate, the above syllable duration modeling is extended to the duration modeling of *Initial* and *Final*. There are two approaches. One approach is to keep the prosodic states of the three models independent because the optimal prosodic states of both *Initial* and *Final* duration models may not match with those of the syllable duration model. The mismatch may results from the inconsistency in the effect of linguistic features on the *Initial* duration and on the *Final* duration. A previous study [24] found that consonant-lengthening can happen at all initial positions especially at the beginning of a word, while vowel-lengthening can occur only at phrasal final. The other approach is to share the same prosodic states so the relationship between the impact value of prosodic states of syllable and those of *Final* and *Initial* can be observed more conveniently. For the first approach, *Initial* and *Final* models could adopt the similar method as syllable. For the sharing model, an additional constraint is set in *Initial* and *Final* models to let their prosodic states the same as in syllable model. The training algorithm of *Initial* and *Final* models is then modified to an ML (Maximum Likelihood) one with all prosodic states being predetermined by the training procedure of syllable model.

III. Experimental Results

The corpus is recorded in fast, median, and slow speed by a professional female announcer in reading style using WaveSurfer software on personal computer. The median speed was recorded first. The material contains 359 short paragraphs including news, blogs and text books of elementary school. There are totally 44934 syllables. Averagely, every sentence contains 10.37 syllables. The sentence length ranges from 80 to 272 syllables. The sampling rate is 20 kHz and the file format is 16 bit PCM. The pronunciations have been labeled. The boundaries of syllable, *Initial* and *Final* have also been marked by automatic segmentation based on Hidden Markov Model ToolKit (HTK) [25], and then corrected manually.

Table 1 shows the duration mean, standard deviation and the ratio of standard deviation to mean of syllable, *Initial* and *Final* in fast, median, and slow speed. The experiment of *Initial* was done without considering the null *Initial* and the very short *Initial*s of {b, d, g} which are generally difficult to be segmented accurately. As shown in Table 1(a) and (b), the duration mean of syllable, *Initial* and *Final* lengthen and the standard deviation become larger as the speed slows down. Besides, from Table 1(c), the ratio of standard deviation to mean of fast speed is the largest. That is, the relative variation is the greatest in high speed.

The normal distribution assumption is then checked. Take syllable durations in slow speed as an example. Fig. 1(a) shows the histogram with a fitted normal distribution and Fig. 1(b) shows a normal Q-Q plot with an RMA (Reduced Major Axis) regression line, together with the Probability Plot Correlation Coefficient (PPCC) equals 0.9967. (Basically, a normal distribution will plot on a straight line.) Besides, Shapiro-Wilk normality test returns a test statistic W = 0.9934, ($0 < W \leqq 1$, W is small for non-normal samples). Jarque-Bera normality test returns *JB* = 870.1 and chi-square normality test returns value 200.01. The p values of the three tests are much smaller than 0.05. From the above observations, except some outliers (those samples with much longer and shorter durations), most samples actually fit the normal distribution. To make the model simple, the assumption of Gaussian distribution is still adopted in this study. A mixture Gaussian distribution may fit better and could be put as a future study.

Table 2 shows the mean, standard deviation, and RMSE (Root Mean Square Error) of the normalized duration of syllable, *Final*, and *Initial* in fast, median, and slow speed in EM modeling. The *Final* and *Initial* models used in this experiment are not sharing prosodic states with syllable model. After excluding the impact of factors by EM modeling, the standard deviation of the normalized duration in Table 2(b) greatly reduced compared with the original standard deviation in Table 1(b), while the mean of the normalized duration in Table 2(a) is almost the same with the mean in Table 1(a). Therefore, the EM modeling can successfully exclude the impact of factors. The RMSE of prediction duration by additive model is shown in Table 2(c). From Table 1(b), 2(b), and 2(c), though the original syllable standard deviation of high speed speaking rate is higher than the deviation of *Final*, the normalized syllable standard deviation of high speed speaking rate becomes lower than the deviation of *Final*, and in prediction stage, the RMSE of syllable prediction is lower than the RMSE of *Final*. The relatively high deviation of the normalized *Final* duration and RMSE show that the *Final* duration in fast speed is more difficult to model than syllable duration.

At last, the relationship between syllable duration and the structure of *Initial* and *Final* durations after excluding the

TABLE I. THE DURATION (A) MEAN (UNIT: MS), (B) STANDARD DEVIATION (UNIT: MS) AND (C) RATIO OF STANDARD DEVIATION/MEAN OF SYLLABLE, *INITIAL* AND *FINAL* IN FAST, MEDIAN, AND SLOW SPEED.

