Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 266–277

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A consistency analysis on an acoustic module for Mandarin text-to-speech

Cheng-Yu Yeh [a,*], Shun-Chieh Chang [b], Shaw-Hwa Hwang [b]

[a] Department of Electrical Engineering, National Chin-Yi University of Technology, 57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 41170, Taiwan, ROC
[b] Department of Electrical Engineering, National Taipei University of Technology, 1, Sec. 3, Chung-hsiao E. Rd., Taipei 10608, Taiwan, ROC

Received 19 January 2012; received in revised form 4 July 2012; accepted 14 August 2012
Available online 25 August 2012

Abstract

In this work, a consistency analysis on an acoustic module for a Mandarin text-to-speech (TTS) is presented as a way to improve the speech quality. 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 Linde–Buzo–Gray (LBG) algorithm is used to train the VQ codebooks of each segment. Thirdly, the prosodic 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 warping process between the spectrum and the prosody intra a syllable must be considered in a TTS system to improve the speech quality.
© 2012 Elsevier B.V. All rights reserved.

Keywords: Consistency analysis; Hidden Markov model (HMM); Vector quantization (VQ); Acoustic module; Text-to-speech (TTS); Speech synthesis

1. Introduction

A text-to-speech (TTS) system (Klatt, 1987; Lee et al., 1989; O'Malley, 1990; Hwang et al., 1996; Chou et al., 1997) 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. mobile phones, smartphones, e-books 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 (Dey et al., 2007; Guo et al., 2008; Karabetsos et al., 2009; Chalamandaris et al., 2010; Spelta et al., 2010; Yue, 2010). 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 (Chou and Tseng, 1998; Wu and Chen, 2001; Chou et al., 2002; Bellegarda, 2010; Moulines and Charpentier, 1990; Zhu et al., 2002; Chen et al., 1998; Ying and Shi, 2001; Chou et al.,

* Corresponding author. Tel.: +886 4 23924505x7236; fax: +886 4 23924419.
E-mail addresses: cy.yeh@ncut.edu.tw (C.-Y. Yeh), t6319011@ntut.edu.tw (S.-C. Chang), hsf@ntut.edu.tw (S.-H. Hwang).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2012.08.009

1996; Hwang and Yeh, 2005; Yeh and Hwang, 2005; Yeh and Chen, 2010) 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 (Chalamandaris et al., 2010; Chou and Tseng, 1998; Wu and Chen, 2001; Chou et al., 2002; Bellegarda, 2010), and the other is the model-based synthesis units approaches (Hwang et al., 1996; Moulines and Charpentier, 1990; Zhu et al., 2002; Chen et al., 1998; Ying and Shi, 2001; Chou et al., 1996; Hwang and Yeh, 2005; Yeh and Hwang, 2005; Yeh and Chen, 2010). 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. 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, that is, a large deal of storage space is needed to reach a superior speech quality.

In contrast, the model-based speech synthesis 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 pitch-synchronous overlap-add (PSOLA) algorithms (Moulines and Charpentier, 1990; Zhu et al., 2002). 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 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 prosody and the spectrum. Experimental results are demonstrated and discussed in Section 4. A system description of improved Mandarin TTS and its performance assessment are presented in Section 5. This work is summarized at the end of this paper.

2. Modeling of the consistency analysis in Mandarin speech

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 prosody 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 *initial* and a vowel *final* 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 prosody intra a syllable is the one of interest. The warping curve within a syllable can be obtained by exploring the prosodic 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 (Huang et al., 2001). For evaluation of the MFCCs, the discrete Fourier transform (DFT) is first performed to obtain its spectrum

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

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

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

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

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

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:

$$B(f) = 1125\ln(1 + f/700) \tag{4}$$

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

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

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

and then MFCCs are obtained as

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

where L is the order of MFCC, $L < M$. In this work, the L is set to 24, N 512, M 64, $F_s = 8000$ Hz, $f_l = 0$ Hz, and $f_h = 4000$ Hz.

On the other hand, for exploring the prosodic 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 prosodies are employed as a prosodic vector, and then a clustering algorithm is used to analyze the prosodic vector. Thus, the warping curve can be analyzed by exploring clustering result of the prosodic vector within a spectral segment.

