A Study on the Consistency Analysis of Duration Parameter for Mandarin Speech

Cheng-Yu Yeh[1], Kuan-Lin Chen[2], Shaw-Hwa Hwang[2]
[1, 2]Department of Electrical Engineering
[1]National Chin-Yi University of Technology
Taichung, Taiwan
[2]National Taipei University of Technology
Taipei, Taiwan
cy.yeh@ncut.edu.tw, kent_kuan@yahoo.com.tw, hsf@ntut.edu.tw

ABSTRACT: *In this work, a consistency analysis of duration parameter for Mandarin speech is presented. Found by an inspection on the pronunciation process of human beings, the consistency can be interpreted as a high correlation of a warping curve between 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 LBG algorithm is used 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 speech quality.*

Keywords: Consistency Analysis, Hidden Markov Model (HMM), Vector Quantization (VQ), Text-to-Speech (TTS)

Received: 19 December 2011, Revised 2 February 2012, Accepted 9 February 2012

1. Introduction

A text-to-speech (TTS) system [1-3] is a system converting a text input into a speech output, and applied to smart human computer interfaces and auxiliary speech systems for the visual impaired. In the era of multimedia communications, the growing significance of TTS is seen definitely for the reason that it can be found in a wide variety of applications such as general consumer electronics, robots, virtual anchors, text messages of cell phone, and smart speech service systems.

Moreover, due to the growing demand of embedded systems, a wide range of portable devices, e.g. smartphones, ebooks and relevant products, have been popularized in the market, and extended developments look promising. Consequently, integration of TTS systems into embedded systems becomes one of the hottest research issues these days [4-6]. In an attempt to implement TTS technology on an embedded system, there are two additional requirements imposed on such integrated system, that is, a low memory requirement and a low computational complexity.

Reviewing the history of TTS technology development, the waveform-based synthesis units approach [7-13] is one of the most commonly used technology in TTS. This approach is further classified into two types in terms of the way it is synthesized. One is the corpus-based synthesis units [7-9], and the other is the model-based synthesis units approaches [10-13]. This corpus-based speech synthesis technique relies on a unit selection method and compilation of speech units from a large speech

database. The speech database usually derives from a sufficiently large corpus where appropriately selected spoken utterances are carefully annotated to the unit level. However, this approach requires a great number of speech units, that is, a large deal of storage space is needed to reach a superior speech quality. In contrast, the model-based technique adopts a small size synthesis unit, which treats a set of fundamental speech elements, e.g. phonemes, diphones or syllables as synthesis units, then a synthesized speech is made through a prosodic modification conducted on synthesized units by pitchsynchronous overlap-add (PSOLA) algorithms [10, 11]. Accordingly, a double advantage of requiring a low memory and a low computation load is reached with an inferior but comparable speech quality relative to corpus-based methods.

However, the TTS with the waveform-based synthesis units approach necessitates a prosody model all the time to deal with the prosodic modification on synthesized units. Exploring the pronunciation process of human beings, the speech is made by an excitation source flowing through the vocal tract and emanating from the mouth and the nostrils of a speaker. The excitation source containing the airflow and the vibration of vocal cords reflects the prosodic information. Both the vocal tract, affecting the voice spectrum, and the excitation source couple to generate a natural and fluent speech. Thus, an inspection result is seen, which is the prosody and the spectrum embedded in the running speech is consistent. Definitely as one of significant issues for a TTS system, the spectrum and prosody modules are addressed separately in most cases, leading to an inconsistency between both of them. Therefore, it motivates us to demonstrate the consistency between the prosody and the spectrum embedded in the running speech is existent.

In the cause of verifying the consistency property, the definition of consistency will be firstly explained in this work. Subsequently, the consistency analysis between the duration parameter of prosody and the spectrum is focused and discussed. The analytic methods, procedures, and practical experiments are presented to demonstrate the proposed deduction. It is also expected to upgrade the performance of Mandarin TTS system through the research in this paper.

The rest of this paper is outlined as follows. The modeling of the consistency analysis in Mandarin speech is described in Section 2. Presented in Section 3 is a procedure of the consistency analysis between the duration parameter of prosody and the spectrum. Experimental results are demonstrated and discussed in Section 4. This work is summarized at the end of this paper.

2. Modeling of the Consistency Analysis

As referred to previously, an inspection on the pronunciation process of human beings, both the excitation source and the vocal tract couple to generate a natural and fluent speech. The excitation source reflects the prosodic information, and the vocal tract affects the voice spectrum. The prosodic information usually contains the pitch contour, duration, and energy parameters. In this work, the consistency property between the duration and the spectrum is analyzed. The definition and modeling of the consistency analysis in Mandarin speech is presented.

