Consistency Analysis of the Duration Parameter Within a Syllable for Mandarin Speech

Cheng-Yu Yeh[1,*], Kuan-Lin Chen[2], Shaw-Hwa Hwang[2]

[1] *Department of Electrical Engineering, National Chin-Yi University of Technology*
57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 41170, Taiwan, R.O.C.
e-mail: cy.yeh@ncut.edu.tw

[2] *Department of Electrical Engineering, National Taipei University of Technology*
1, Sec. 3, Chung-hsiao E. Rd., Taipei 10608, Taiwan, R.O.C.
e-mail: hsf@ntut.edu.twAdresas

crossref http://dx.doi.org/10.5755/j01.itc.42.2.1745

Abstract. This work presents a study of Mandarin speech focusing on consistency analysis of the duration parameter within syllables. Identified as a result of inspection of the human pronunciation process, this consistency can be interpreted as a high correlation between the warping curves of the spectrum and the prosody intra a syllable. Through three steps in the procedure of the consistency analysis, the HMM algorithm is used firstly to decode HMM-state sequences within a syllable at the same time as to divide them into three segments. Secondly, based on a designated syllable, the vector quantization (VQ) with the Linde-Buzo-Gray algorithm is employed to train the VQ codebooks of each segment. Thirdly, the duration vector of each segment is encoded as an index by VQ codebooks, and then the probability of each possible path is evaluated as a prerequisite to analyze the consistency. It is demonstrated experimentally that a consistency is definitely acquired in case the syllable is located exactly in the same word. These results offer a research direction that the time warping process intra a syllable must be considered in a TTS system to improve the synthesized speech quality.

Keywords: consistency analysis, hidden Markov model (HMM), vector quantization (VQ), text-to-speech (TTS), speech synthesis.

1. Introduction

A text-to-speech (TTS) system [1-5], also known as speech synthesis, is a system which converts text input to speech output, and which has a number of applications including intelligent human computer interfaces and auxiliary speech systems for the visually impaired. Moreover, due to the growing demand of embedded systems, a wide range of portable devices, such as smart phones, have increased dramatically in popularity, and thus the potential for extended TTS application developments looks promising. Consequently, the integration of TTS systems into embedded systems has become one of the leading research issues in recent years [6-9]. The growing significance of TTS lies in the fact that it can provide a variety of speech-based applications.

A review of the development of TTS techniques shows that the waveform-based synthesis units approach [10-17] is one of the most commonly used techniques in this field. This approach is further classified into two types in terms of the size of the synthesis units, namely corpus-based [9-12], and small footprint [13-17] units. The corpus-based speech synthesis approach relies on a unit selection method and modification of speech units from a large speech database, which is usually derived from a sufficiently large corpus where appropriately selected spoken utterances are carefully annotated to the unit level. The selection of the units aims to cover as many units as possible in different phonetic and prosodic contexts in order to provide the necessary variability in the synthetic speech output. However, this approach requires a great number of speech units and a large storage space to reach a superior speech quality.

In contrast, the footprint TTS approach adopts a small size synthesis unit, which treats a set of fundamental speech elements, e.g. phonemes, diphones or syllables, as synthesis units. Synthesized

[*] Corresponding author

speech is then made through a prosodic modification conducted on the synthesized units using pitch-synchronous overlap-add (PSOLA) algorithms [13, 14]. Accordingly, this approach affords the double advantage of requiring low memory and a low computation load with an inferior but comparable speech quality relative to the corpus-based methods. An additional advantage is that a footprint TTS is more suitable for implementation on embedded platforms and portable devices.

However, a TTS adopting the waveform-based synthesis units approach necessitates a permanent prosody model to deal with the prosodic modification of the synthesized units. Exploration of the human pronunciation process has indicated that speech is made by an excitation source flowing through the vocal tract and emanating from the mouth and nostrils of the speaker. The excitation source containing the airflow and the vibration of the vocal cords reflects the prosodic information. Both the vocal tract, affecting the voice spectrum, and the excitation source, are coupled together to generate natural and fluent speech. Thus, an inspection result is seen, which is that the prosody and the spectrum embedded in the running speech are consistent. One of the significant issues for TTS systems is that the spectrum and prosody modules are addressed separately in most cases, leading to an inconsistency between them. Therefore, we were motivated to demonstrate that consistency between the prosody and the spectrum embedded in running speech does in fact exist.