(a)

	Fast	Median	Slow
Syllable	185.253	245.541	271.494
Final	135.001	176.731	195.280
Initial	73.265	97.643	107.517

(b)

	Fast	Median	Slow
Syllable	67.683	77.259	83.390
Final	62.291	67.527	75.388
Initial	37.679	44.210	48.407

(c)

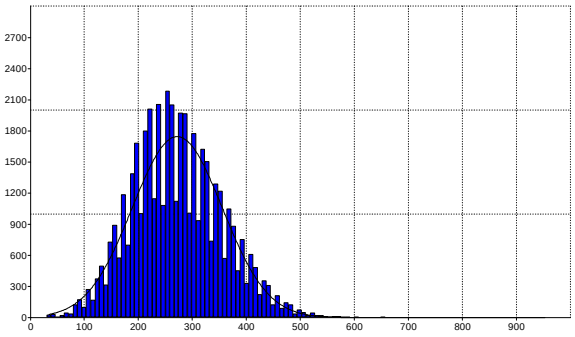
	Fast	Median	Slow
Syllable	0.365	0.315	0.307
Final	0.461	0.382	0.386
Initial	0.514	0.453	0.450

interference from acoustical factors for different speaking rate is examined. The impact value of prosodic states plus the mean of duration is taken as the duration excluding the impact from acoustic factors. Specifically, the ratio of $(\gamma_y^f + \mu_f)/(\gamma_y^i + \mu_i)$ versus $(\gamma_y + \mu)$ is observed. γ_y^f, γ_y^i, and γ_y are the impact value of prosodic states of *Final*, *Initial* and syllable duration models. μ_f, μ_i, and μ are the mean of *Final*, *Initial* and syllable durations. $(\gamma_y^f + \mu^f)/(\gamma_y^i + \mu^i)$ can be considered as the duration component ratio of *Final* to *Initial* in syllable structure. For easy comparing, the *Final* and *Initial* models used in this experiment are sharing prosodic states with syllable model.

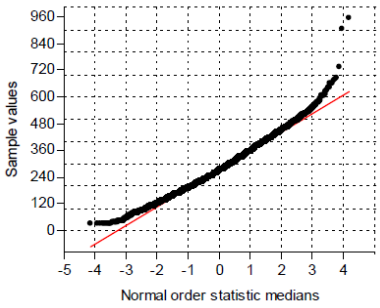
Fig. 2 displays the figure of $(\gamma_y^f + \mu_f)/(\gamma_y^i + \mu_i)$ versus $(\gamma_y + \mu)$, or the ratio of *Final* to *Initial* versus the syllable duration after excluding the interference from acoustical factors. The vertical axis is $(\gamma_y^f + \mu_f)/(\gamma_y^i + \mu_i)$ and the horizontal axis is $(\gamma_y + \mu)$. From Fig. 2, it is easy to see that for the same syllable duration, the duration ratio of *Final* to *Initial* of fast speaking rate is highest. It is followed by the ratio of median rate, and the ratio of slow rate is lowest. That is, the ratio of *Final* to *Initial,* generally, becomes larger as the speaking rate increases. It may be because in fast speed, the pronunciation is more relaxed and *Final* dominates. Besides, generally, the ratio becomes larger as the syllable becomes longer. But, for extremely short syllable in fast speed, about less than 100 ms, the ratio becomes large. Besides, in syllable duration larger than about 350 ms in median and slow speed, the value of the ratio gets almost saturated and becomes almost a constant.

Our objective of this modeling is to find out the change of initial and final structure under various speaking rate. However, it is difficult to get an obvious pattern from the original observed duration. Observing the results of our experiment, the value of acoustic factors including lexical tone, base-syllable, and utterance, did not show particular different pattern among different speaking rate. Therefore, we assume that speaking rate does not have big impact on the acoustic factors including lexical tone, base-syllable, and utterance in our experiment, and we observed the change of initial and final structure under various speaking rate as in Fig.2. After excluding the interference of acoustic factors, it is easy to find out that for the same syllable duration, when we increase the speaking rate, the duration ratio of *Final* to *Initial* becomes larger.

At last, speaker factor may be also important for speaking rate. Since our experiment is based on a corpus recorded by a single professional speaker, the impact of speaker is not included in our modeling, therefore, the further study of speaker factor will be put as our future work.



(a)



(b)

Figure 1. The (a) histogram (unit of horizontal axis: ms) (b) normal Q-Q plot with an RMA regression line (unit of vertical axis: ms) of the observed syllable durations in slow speed.