In the Chinese language phonology, there is a total of 411 distinguishable syllables composed of an optional consonant and a vowel as basic pronunciation units in a Mandarin speech. However, a Chinese word consisting of a minimum of one syllable is regarded as the smallest unit that is meaningful. Besides, the waveform and the spectrum of all the same pronunciation units are definitely not identical because the speech signal is a non-stationary signal. Thus, the consistency can be interpreted as the high correlation of a warping curve between the spectrum and the prosody intra a syllable. In other words, the warping curves are consistent as long as the same pronunciations are located in the same Chinese word, implying that the same pronunciations located in different Chinese words brings about distinct consistency, that is, different warping curves are made. Observing the warping curve can help us to further acquire the knowledge of the detail 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 curve between the spectrum and the duration intra a syllable is the one 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 Melfrequency cepstral coefficients (MFCCs) are used as spectral feature and the HMMs are employed to decode the state sequence within a syllable [14]. For evaluation of the MFCCs, the discrete Fourier transform (DFT) is first performed to obtain its spectrum

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, 0 \le k \le N \quad (1)$$

then, a filterbank with M filters according to Mel-scale is defined by:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{(f[m-1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (2)$$

where $1 \le m \le M$ and the boundary points $f[m]$ are uniformly spaced in the Mel-scale:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \quad (3)$$

where f_l and f_h are the lowest and the highest frequencies of the filterbank, F_s is the sampling rate, and the Mel-scale B and its inverse B^{-1} are given by:

$$B(f) = 1125 \ln(1 + f/700) \quad (4)$$

$$B^{-1}(b) = 700\,(e^{(b/1125)} - 1) \quad (5)$$

Thus, the log-energy at the output of each filter is computed as

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right] \quad (6)$$

and then MFCCs are obtained as

$$c[n] = \sum_{m=1}^{M} S[m] \cos(\pi n (m - 1/2)/M), \quad 0 \le n < L \quad (7)$$

where L is the order of MFCC, $L < M$. In this work, the L is set to 12, N 512, M 64, $F_s = 8000\ Hz$, $f_l = 0\ Hz$, and $f_h = 4000\ Hz$.

On the other hand, for exploring the duration information within a spectral segment, each syllable will be divided into three spectral segments, and each spectral segment contains two to three HMM states. Based on spectral segment, all the state durations are employed as a duration vector, and then a clustering algorithm is used to analyze the duration vector. Thus, the warping curve can be analyzed by exploring clustering result of the duration vector within a spectral segment.

3. Procedure of the Consistency Analysis

Presented in Figure 1 is a flowchart of the procedure of consistency analysis. There are three steps required in the procedure. Firstly, the feature extraction such as MFCCs, energy parameter etc. are computed from a large speech database. Then, dividing them into three segments, the HMM decoding algorithm [14-17] is used to decode the state sequences within a syllable at the same time.

The first segment represents the consonant *initial* which is an INITIAL model with three states. The second segment and the third represent the vowel *final* segment and occupy two and three states in the FINAL model respectively. In this paper, the MFCCs' dimension is set to 12. There are 22 types of INITIAL, 39 types of FINAL, one type of silence and breath models included in the HMMs. Each INITIAL model, each FINAL model, the silence and breath models contain 3 states, 5 states and one state, respectively, with each composed of two mixture Gaussian density functions.

As the second step, based on a designated syllable, the vector quantization (VQ) with the Linde-Buzo-Gray (LBG) algorithm [18] is used to train the VQ codebooks of each spectral segment with respect to the duration vector. Thus, there is a total of three codebooks constructed in each syllable. In this paper, setting each codebook to the size of 4 during the training process, the codeword dimension within the codebook is determined according to the number of HMM-states in individual spectral segment. That is, the first and the last segments hold the codebook in three dimensions, respectively, and the second segment holds the codebook in two dimensions.