With our aim of verifying this property of consistency, and taking Mandarin speech as an example, a definition of consistency is first provided. Subsequently, the consistency analysis of the spectrum and duration parameter of prosody within specific syllables is discussed. The analytic methods, procedures, and practical experiments are presented to demonstrate the proposed deduction. It is expected that the findings of this research will contribute to improvements in the performance of Mandarin TTS systems.

The rest of this paper is outlined as follows. The modeling of the consistency analysis in Mandarin speech is described in Section 2. A procedure of the consistency analysis is presented in Section 3. The experimental results are demonstrated and discussed in Section 4, followed by the conclusions in the last section.

2. Modeling of the consistency analysis in Mandarin speech

As referred to previously, examination of the human pronunciation process has revealed that both the excitation source and the vocal tract couple to generate natural and fluent speech. The excitation source reflects the prosodic information, while the vocal tract affects the voice spectrum. The prosodic information usually consists of the pitch contour, duration, and energy parameters. In this work, the property of consistency between the duration and the spectrum is analyzed. A definition and modeling of the consistency analysis in Mandarin speech is therefore presented.

In the phonology of Mandarin Chinese, there are a total of 411 distinguishable syllables composed of an optional consonant *initial* and a vowel *final* as the basic pronunciation units. However, a Chinese word consisting of a minimum of one syllable is regarded as the smallest meaningful unit. Besides, the waveform and the spectrum of all the same pronunciation units are definitely not identical because the speech signal is non-stationary. Thus, consistency can be interpreted as a high correlation between the warping curves of the spectrum and the prosody intra a syllable. The warping curve means a curve that the prosodic information shifted along the spectrum within a syllable. For further explanation, the warping curves are consistent as long as the same pronunciations are located in the same word, implying that the same pronunciations located in different words bring about distinct consistencies, that is, different warping curves are made. Observing the warping curve can help us to further acquire an understanding of the subtle variation between the spectrum and the prosody intra a syllable.

Subsequently, the following analysis is made on a syllabic basis, according to which the warping curves of the spectrum and the duration intra a syllable are the focus of interest. The warping curve within a syllable can be obtained by exploring the duration information under a sequence of hidden Markov model (HMM)-state based spectral segments.

In the HMM-state based spectral segments, the Mel-frequency cepstral coefficients (MFCCs) are used as spectral feature and the HMMs are employed to decode the state sequence within a syllable [18]. On the other hand, for exploring the duration information within a spectral segment, each syllable is divided into three spectral segments, with each consisting of two to three HMM states. Based on the spectral segments, all the state durations are employed as a duration vector, and then a clustering algorithm is used to analyze the vector. Thus, the warping curve can be analyzed by exploring the clustering results of the duration vector within a spectral segment. In this work, the dimension is set to 12 for evaluating MFCCs under speech database with 8 kHz sampling frequency.

3. Procedure of consistency analysis

In Figure 1, a flowchart of the procedure of consistency analysis is presented. There are three steps required in the procedure. Firstly, the feature extraction such as MFCCs, duration parameter etc. are computed from a large speech database. Then, dividing them into three segments, the HMM decoding algorithm [18-21] is used to simultaneously decode the state sequences within a syllable. In the decoding process, the HMM is a phone-based model. Each single syllable consists of two models, namely the INITIAL and FINAL models, and a decoding process is performed on the state sequences. Hence, if a syllable belongs to a consonant-vowel type, then the INITIAL and FINAL represent the consonant and vowel parts, respectively. On the contrary, if a syllable belongs to a vowel-only type, e.g. a main-vowel, then the INITIAL and FINAL both represent the vowel part. Setting the dimension of the MFCCs to 12 in the input features, there are 59 types of INITIAL and 45 types of FINAL models included in the HMMs. Each INITIAL and each FINAL model contains 3 and 5 states, respectively, with each composed of two mixture Gaussian density functions. Hence, intra a syllable, the first segment represents an INITIAL model with three states, while the second segment and the third one occupy two and three states in the FINAL model, respectively.