3. Procedure of the consistency analysis

Presented in Fig. 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, prosodic information including the pitch contour, duration, and energy parameters etc. are computed from a large speech database. Hence, the consistency analyses are made in four aspects, that is, the warping curve (1) between the spectrum and duration intra a syllable, (2) between the spectrum and energy, (3) between the spectrum and pitch mean, defined as the average value of the pitch contour, and (4) between the spectrum and a prosodic unit consisting of three of the prosodic parameters together.

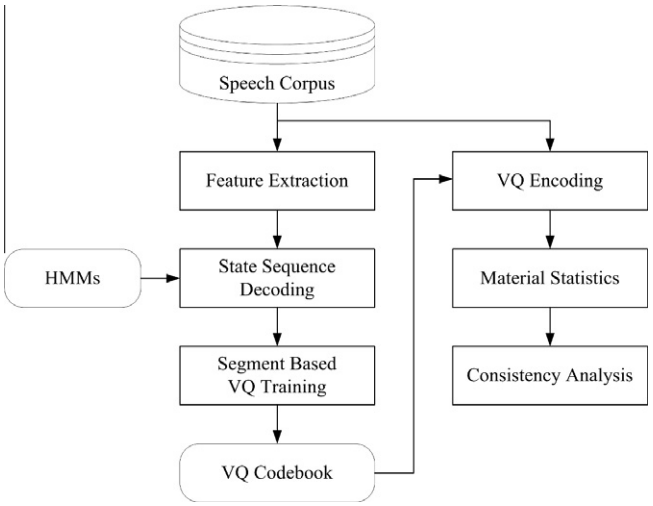

Fig. 1. A flowchart of the procedure of consistency analysis.

Then, dividing them into three segments, the HMM decoding algorithm (Huang et al., 2001; Rabiner, 1989; Yoshimura et al., 2000; Zen et al., 2009) is used to decode the state sequences within a syllable at the same time. 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 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 24 in input features, there are 59 types of INITIAL and 45 types of FINAL models included in the HMMs. Each INITIAL model and each FINAL model contain 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 and the third occupy two and three states in the FINAL model, respectively.

As the second step, based on a designated syllable, the vector quantization (VQ) with the Linde–Buzo–Gray (LBG) algorithm (Linde et al., 1980) is used to train the VQ codebooks of each spectral segment with respect to three of prosodic vectors containing the duration, energy, and the pitch mean respectively. Thus, there is a total of nine 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 three codebooks in three dimensions, respectively, and the second segment holds three codebooks in two dimensions.

The forms of $\mathbf{Pth}_{jk}$, $\mathbf{Dur}_{jk}$, and $\mathbf{Eng}_{jk}$, representing the vectors of pitch mean, duration, and energy of the jth pattern in the kth syllabic cluster respectively, are defined as

$$\mathbf{Pth}_{jk} = \begin{cases} [p_{jk}(s_1)p_{jk}(s_2)p_{jk}(s_3)], & \text{for segment\#1} \\ [p_{jk}(s_4)p_{jk}(s_5)], & \text{for segment\#2} \\ [p_{jk}(s_6)p_{jk}(s_7)p_{jk}(s_8)], & \text{for segment\#3} \end{cases} \quad (8)$$

$$\mathbf{Dur}_{jk} = \begin{cases} [d_{jk}(s_1)d_{jk}(s_2)d_{jk}(s_3)], & \text{for segment\#1} \\ [d_{jk}(s_4)d_{jk}(s_5)], & \text{for segment\#2} \\ [d_{jk}(s_6)d_{jk}(s_7)d_{jk}(s_8)], & \text{for segment\#3} \end{cases} \quad (9)$$

$$\mathbf{Eng}_{jk} = \begin{cases} [e_{jk}(s_1)e_{jk}(s_2)e_{jk}(s_3)], & \text{for segment\#1} \\ [e_{jk}(s_4)e_{jk}(s_5)], & \text{for segment\#2} \\ [e_{jk}(s_6)e_{jk}(s_7)e_{jk}(s_8)], & \text{for segment\#3} \end{cases} \quad (10)$$

where $p_{jk}(s_i)$, $d_{jk}(s_i)$, and $e_{jk}(s_i)$, $1 \leqslant i \leqslant 8$, are the values of pitch mean, duration, and energy in the ith HMM-state respectively. The index k indicates one of the 411 distinguishable syllables, i.e. $1 \leqslant k \leqslant 411$. The number of the kth syllabic cluster is referred to the N_k and $1 \leqslant j \leqslant N_k$.