TABLE II. THE (A) MEAN, (B) STANDARD DEVIATION, AND (C) RMSE OF THE NORMALIZED DURATION OF SYLLABLE, *FINAL*, AND *INITIAL* IN FAST, MEDIAN, AND SLOW SPEED (UNIT: MS).

(a)

	Fast	Median	Slow
Syllable	183.291	242.985	268.032
Final	133.847	175.115	192.651
Initial	73.539	97.956	107.382

(b)

	Fast	Median	Slow
Syllable	8.928	11.602	11.596
Final	11.660	10.544	11.447
Initial	5.204	6.489	7.051

(c)

	Fast	Median	Slow
Syllable	8.933	11.608	11.600
Final	11.677	10.569	11.450
Initial	5.233	6.520	7.087

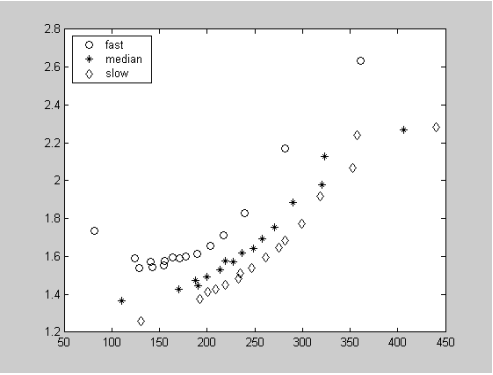


Figure 2. The ratio of *Final* to *Initial* versus syllable duration (unit: ms) after excluding acoustical factors.

IV. CONCLUSIONS

In this paper, the duration variation was studied and duration models are built for syllable, *Initial* and *Final* in different speaking rate for Mandarin Chinese. An EM algorithm is applied to syllable duration modeling. Extensions of the syllable duration modeling method are also performed on *Initial* and *Final*. From the experimental results, the impact of factors on syllable, *Initial* and *Final* duration in different speaking rate are explored. By observing the relationship between syllable duration and the structure of *Initial* and *Final* durations after excluding the interference from acoustical factors for different speaking rate, an important conclusion is that for the same syllable duration, the duration ratio of *Final* to *Initial* becomes larger as the speaking rate increases. In addition, the ratio basically becomes larger as the syllable becomes longer. But for extremely short syllable in fast speed, the ratio becomes large; in syllable duration larger than about 350 ms in median and slow speed, the ratio becomes almost saturated.

ACKNOWLEDGMENT

This work was supported by NSC, Taiwan under Contract NSC 99-2221-E-327-044. The author thanks Dept. of Electrical and Computer Engineering, National Chiao Tung University, Taiwan for supplying the speech databases.

REFERENCES

[1] M. Pitermann, “Effect of Speaking Rate and Contrastive Stress on Formant Dynamics and Vowel Perception,” The Journal of the Acoustical Society of America, vol. 107, no. 6, Jul. 2000, pp. 3425-3437, doi:10.1121/1.429413.

[2] T. Gay, “Effect of Speaking Rate on Vowel Formant Movements,” The Journal of the Acoustical Society of America, vol. 63, no. 1, Feb. 1978, pp. 223-230, doi:10.1121/1.381717.

[3] G. Weismer and J. Berry, “Effects of Speaking Rate on Second Formant Trajectories of Selected Vocalic Nuclei,” The Journal of the Acoustical Society of America, vol. 113, no. 6, Jul. 2003, pp. 3362-3378, doi:10.1121/1.1572142.

[4] D. O’Shaughnessy, “The Effects of Speaking Rate on Formant Transitions in French Synthesis-by-Rule,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 86), vol. 11, Apr. 1986, pp. 2027-2030, doi: 10.1109/ICASSP.1986.1168797.

[5] A. H. S. Chan and P. S. K. Lee, “Intelligibility and Preferred Rate of Chinese Speaking,” International Journal of Industrial Ergonomics, vol. 35, no. 3, Jan. 2005, pp. 217-228, doi:10.1016/j.ergon.2004.09.001.

[6] J. C. Krause and L. D. Braida, “Investigating Alternative Forms of Clear Speech: The Effects of Speaking Rate and Speaking Mode on Intelligibility,” The Journal of the Acoustical Society of America, vol. 112, no. 5, pt. 1, Nov. 2002, pp. 2165-2172, doi: 10.1121/1.1509432.

[7] C. Jones, L. Berry, and C. Stevens, “Synthesized Speech Intelligibility and Persuasion: Speech Rate and Non-native Listeners,” Computer Speech and Language, vol. 21, 2007, pp. 641-651, doi:10.1016/j.csl.2007.03.001.