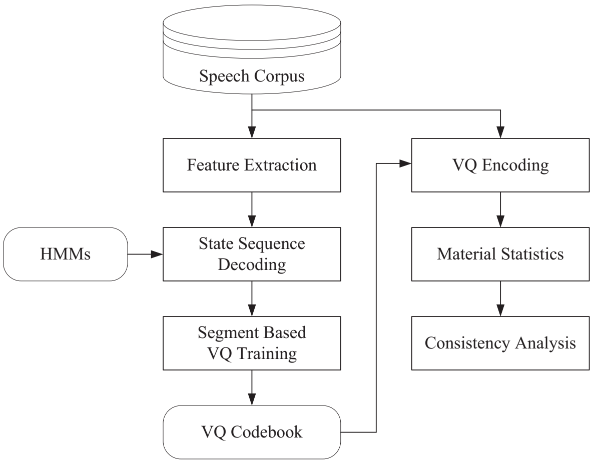


Figure 1. A flowchart of the procedure of consistency analysis

As the second step, based on a designated syllable, the vector quantization (VQ) with the Linde-Buzo-Gray algorithm [22] is used to train the VQ codebooks of each spectral segment with respect to the duration vector. Thus, a total of three codebooks are constructed for each syllable. In this paper, each codebook is set to the size of 4 during the training process, while the codeword dimension within the codebook is determined according to the number of HMM-states in the individual spectral segments. That is, the first and last segments hold the codebook in three dimensions, while the second segment holds it in two dimensions.

The form of $\mathbf{Dur}_{jk}$ representing the duration vector of the jth pattern in the kth syllabic cluster is defined as

$$\mathbf{Dur}_{jk} = \begin{cases} \left[d_{jk}(s_1)\ d_{jk}(s_2)\ d_{jk}(s_3)\right], & \text{for segment \#1} \\ \left[d_{jk}(s_4)\ d_{jk}(s_5)\right], & \text{for segment \#2} \\ \left[d_{jk}(s_6)\ d_{jk}(s_7)\ d_{jk}(s_8)\right], & \text{for segment \#3} \end{cases} \tag{1}$$

where $d_{jk}(s_i)$, $1 \le i \le 8$, is the value of duration in the ith state. The k indicates one of the 411 distinguishable syllables, i.e. $1 \le k \le 411$. The number of the kth syllabic cluster is referred to N_k, $1 \le j \le N_k$.

As the last step, the duration vector of each segment is encoded as an index by a VQ search algorithm. Then, the probability of each possible path, which represents the index seen all the way from the first to the last segment, is evaluated for a designated syllable. Finally, a number of consistency properties can be found and extracted from the probability of a segment sequence.

4. Experimental results and discussions

There are two experiments conducted in this paper. The first consistency analysis is tested on a main-vowel syllable, i.e. the Mandarin syllable "ㄩ", which its international phonetic alphabet (IPA) is labeled as "y". The second is tested on an *initial-final* syllable, i.e. the Mandarin syllable "ㄉㄚ", which its IPA is labeled as "t-a". All the experiments are conducted on a Chinese speech database with 8 kHz sampling frequency and 16-bit PCM format, containing 74,402 syllables out of 4020 sentences by one male speaker, taking 308 MB of storage space and a running time of 336 minutes.

4.1. Consistency analysis for the case of Mandarin syllable "ㄩ"

Taking the Mandarin syllable "ㄩ (y)" as an example to analyze the consistency between the duration and the spectrum in this experiment, the trained VQ codebooks of duration for the syllable "ㄩ (y)" are tabulated in Table 1.