Nevertheless, the prosodic unit comprising the pitch mean, duration, and energy together is taken into different consideration. That is, there are three different circumstances in comparison with the above procedure. Firstly, based on every HMM-state, the normalization of each prosodic parameter within the prosodic vector is performed separately in an attempt to remove the effect stemming from each prosodic parameter in the course of the training procedure of VQ. Secondly, the vector dimension of 9 is given in the first and the third segments respectively and the vector dimension of 6 is given in the second segment. Thirdly, the size of 8 is set in the VQ codebook. The prosodic vector, $\mathbf{Prosody}_{jk}$, of the jth pattern in the kth syllabic cluster is defined as

$$\mathbf{Prosody}_{jk} = [\mathbf{Pth}_{jk}|\mathbf{Dur}_{jk}|\mathbf{Eng}_{jk}] \quad (11)$$

and the corresponding normalized prosodic vector is obtained as

$$\mathbf{Prosody}_{jk}^{(Nor)} = [\mathbf{Pth}_{jk}^{(Nor)}|\mathbf{Dur}_{jk}^{(Nor)}|\mathbf{Eng}_{jk}^{(Nor)}] \quad (12)$$

where $\mathbf{Pth}_{jk}^{(Nor)}$, $\mathbf{Dur}_{jk}^{(Nor)}$, and $\mathbf{Eng}_{jk}^{(Nor)}$ are the normalized vectors of respective prosodic parameters. The element inside each vector is given by

$$\tilde{p}_{jk}(s_i) = \left(p_{jk}(s_i) - m_k^{(p)}(s_i)\right)/\sigma_k^{(p)}(s_i), \quad 1 \leqslant i \leqslant 8 \quad (13)$$

$$\tilde{d}_{jk}(s_i) = \left(d_{jk}(s_i) - m_k^{(d)}(s_i)\right)/\sigma_k^{(d)}(s_i), \quad 1 \leqslant i \leqslant 8 \quad (14)$$

$$\tilde{e}_{jk}(s_i) = \left(e_{jk}(s_i) - m_k^{(e)}(s_i)\right)/\sigma_k^{(e)}(s_i), \quad 1 \leqslant i \leqslant 8 \quad (15)$$

where $\tilde{p}_{jk}(s_i)$, $\tilde{d}_{jk}(s_i)$, and $\tilde{e}_{jk}(s_i)$ represent the normalized elements of the pitch mean, duration, and energy respectively. $m_k(s_i)$ and $\sigma_k(s_i)$ are the mean and standard deviation in the ith HMM-state of the kth syllabic cluster, respectively.

As the last step, the prosodic 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 four experiments conducted in this paper. The first three are the consistency analyses concerning the duration, energy, and the pitch mean respectively, tested on a main-vowel syllable. The fourth is the consistency analysis concerning a prosodic unit made up of such three prosodic parameters together, tested on an *initial–final* syllable. 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 min.

4.1. Consistency analysis concerning the duration

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

Taking a further step to analyze the whole pronunciations with "i-2", meaning the syllable "i" with the second tone and a subset in the syllable "i", the possible paths and their probabilities for the segment sequences within the syllable "i-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, which its value is set from 1 to 4, represents a corresponding codeword in the codebook. There are 522 of the whole pronunciations with "i-2" tested in Table 2, and there is a total of 64 (4∗4∗4) combinations found in the seg-

Table 1
Codebooks of a duration pattern in the syllable "i" (Number of training data: 1988; codeword unit: 10 ms).