The forms of $\mathbf{Dur}_{jk}$ representing the duration vector of the *j*th pattern in the *k*th syllabic cluster is defined as

$$\mathbf{Dur}_{jk} = \begin{cases} \left[d_{jk}(s_1) \;\; d_{jk}(s_2) \;\; d_{jk}(s_3) \right], & \text{for segment \#1} \\ \left[d_{jk}(s_4) \;\; d_{jk}(s_5) \;\; \right], & \text{for segment \#2} \\ \left[d_{jk}(s_6) \;\; d_{jk}(s_7) \;\; d_{jk}(s_8) \right], & \text{for segment \#3} \end{cases} \tag{8}$$

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

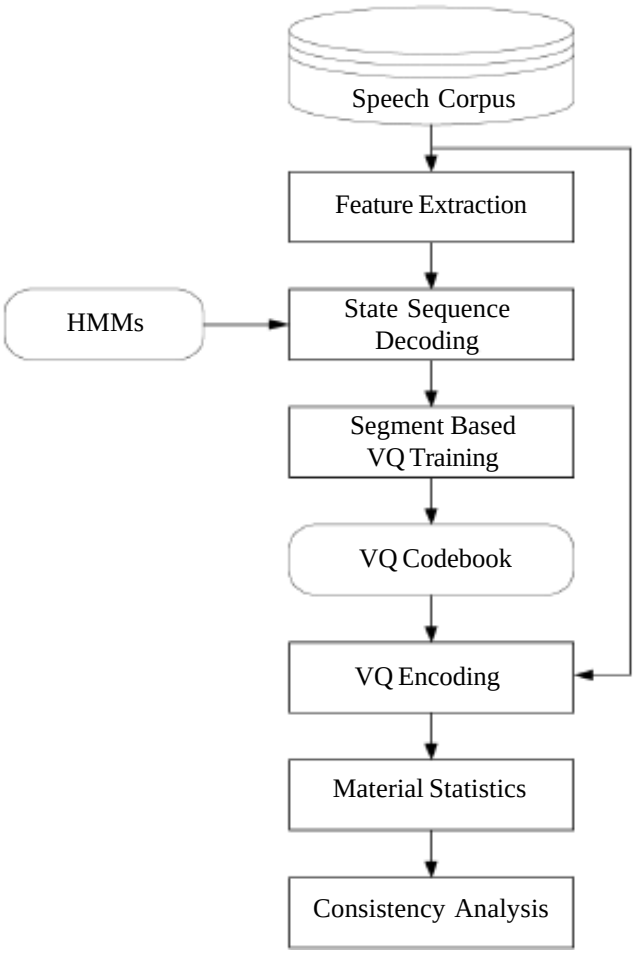

Figure 1. A flowchart of the procedure of consistency analysis

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

In this work, the consistency analysis concerning the duration is examined. The experiment is conducted on a Chinese speech database, containing 39,159 syllables out of 3,365 sentences by one male speaker, taking 398 MB of storage space and a running time of 175 minutes.

Taking the Mandarin syllable "ㄓ", the international phonetic alphabet (IPA) is labeled as "ʈʂ", 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 "ㄓ (ʈʂ)" are tabulated in Table 1.

Taking a further step to analyze the whole pronunciations with "ʈʂ-1", meaning the syllable "ʈʂ" with the first tone and a subset in the syllable "ʈʂ", the possible paths and their probabilities for the segment sequences within the syllable "ʈʂ- " are tabulated in Table 2. Items "*Index*1", "*Index*2", and "*Index*3" represent the codebook indices in the first, the second, and the last segment respectively. Each index, which its value is set from 1 to 4, represents a corresponding codeword in the codebook. There are 216 of the whole pronunciations with "ʈʂ - 1" tested in Table II, and there is a total of 64 (4 * 4 * 4) combinations found in the segment sequences, but a zero probability for 24 paths which meaning some strange timings for pronunciation are not occurred. Given a path with Index1 = 1, Index2 = 1, and Index3 = 1 as an example, it indicates that the duration vectors of all segments located in the first cluster respectively has 0.074074 of probability. It also means that all segments belong to lower durations can be seen according to Table I. Besides, the various path transitions within the syllable demonstrate the different time warping in the same syllable.

	Codewords in each codebook
Segment #1	[1.071429 1.803571 1.571429] [2.066667 1.493333 2.306667] [2.016667 2.000000 3.150000] [2.067227 3.378151 3.050420]
Segment #2	[1.000000 1.212903] [2.050584 1.194553] [1.627907 3.116279] [2.304348 4.739130]
Segment #3	[1.265306 1.265306 1.931973] [1.457317 1.591463 4.524390] [2.478261 3.760870 2.217391] [4.857143 4.734694 4.163265]

Table 1. Codebooks of a Duration Pattern in the Syllable "ʈʂ" (Number of Training Data: 501; Codeword Unit: 10 ms)

In most cases, the syllabic waveform of a Mandarin speech is composed of an *initial* part and a *final* part. Most of the *initial* part is an unvoiced speech, while the *final* part is a voiced speech, dominating the syllabic waveform. Thus, a further discussion and analysis are made on the *final* part merely, including the second and the third segments. Tabulated in Table 3 are the path probabilities of a voiced segment concerning the duration pattern in (a) the whole syllable "ʈʂ-1", (b) the syllable "知(ʈʂ-1)" located in the word "知道(ʈʂ-1, tau- 4)", and (c) the "之(ʈʂ-1)" located in the word "之後(ʈʂ-1, x-ou- 4)".