Table 1. Codebooks of a duration pattern in the syllable "y" (Number of training data: 614; codeword unit: 10 ms)

	Codewords in each codebook
Segment #1	[7.950000 1.800000 2.250000] [1.433121 1.566879 2.089172] [1.506024 1.855422 7.542169] [1.468085 6.914894 2.021277]
Segment #2	[4.071429 4.642857] [1.177419 7.725806] [5.736842 1.473684] [1.150259 1.642487]
Segment #3	[1.444444 6.962963 2.111111] [1.104167 1.562500 5.416667] [1.335052 1.335052 1.670103] [5.447369 1.263158 2.236842]

Taking a further step to analyze the whole pronunciation with "y-2", meaning the syllable "y" with the second tone and a subset in the syllable "y", the possible paths and their probabilities for the segment sequences within the syllable "y-2" are tabulated in Table 2. Items "Index1", "Index2", and

"Index3" represent the codebook indices in the first, the second, and the last segment, respectively. Each index, whose value is set from 1 to 4, represents a corresponding codeword in the codebook. There are 268 of the whole pronunciations with "y-2" tested in Table 2, and there are a total of 64 (4*4*4) combinations found in the segment sequences, but a random-like probability distribution is seen as expected on the ground that these syllables embedded in different context bring about different prosodic information. Given a path with Index1=2, Index2=4, and Index3=3 as an example, it indicates that the duration vectors of three segments located in the second, fourth, and third clusters respectively has 0.194030 of probability. It also means that all segments belong to the shortest durations can be seen according to Table 1. Besides, the various path transitions within the syllable demonstrate the different time warping in the same syllable.

Table 2. Possible paths and corresponding probabilities for a segment sequence within the syllable "y-2" (Number for statistic: 268)

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0.007463	0	0.022388	0
Index1=2	Index2=1	0	0	0.029851	0
	Index2=2	0	0.007463	0.111940	0.014925
	Index2=3	0	0	0.067164	0
	Index2=4	0.014925	0.059701	0.194030	0.059701
Index1=3	Index2=1	0	0	0.029851	0
	Index2=2	0	0	0.037313	0
	Index2=3	0	0.007463	0.022388	0
	Index2=4	0.014925	0.029851	0.134328	0.029851
Index1=4	Index2=1	0	0	0	0
	Index2=2	0	0	0.037313	0.014925
	Index2=3	0	0.007463	0	0
	Index2=4	0	0.014925	0.022388	0.007463

Table 3 shows the possible paths and associated probabilities for the segment sequence within the syllable "於 (y-2)" located in the word "終於 (tʂ-uəŋ-1, y-2)". A total of 26 syllables are counted out of the speech database but merely 5 paths are found, which indicates a strongly non-uniform distribution among such probabilities. The largest probability is 0.538, meaning that the duration pattern for syllable "於 (y-2)" embedded in the word "終於 (tʂ-uəŋ-1, y-2)" is consistent.

Moreover, Table 4 shows the possible paths and corresponding probabilities for the segment sequence within the syllable "於 (y-2)" embedded into the word "對於 (t-uei-4, y-2)". As little as 4 paths are found with the largest probability of 0.625 among such paths. As before, it is also indicated that the duration pattern for the syllable "於 (y-2)" located in the word "對於 (t-uei-4, y-2)" is consistent.

Table 3. Possible paths and corresponding probabilities for a segment sequence within the syllable "於 (y-2)" located in the word "終於 (tʂ-uəŋ-1, y-2)" (Number for statistic: 26)

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0	0	0
	Index2=2	0	0	0.538462	0
	Index2=3	0	0	0.115385	0
	Index2=4	0	0	0	0.192308
Index1=3	Index2=1	0	0	0	0
	Index2=2	0	0	0.076923	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0.076923	0
Index1=4	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0	0

Table 4. Possible paths and corresponding probabilities for a segment sequence within the syllable "於 (y-2)" located in the word "對於 (t-uei-4, y-2)" (Number for statistic: 16)

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0	0	0
	Index2=2	0	0	0.125000	0
	Index2=3	0	0	0.625000	0
	Index2=4	0	0	0.187500	0
Index1=3	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0.062500	0
Index1=4	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0	0