	Codewords in each codebook		
Segment #1	[1.625000	5.500000	1.897727]
	[1.345178	1.383249	6.007614]
	[1.259508	1.281879	1.626398]
	[5.676923	1.369231	1.938462]
Segment #2	[1.121429	5.964286]	
	[5.235294	3.588235]	
	[1.069463	1.225426]	
	[5.702703	1.175676]	
Segment #3	[7.878049	1.317073	2.109756]
	[1.516667	2.975000	2.350000]
	[1.236111	1.086806	4.100695]
	[1.295635	1.073413	1.267857]

Table 2
Path probability of a duration segment sequence within the syllable "i-2" (Number for statistic: 522).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0.003831
	Index2=2	0	0	0	0
	Index2=3	0.003831	0.011494	0.026820	0.057471
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0.003831	0.007663	0.007663
	Index2=2	0	0	0	0
	Index2=3	0.011494	0.022989	0.084291	0.222222
	Index2=4	0	0	0	0.003831
Index1=3	Index2=1	0	0.015326	0.003831	0.011494
	Index2=2	0	0	0	0
	Index2=3	0.022989	0.042146	0.164751	0.157088
	Index2=4	0	0	0	0
Index1=4	Index2=1	0	0	0	0.007663
	Index2=2	0	0	0	0
	Index2=3	0	0.007663	0.019157	0.080460
	Index2=4	0	0	0	0

Table 3
Path probability of a duration segment sequence within the syllable "— (i-2)" located in the word "一個 (i-2, k-ə-5)" (Number for statistic: 80).

		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	0.012500	0	0
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0.062500	0	0.112500
	Index2=4	0	0	0	0.025000
Index1=3	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0.675000	0.025000
	Index2=4	0	0	0	0
Index1=4	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0.012500	0.025000	0.050000
	Index2=4	0	0	0	0

ment 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=3, Index2=3, and Index3=3 as an example, it indicates that the duration vectors of all segments located in the third cluster respectively has 0.164751 of probability. It also means that all segments belong to lower 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.

Tabulated in Table 3 are the possible paths and associated probabilities for the segment sequence within the syllable "— (i-2)" located in the word "一個 (i-2, k-ə-5)". A total of 80 syllables are counted out of the speech database but merely 9 segment trees are found, which indicates a strongly non-uniform distribution among such probabilities. The largest probability is 0.675, meaning that the duration pattern for syllable "— (i-2)" embedded in the word "一個 (i-2, k-ə-5)" is consistent.

Moreover, tabulated in Table 4 are the possible paths and corresponding probabilities for the segment sequence within the syllable "— (i-2)" embedded into the word "一定 (i-2, t-iəŋ-4)". As little as 7 segment trees are found with the largest probability of 0.529412 among such segment trees. As before, it is also indicated that the duration pattern for the syllable "— (i-2)" located in the word "一定 (i-2, t-iəŋ-4)" is consistent.

As can be seen from Tables 3 and 4, there is a strong consistency between the duration pattern and the spectrum. It is validated as well that the same pronunciation in different word acquires a distinct duration warping curve.

In addition, a state diagram of the best path in relation to a duration pattern distributed is made in Fig. 2. There is

Table 4
Path probability of a duration segment sequence within the syllable "— (i-2)" located in the word "一定 (i-2, t-iəŋ-4)" (Number for statistic: 51).

		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	0.058824	0
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0.039216	0.019608	0.529412
	Index2=4	0	0	0	0
Index1=3	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0.058824	0.176471	0
	Index2=4	0	0	0	0
Index1=4	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0.117647
	Index2=4	0	0	0	0

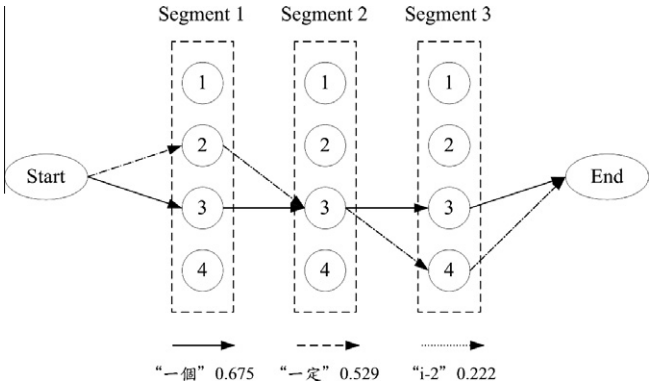

Fig. 2. A state diagram of the best path in relation to the duration pattern distributed.