As tabulated in Table III(a), a random-like probability distribution is seen as expected on the ground that these syllables embedded in different context bring about different prosodic information. However, the distribution is concentrated in certain circumstances in case these syllables lie in the same word. This argument is illustrated with two examples of the words "知道(ʈʂ-1, tau- 4)" and "之後(ʈʂ-1, x-ou- 4)", as shown in Table III(b) and III(c). Accordingly, a strong consistency between the duration and the spectrum is obtained when the same syllables located in the same word.

Furthermore, a state diagram of the best path in relation to a duration pattern distributed is made in Figure 2. There is a 0.6875

probability that the best path of the syllable "知(ʈʂ-1)" is embedded into the word "知道(ʈʂ-1, *tau*- 4)", while a 0.833 probability that the best path of the syllable "之 (ʈʂ -1)" is into the word "之後(ʈʂ-1, *x-ou*- 4)", and a 0.199 probability for the best path in the whole syllable "ʈʂ -1". The higher probability represents the consistency is more explicit. It is validated as well that the same syllable located in different word acquires a distinct duration warping curve, that is, the diverse time warping occurred when the same syllable in different word. These results demonstrate that the influence of warping curve is not only on the global sentence, but also on the intra-syllable.

		Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index1 = 1	Index2 = 1	0.074074	0.032407	0	0
	Index2 = 2	0.101852	0.060185	0.023148	0.004630
	Index2 = 3	0	0.009259	0	0
	Index2 = 4	0	0	0	0
Index1 = 1	Index2 = 1	0.046296	0.060185	0.009259	0
	Index2 = 2	0.055556	0.083333	0.009259	0.046296
	Index2 = 3	0.004630	0.009259	0	0
	Index2 = 4	0	0	0	0
Index1 = 1	Index2 = 1	0.009259	0.074074	0.013889	0.037037
	Index2 = 2	0.032407	0.009259	0.018519	0.013889
	Index2 = 3	0.004630	0	0	0.009259
	Index2 = 4	0	0	0.004630	0.004630
Index1 = 1	Index2 = 1	0.004630	0.004630	0	0.018519
	Index2 = 2	0.009259	0.018519	0.018519	0.041667
	Index2 = 3	0.004630	0	0.004630	0.009259
	Index2 = 4	0	0	0	0.004630

Table 2. Path Probabilities for a Segment Sequence Within the Syllable "ʈʂ -1" (Number for Statistic: 216)

Table 3(a)	Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index2 = 1	0.134259	0.171296	0.023148	0.055556
Index2 = 2	0.199074	0.171296	0.069444	0.106481
Index2 = 3	0.013889	0.018519	0.004630	0.018519
Index2 = 4	0	0	0.004630	0.009259
Table 3(b)	Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index2 = 1	0.062500	0.687500	0	0
Index2 = 2	0	0.250000	0	0
Index2 = 3	0	0	0	0
Index2 = 4	0	0	0	0
Table 3(c)	Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index2 = 1	0.833333	0	0	0
Index2 = 2	0	0	0.166667	0
Index2 = 3	0	0	0	0
Index2 = 4	0	0	0	0

Table 3. Path Probability of a Voiced Segment Concerning the Duration Pattern in (A) The Syllable "ʈʂ -1", (B) The Word "知道(ʈʂ -1, *tau*- 4)", and (C) The Word "之後 (ʈʂ -1, *x-ou*- 4)" (Numbers for Statistic are 216, 16, and 6 Respectively)