[8] S. Liu and F. G. Zeng, “Temporal Properties in Clear Speech Perception,” The Journal of the Acoustical Society of America, vol. 120, no. 1, Jul. 2006, pp. 424-432, doi: 10.1121/1.2208427.

[9] C. Y. Tseng and Y. L. Lee, “Speech rate and prosody units: Evidence of interaction from Mandarin Chinese,” Proceedings of the International Conference on Speech Prosody, Mar. 2004, pp. 251-254.

[10] E. Fosler-Lussier and N. Morgan, “Effects of Speaking Rate and Word Frequency on Pronunciations in Convertional Speech,” Speech Communication, vol. 29, no. 2-4, Jan. 1999, pp. 137-158, doi:10.1016/S0167-6393(99)00035-7.

[11] T. Shinozaki and S. Furui, “Hidden Mode HMM using Bayesian Network for Modeling Speaking Rate Fluctuation,”

 ISBN: 978-1-61208-342-1

Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU03), Jan. 2003, pp. 417-422, doi:10.1109/ASRU.2003.1318477.

[12] H. Nanjo and T. Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," IEEE Transactions on Speech and Audio Processing, vol. 12, issue 4, Jul. 2004, pp. 391-400, doi: 10.1109/TSA.2004.828641.

[13] M. S. Sommers, L. C. Nygaarda, and D. B. Pisoni, "Stimulus Variability and Spoken Word Recognition. I. Effects of Variability in Speaking Rate and Overall Amplitude," The Journal of the Acoustical Society of America, vol. 96, no. 3, Sep. 1994, pp. 1314-1324, doi: 10.1121/1.411453.

[14] M. S. Sommers and J. Barcroft, "Stimulus Variability and the Phonetic Relevance Hypothesis: Effects of Variability in Speaking Style, Fundamental Frequency and Speaking Rate on Spoken Word Identification," The Journal of the Acoustical Society of America, vol. 119, no. 4, Apr. 2006, pp. 2406-2416, doi: 10.1121/1.2171836.

[15] M. Radeau, J. Morais, P. Mousty, and P. Bertelson, "The Effect of Speaking Rate on the Role of the Uniqueness Point in Spoken Word Recognition," Journal of Memory and Language, vol. 42, 2000, pp. 406-422, doi:10.1006/jmla.1999.2682.

[16] S. M. Ban and H. S. Kim, "Speaking Rate Dependent Multiple Acoustic Models Using Continuous Frame Rate Normalization," Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Dec. 2012, pp. 1-4.

[17] D. Philippou-Hubner, B. Vlasenko, R. Bock, and A. Wendemuth, "The Performance of the Speaking Rate Parameter in Emotion Recognition from Speech," IEEE International Conference on Multimedia and Expo Workshops (ICMEW), July 2012, pp. 296-301, doi: 10.1109/ICMEW.2012.57.

[18] C. H. Hsieh, Y. R. Wang, C. Y. Chiang, and S. H. Chen, "A Speaking rate-Controlled Mandarin TTS System," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2013, pp. 6900-6904, doi: 10.1109/ICASSP.2013.6638999.

[19] Y. Wu, Q. H. He, and Y. X. Li, "Speaking Rate Estimation for Multi-Speakers," International Conference on Audio, Language and Image Processing (ICALIP), July 2012, pp. 976-979, doi: 10.1109/ICALIP.2012.6376756.

[20] M. Wang, W. Shi, R. Huang, and Z. Xiong, "The Temporal Effect of Speaking Rate, Focus and Prosody in Chinese," 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), Dec. 2012, pp. 445-449, doi: 10.1109/ISCSLP.2012.6423481.

[21] S. H. Chen, W. H. Lai, and Y. R. Wang, "A new duration modeling approach for Mandarin speech," IEEE Trans. on Speech and Audio Processing, vol.11, issue 4, July 2003, pp. 308-320, doi: 10.1109/TSA.2003.814377.

[22] M. Chu and Y. Feng, "Study on factors influencing durations of syllables in Mandarin," EUROSPEECH, Sep. 2001, pp. 927-930.

[23] P. Wriggers, Computational Contact Mechanics, 2nd ed., Spring, 2006, pp. 336-337.

[24] C. Shih and B. Ao, "Duration Study for the Bell Laboratories Mandarin Text-to-Speech System," in Progress in Speech Synthesis, J. P. H. van Santen, J. P. Olive, R. W. Sproat, and J. Hirschberg, Eds. Springer, 1997, pp. 383-399.

[25] S. J. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," University of Cambridge, Department of Engineering, 1994, http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=BD1F4D060CB93D921A658DB57F3C1839?doi=10.1.1.17.8190&rep=rep1&type=pdf.

 ISBN: 978-1-61208-342-1