a 0.675 probability that the best path of the syllable "— (i-2)" is embedded into the word "一個 (i-2, k-ə-5)", while a 0.529 probability that the best path of the syllable "— (i-2)" is into the word "一定 (i-2, t-iəŋ-4)", and a 0.222 probability for the best path in the whole syllable "i-2". A further insight into Fig. 2 and Table 1 reveals that the syllable "i-2" in the word "一個 (i-2, k-ə-5)" occurs with the maximum probability in the cases of Index1=3, Index2=3 and Index3=3. Indexing the corresponding codeword out of Table 1, it is found that the segments 1, 2 and 3 acquire durations of 4.167785 (1.259508 + 1.281879 + 1.626398), 2.294889 (1.069463 + 1.225426) and 6.423612 (1.236111 + 1.086806 + 4.100695), respectively. Likewise, for the syllable "i-2" embedded into the word "一定 (i-2, t-iəŋ-4)", the cases of Index1=2, Index2=3 and Index3=4 gain the maximum probability for indices 1–3, and it is as well seen that segments 1, 2 and 3 are of durations of 8.736041, 2.294889 and 3.636905, respectively. It is evident that the same syllable in different word acquires a distinct duration warping curve. These results indicate that the warping curve demonstrates influence not only on the global sentence, but also on the intra-syllable.

4.2. Consistency analysis concerning the energy

In this experiment, given the same syllable in the first experiment as an illustration, the consistency is analyzed between the energy and the spectrum. The trained VQ codebooks of energy for the syllable "i" are presented in Table 5.

Tabulated in Tables 6–8 are the path probabilities of a segment sequence concerning the energy pattern in the syllable "i-2", the word "一個 (i-2, k-ə-5)", and the word "一定 (i-2, t-iəŋ-4)", respectively. In Tables 7 and 8, similar to the experiment conducted on duration, it is confirmed, by two examples of the syllable "i-2" in the words "一個 (i-2, k-ə-5)" and "一定 (i-2, t-iəŋ-4)", that the distribution is concentrated under certain circumstances for these syllables embedded into the same word.

Moreover, Fig. 3 shows a state diagram of the best path in relation to an energy pattern distributed. There is a 0.438 probability that the best path of the syllable "— (i-2)" is found within the word "一個 (i-2, k-ə-5)", while a 0.412 probability that the best path of the syllable "— (i-2)" is within the word "一定 (i-2, t-iəŋ-4)", and a 0.084 probability for the best path in the whole syllable "i-2". There is a much higher probabilities that the best path lies in the words "一個 (i-2, k-ə-5)" and "一定 (i-2, t-iəŋ-4)" than there is for the whole syllable. Furthermore, a strong consistency of the energy pattern is validated by these experimental results.

An investigation into Fig. 3 and Table 5 reveals that various energy warping curves are seen due to the fact that identical syllable is located in different word. The syllable "i-2" in the word "一個 (i-2, k-ə-5)" occurs with the maximum probability in the cases of Index1=1, Index2=3 and

Table 5
Codebooks of an energy pattern in the syllable "i" (number of training data: 1988; codeword unit: dB).

	Codewords in each codebook		
Segment #1	[53.845585	56.043999	60.036594]
	[38.925594	58.789761	64.400459]
	[59.865810	61.541767	63.572266]
	[67.151085	66.578117	66.694588]
Segment #2	[62.739014	63.552902]	
	[70.213074	70.958199]	
	[58.356564	58.498104]	
	[66.400383	66.982262]	
Segment #3	[67.430618	66.422226	60.890232]
	[70.193787	70.183205	68.242798]
	[59.942188	56.840134	52.361439]
	[64.464203	61.912445	56.525749]

Table 6
Path probability of an energy segment sequence within the syllable "i-2" (Number for statistic: 522).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0.053640	0.019157	0.015326	0.068966
	Index2=2	0.007663	0.003831	0	0.003831
	Index2=3	0.011494	0	0.084291	0.068966
	Index2=4	0.022989	0.003831	0	0.019157
Index1=2	Index2=1	0.007663	0	0	0.022989
	Index2=2	0.003831	0	0	0.011494
	Index2=3	0	0	0	0.003831
	Index2=4	0.019157	0.015326	0	0.026820
Index1=3	Index2=1	0.026820	0.011494	0.019157	0.072797
	Index2=2	0.015326	0.007663	0	0
	Index2=3	0	0	0.038314	0.030651
	Index2=4	0.049808	0.019157	0	0.034483
Index1=4	Index2=1	0.011494	0	0	0.038314
	Index2=2	0.011494	0.011494	0	0.003831
	Index2=3	0	0	0	0
	Index2=4	0.061303	0.011494	0	0.030651