In addition, a state diagram of the best path in relation to a duration pattern distributed is made in Figure 2. There is a 0.538 probability that the best path of the syllable "於 (y-2)" is found within the word "終於 (tʂ-uəŋ-1, y-2)", while a 0.625 probability that the best path of the syllable "於 (y-2)" is within the word "對於 (t-uei-4, y-2)", and a 0.194 probability for the best path in the whole syllable "y-2". There is a much higher probabilities that the best path lies in the words "終於 (tʂ-uəŋ-1, y-2)" and "對於 (t-uei-4, y-2)" than there is for the whole syllable. A strong consistency between the duration pattern and the spectrum is validated by these experimental results.

Finally, presented in Figure 3 are duration warping curves for the best path within the whole syllable "y-2", the syllable "於 (y-2)" located in the word "終於 (tʂ-uəŋ-1, y-2)", and the syllable "於 (y-2)" located in the word "對於 (t-uei-4, y-2)", respectively. Each warping curve in Figure 3 is obtained by observation of the Figure 2 and Table 1. It is evident that the same syllable in different word acquires a distinct duration warping curve. These results demonstrate that the influence of time warping curve is not only on the global sentence, but also on the intra-syllable.

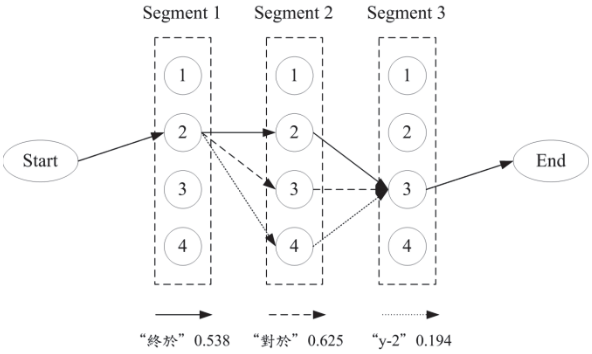


Figure 2. A state diagram of the best path in relation to the duration pattern distributed in the case of the syllable "y-2"

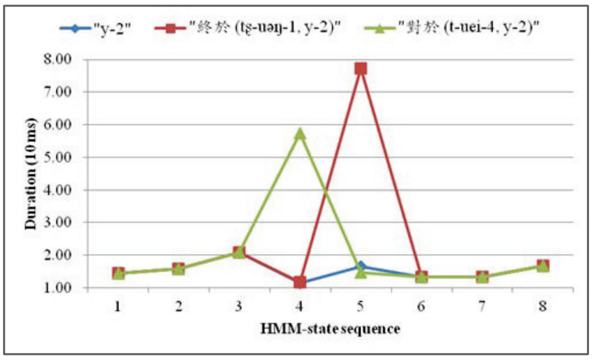


Figure 3. Duration warping curves for the best path within the whole syllable "y-2", the syllable "於 (y-2)" located in the word "終於 (tʂ-uəŋ-1, y-2)", and the syllable "於 (y-2)" located in the word "對於 (t-uei-4, y-2)", respectively

Table 5. Codebooks of a duration pattern in the syllable "t-a" (Number of training data: 770; codeword unit: 10 ms)

	Codewords in each codebook
Segment #1	[1.067308 1.004808 2.591346] [2.038461 1.000000 1.480769] [1.062500 2.125000 1.937500] [1.000000 1.000000 1.000000]
Segment #2	[1.086022 9.860215] [1.421488 5.685950] [1.509434 1.490566] [7.492308 1.430769]
Segment #3	[5.373134 2.313433 2.208955] [10.153846 1.692308 2.961539] [1.571429 1.619048 4.000000] [1.475904 2.090361 1.481928]