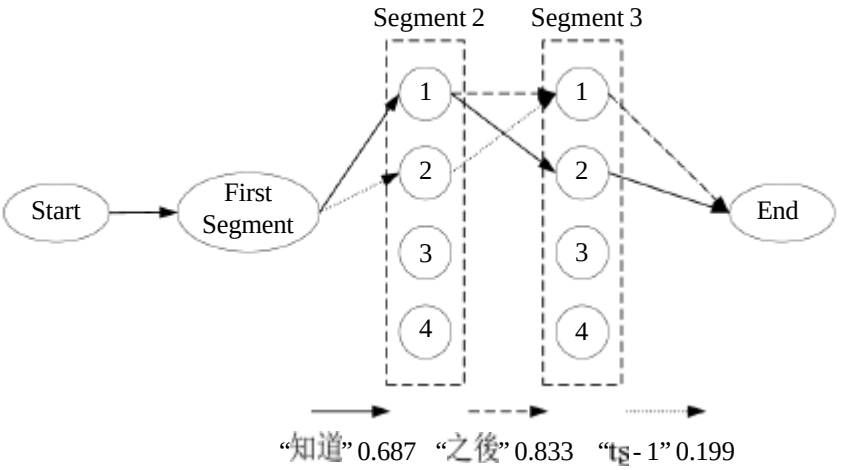

Figure 2. A state diagram of the best path in relation to the duration pattern distributed

5. Conclusions

This paper is proposed mainly with a focus on the consistency analysis of duration parameter for Mandarin speech. It is validated experimentally that the warping curve between the duration and the spectrum intra a syllable is of the consistency in case the syllable lies exactly in the same word. It is also concluded that various words hold various characteristics of consistency, giving rise to a research direction that the time warping process intra a syllable must be taken into account in a TTS system as a way to improve the synthesized speech quality. Besides, the consistency analyses concerning other prosodic parameters will be further studied in the future.

References

[1] Klatt, D. H. (1987). Review of text-to-speech conversion for English, *Acoust. J., Soc. Am.*, 82 (3) 737-793.

[2] Lee, L. S., Tseng, C. Y., Ming, O.Y. (1989). The synthesis rules in a Chinese text-to-speech system, *IEEE Trans. Acoust. Speech Signal Process.*, 37 (9) 1309-1320.

[3] O’Malley, M. H. (1990). Text-to-speech conversion technology, *Computer*, 23 (8) 17-23.

[4] Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., Raptis, S. (2009). Embedded unit selection text-to-speech synthesis for mobile devices, *IEEE Trans. Consum. Electron.*, 55 (2) 613-621.

[5] Chalamandaris, A., Karabetsos, S., Tsiakoulis, P., Raptis, S. (2010). A unit selection text-to-speech synthesis system optimized for use with screen readers, *IEEE Trans. Consum. Electron*., 56 (3) 1890-1897.

[6] Yue, D. J. (2010). Two stage concatenation speech synthesis for embedded devices, *In*: Proc. ICALIP, p. 1652-1656.

[7] Wu, C. H., Chen, J. H. (2001). Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis, Speech Commun., 35 (3-4) 219-237.

[8] Chou, F. C., Tseng, C. Y., Lee, L. S. (2002). A set of corpus-based text-tospeech synthesis technologies for Mandarin Chinese, *IEEE Trans. Speech Audio Process.*, 10 (7) 481-494.

[9] Bellegarda, J. R. (2010). A Dynamic Cost Weighting Framework for Unit Selection Text–to–Speech Synthesis, *IEEE Trans. Audio Speech Lang. Process.*, 18 (6) 1455-1463.

[10] Moulines, E., Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Commun.*, 9 (5-6) 453-467.

[11] Zhu, Y., Zhao, L., Xu, Y., Niimi, Y. (2002). A Chinese text-to-speech system based on TD-PSOLA, *In*: Proc. TENCON, p. 204-207.

[12] Cheng, S. H., Hwang, S. H., Wang, Y. R. (1998). An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech, *IEEE Trans. Speech Audio Process.*, 6 (3) 226-239.

[13] Yeh, C. Y., Hwang, S. H. (2005). Efficient text analyzer with prosody generator-driven approach for Mandarin text-to-speech, *IEEE Proc. Vis. Image Signal Process.*, 152 (6) 793-799.

[14] Huang, X. D., Acero, A., Hon, H. W. (2001). Spoken Language Processing, Prentice Hall PTR, New Jersey.

[15] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *In*: Proc. IEEE, 77 (2) 257- 286.

[16] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis, *In*: Proc. ICASSP, 3, 1315-1318.

[17] Zen, H., Tokuda, K., Black, A.W. (2009). Statistical parametric speech synthesis, *Speech Commun*., 51 (11) 1039-1064.

[18] Linde, Y., Buzo, A., Gray, R. (1980). An algorithm for vector quantizer design, *IEEE Trans. Commun*., 28 (1) 84-95.