Table 7
Path probability of an energy segment sequence within the syllable "— (i-2)" located in the word "一個 (i-2, k-ə-5)" (Number for statistic: 80).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0.100000
	Index2=2	0.012500	0	0	0
	Index2=3	0	0	0.437500	0
	Index2=4	0	0	0	0
Index1=2	Index2=1	0	0	0	0
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0.025000
	Index2=4	0.050000	0	0	0
Index1=3	Index2=1	0	0	0.012500	0.137500
	Index2=2	0.012500	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0	0	0	0.050000
Index1=4	Index2=1	0	0	0	0.050000
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0
	Index2=4	0.075000	0	0	0.037500

Table 8
Path probability of an energy segment sequence within the syllable "一 (i-2)" located in the word "一定 (i-2, t-iəŋ-4)" (Number for statistic: 51).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0.411765
	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	0.019608
	Index2=3	0	0	0	0
	Index2=4	0.019608	0	0	0.078431
Index1=3	Index2=1	0	0	0	0.019608
	Index2=2	0	0	0	0
	Index2=3	0	0	0	0.039216
	Index2=4	0.176471	0	0	0.235294
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

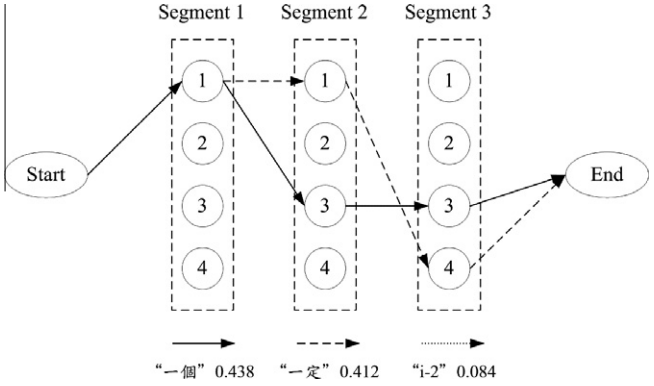

Fig. 3. A state diagram of the best path in relation to the energy pattern distributed.

Index3=3. Indexing the corresponding codeword out of Table 5, it is found that a sequence of energy level with 53.845585, 56.043999, 60.036594, 58.356564, 58.498104, 59.942188, 56.840134, 52.361439, is acquired along the state transition in the segments 1–3. Likewise, for the syllable "i-2" embedded into the word "一定 (i-2, t-iəŋ-4)", the cases of Index1=1, Index2=1 and Index3=4 gain the maximum probability for indices 1 to 3. In the second and the third segments, the energy of the syllable "i-2" in the word "一個 (i-2, k-ə-5)" is concentrated around the smallest energy level, nevertheless it is concentrated near the middle energy level for the word "一定 (i-2, t-iəŋ-4)".

4.3. Consistency analysis concerning the pitch mean

In this experiment, the same syllable is also given as an illustration to analyze the consistency between the pitch mean and the spectrum, with the trained VQ codebooks of the pitch mean for the syllable "i" tabulated in Table 9.

Tabulated in Tables 10–12 are the path probabilities of a segment sequence concerning the pitch mean in the syllable

Table 9
Codebooks of the pitch mean in the syllable "i" (Number of training data: 1988; codeword unit: 125μs).