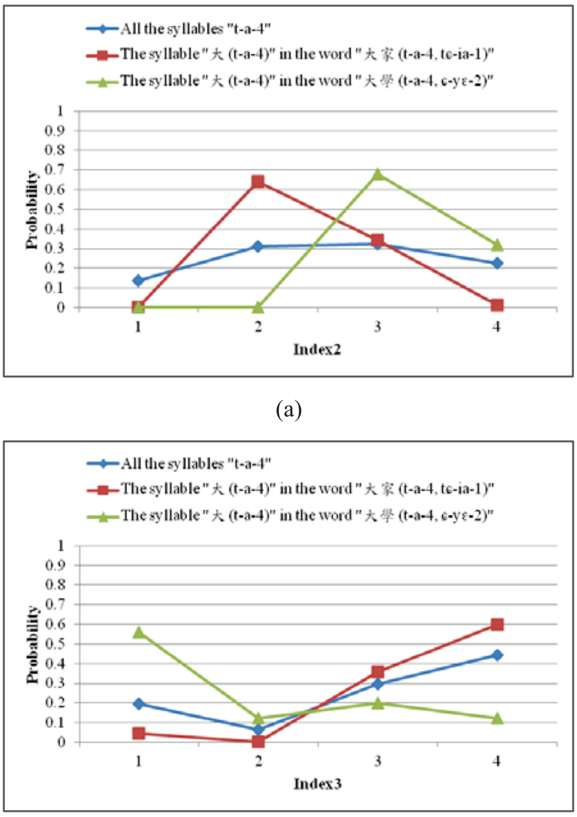
4.2. Consistency analysis for the case of Mandarin syllable "ㄉㄚ"

Taking the Mandarin syllable "ㄉㄚ (t-a)" as the second example to analyze the consistency between the duration and the spectrum in this experiment. The trained VQ codebooks of duration for the syllable "t-a" are presented in Table 5. In this case, the waveform of syllable "ㄉㄚ (t-a)" is composed of an *initial* part and a *final* part. The *initial* part is an unvoiced speech, while the *final* part is a voiced speech, dominating the syllabic waveform. Thus, the consistency analysis is made on the *final* part merely, including the second and the third segments.

To analyze the *final* part, Table 6 lists the path probabilities of a voiced segment concerning the duration pattern in (a) the whole syllable "t-a-4", (b) the syllable "大 (t-a-4)" located in the word "大家 (t-a-4, tɕ-ia-1)", and (c) the syllable "大 (t-a-4)" located in the word "大學 (t-a-4, ɕ-yɛ-2)". Moreover, as illustrated in Figure 4, the distribution of individual segments in Table 6 is alternatively presented in graphic form as Figure 4. Presented in Figure 4(a) is the probability distribution of the duration pattern in the second segment. Thus, each value in Figure 4(a) is obtained by summation of the probabilities of each row in Table 6, that is, the summation of the probabilities from Index3=1 to Index3=4. Likewise, each value in Figure 4(b), meaning the probability distribution of the duration pattern in the third segment, is obtained by summation of the probabilities from Index2=1 to Index2=4. As such, a difference in consistency is seen as before between the words "大家 (t-a-4, tɕ-ia-1)" and "大學 (t-a-4, ɕ-yɛ-2)".

Table 6. Path probability of a voiced segment concerning the duration pattern in (a) the syllable "t-a-4", (b) the word "大家 (t-a-4, tɕ-ia-1)", and (c) the word "大學 (t-a-4, ɕ-yɛ-2)" (Numbers for statistic are 544, 70, and 25 respectively)

Table 6(a)	Index3=1	Index3=2	Index3=3	Index3=4
Index2=1	0.011029	0.007353	0.033088	0.084559
Index2=2	0.029412	0.003676	0.110294	0.169118
Index2=3	0.139706	0.047794	0.080882	0.055147
Index2=4	0.014706	0.003676	0.073529	0.136029
Table 6(b)	Index3=1	Index3=2	Index3=3	Index3=4
Index2=1	0	0	0	0
Index2=2	0	0	0.114286	0.528571
Index2=3	0.042857	0	0.228571	0.071429
Index2=4	0	0	0.014286	0
Table 6(c)	Index3=1	Index3=2	Index3=3	Index3=4
Index2=1	0	0	0	0
Index2=2	0	0	0	0
Index2=3	0.560000	0.120000	0	0
Index2=4	0	0	0.200000	0.120000



(a)

(b)

Figure 4. Probability distribution of the duration pattern in (a) the second segment and (b) the third segment within the syllable “t-a-4”

In addition, Figure 5 shows a state diagram of the best path in relation to the duration pattern distributed. There is a 0.528 probability that the best path of the syllable “t-a-4” is found within the word “大家 (t-a-4, tɕ-ia-1)”, while a 0.560 probability that the best path of the syllable “t-a-4” is within the word “大學 (t-a-4, ɕ-yɛ-2)”, and a 0.169 probability that the best path is in the whole syllable “t-a-4”. A strong consistency of the duration pattern is verified by these experimental results.

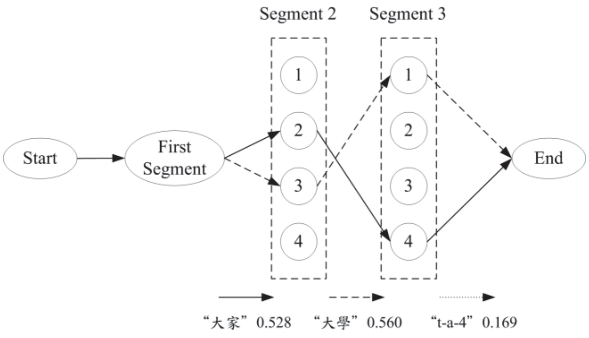


Figure 5. A state diagram of the best path in relation to the duration pattern distributed in the case of the syllable “t-a-4”

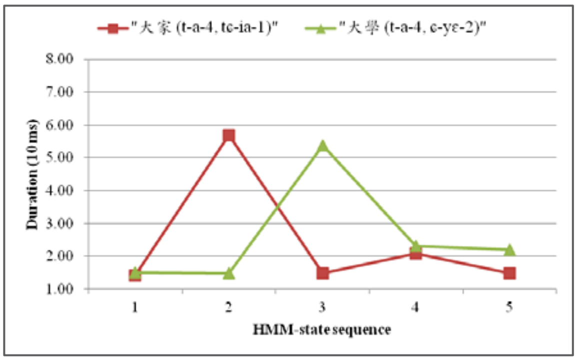


Figure 6. Duration warping curves for the best path within the syllable “大 (t-a-4)” located in the word “大家 (t-a-4, tɕ-ia-1)” and the syllable “大 (t-a-4)” located in the word “大學 (t-a-4, ɕ-yɛ-2)”, respectively

Finally, presented in Figure 6 are duration warping curves for the best path within the syllable “大 (t-a-4)” located in the word “大家 (t-a-4, tɕ-ia-1)” and the syllable “大 (t-a-4)” located in the word “大學 (t-a-4, ɕ-yɛ-2)”, respectively. Each warping curve in Figure 6 is obtained by observation of the Figure 5 and Table 5. The warping curve for the best path within the whole syllable “t-a-4” is identical to which within the word “大家 (t-a-4, tɕ-ia-1)”. The same syllable in different word acquires a distinct duration warping curve is again verified.

5. Conclusions

This paper focuses on the consistency analysis of duration parameter for Mandarin speech. It is validated experimentally that the warping curves of the duration and the spectrum intra a syllable are consistent in the case where the syllable lies in exactly the same word. It is also concluded that various words have various characteristics of consistency, giving rise to the research direction that the time warping process intra a syllable must be taken into account in a TTS system as a way to improve synthesized speech quality.

Acknowledgment

This research was financially supported by the Ministry of Economic Affairs under Grant No. 100-EC-17-A-03-S1-123, Taiwan, R.O.C.

References

[1] **D. H. Klatt.** Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 1987, Vol. 82, No. 3, pp. 737–793.

[2] **L. S. Lee, C. Y. Tseng, O. Y. Ming.** The synthesis rules in a Chinese text-to-speech system. In: *IEEE Trans. Acoustics, Speech and Signal Processing*, 1989, Vol. 37, No. 9, pp. 1309–1320.

[3] **M. H. O'Malley.** Text-to-speech conversion technology. *Computer*, 1990, Vol. 23, No. 8, pp. 17–23.

[4] **S. H. Hwang, S. H. Chen, Y. R. Wang.** A Mandarin text-to-speech system. In: *Proceedings of ICSLP*, 1996, pp. 1421–1424.

[5] **T. Anbinderis.** Automatic stressing of Lithuanian text using decision treles. *Information Technology and Control*, 2010, Vol. 39, No. 1, pp. 61–67.

[6] **S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis.** Embedded unit selection text-to-speech synthesis for mobile devices. In: *IEEE Trans. Consumer Electronics*, 2009, Vol. 55, No. 2, pp. 613-621.

[7] **D. J. Yue.** Two stage concatenation speech synthesis for embedded devices. In: *Proceedings of ICALIP*, 2010, pp. 1652–1656.

[8] **C. Spelta, V. Manzoni, A. Corti, A. Goggi, S. M. Savaresi.** Smartphone-based vehicle-to-driver/ environment interaction system for motorcycles. In: *IEEE Embedded Systems Letters*, 2010, Vol. 2, No. 2, pp. 39–42.

[9] **A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, S. Raptis.** A unit selection text-to-speech synthesis system optimized for use with screen readers. In: *IEEE Trans. Consumer Electronics*, 2010, Vol. 56, No. 3, pp. 1890–1897.

[10] **C. H. Wu, J. H. Chen.** Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis. *Speech Communication*, 2001, Vol. 35, No. 3–4, pp. 219–237.

[11] **F. C. Chou, C. Y. Tseng, L. S. Lee.** A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese. In: *IEEE Trans. Speech and Audio Processing*, 2002, Vol. 10, No. 7, pp. 481–494.

[12] **J. R. Bellegarda.** A Dynamic Cost Weighting Framework for Unit Selection Text–to–Speech Synthesis. In: *IEEE Trans. Audio, Speech and Language Processing*, 2010, Vol. 18, No. 6, pp. 1455–1463.

[13] **E. Moulines, F. Charpentier.** Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 1990, Vol. 9, No. 5–6, pp. 453–467.

[14] **Y. Zhu, L. Zhao, Y. Xu, Y. Niimi.** A Chinese text-to-speech system based on T D-PSOLA. In: *Proc. TENCON*, 2002, pp. 204–207.

[15] **S. H. Chen, S. H. Hwang, Y. R. Wang.** An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech. In: *IEEE Trans. Speech and Audio Processing*, 1998, Vol. 6, No. 3, pp. 226–239.

[16] **Z. Ying, X. Shi.** An RNN-based algorithm to detect prosodic phrase for Chinese TTS. In: *Proc. ICASSP*, 2001, pp. 809–812.

[17] **C. Y. Yeh, S. H. Hwang.** Efficient text analyzer with prosody generator-driven approach for Mandarin text-to-speech. In: *IEE Proc. Vision, Image and Signal Processing*, 2005, Vol. 152, No. 6, pp. 793–799.

[18] **X. D. Huang, A. Acero, H. W. Hon.** Hidden Markov models. In: *Spoken Language Processing. Prentice Hall PTR, New Jersey, USA*, 2001.

[19] **L. R. Rabiner.** A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE*, 1989, Vol. 77, No. 2, pp. 257–286.

[20] **T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura.** Speech parameter generation algorithms for HMM-based speech synthesis. In: *Proc. ICASSP*, 2000, pp. 1315–1318.

[21] **H. Zen, K. Tokuda, A. W. Black.** Statistical parametric speech synthesis. *Speech Communication*, 2009, Vol. 51, No. 11, pp. 1039–1064.

[22] **Y. Linde, A. Buzo, R. Gray.** An algorithm for vector quantizer design. In: *IEEE Trans. Communications*, 1980, Vol. 28, No. 1, pp. 84–95.

Received May 2012.