	Codewords in each codebook		
Segment #1	[75.124023	74.340378	73.807777]
	[45.987888	44.819000	45.883999]
	[94.037048	90.888023	87.397331]
	[59.596931	58.124371	57.734974]
Segment #2	[45.197521	46.160461]	
	[89.140907	89.703789]	
	[60.733631	62.577209]	
	[74.791603	75.797157]	
Segment #3	[91.895409	96.579842	99.355911]
	[65.976860	67.098755	69.253876]
	[49.603638	50.573990	52.524113]
	[80.523560	81.733459	83.634819]

Table 10
Path probability of a segment sequence of the pitch mean within the syllable "i-2" (Number for statistic: 522).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0.015326	0
	Index2=2	0.019157	0	0	0.065134
	Index2=3	0	0.038314	0.022989	0
	Index2=4	0	0.122605	0.011494	0.072797
Index1=2	Index2=1	0	0	0.003831	0
	Index2=2	0	0	0	0
	Index2=3	0	0.007663	0	0
	Index2=4	0.003831	0	0	0
Index1=3	Index2=1	0	0	0.003831	0
	Index2=2	0.145594	0.011494	0	0.187739
	Index2=3	0	0.011494	0.011494	0
	Index2=4	0	0.118774	0.003831	0.022989
Index1=4	Index2=1	0	0	0.019157	0
	Index2=2	0.003831	0	0	0
	Index2=3	0	0.030651	0.026820	0
	Index2=4	0	0.007663	0	0.011494

Table 11
Path probability of a segment sequence of the pitch mean within the syllable "一 (i-2)" located in the word "一個 (i-2, k-ə-5)" (Number for statistic: 80).

		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	0.075000	0	0
	Index2=4	0	0	0.012500	0.125000
Index1=2	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=3	Index2=1	0	0	0	0
	Index2=2	0.212500	0	0	0.475000
	Index2=3	0	0	0.025000	0
	Index2=4	0	0.075000	0	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 12
Path probability of a segment sequence of the pitch mean within the syllable "— (i-2)" located in the word "一定 (i-2, t-iəŋ-4)" (Number for statistic: 51).

		Index3=1	Index3=2	Index3=3	Index3=4
Index1=1	Index2=1	0	0	0	0
	Index2=2	0	0	0	0.019608
	Index2=3	0	0.098039	0	0
	Index2=4	0	0.450980	0	0.058824
Index1=2	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=3	Index2=1	0	0	0	0
	Index2=2	0	0	0	0.215686
	Index2=3	0	0	0	0
	Index2=4	0	0.098039	0	0.058824
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

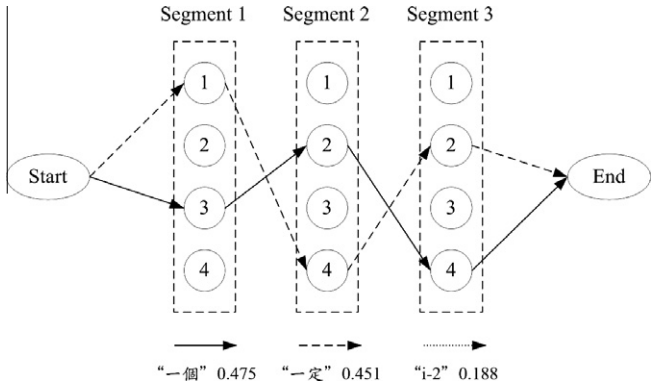

Fig. 4. A state diagram of the best path in relation to the pitch mean distributed.

"i-2", the word "一個 (i-2, k-ə-5)", and the word "一定 (i-2, t-iəŋ-4)", respectively. As such, a difference is seen in the consistency between the words "一個 (i-2, k-ə-5)" and "一定 (i-2, t-iəŋ-4)". The syllable "i-2" in the word "一個 (i-2, k-ə-5)" is of a higher pitch mean than in the word "一定 (i-2, t-iəŋ-4)".

Moreover, Fig. 4 shows a state diagram of the best path in relation to the pitch mean distributed. There is a 0.475 probability that the best path of the syllable "— (i-2)" is embedded into the word "一個 (i-2, k-ə-5)", while a 0.451 probability that the best path of the syllable "— (i-2)" is into the word "一定 (i-2, t-iəŋ-4)", and a 0.188 probability that the best path is into the whole syllable "i-2". The probabilities in the first two cases are larger than that in the case of the whole syllable.

In a brief conclusion, a verified consistency concerning individual prosodic pattern has been presented by the above experiments. The consistency with regard to a prosodic unit, consisting of the pitch mean, duration, and energy together, will be analyzed as follows.

4.4. Consistency analysis concerning the prosodic unit

In this experiment, three of the prosodic parameters, i.e. the pitch mean, duration, and energy, are merged as a prosodic vector, taken into account together, to inspect whether there is a maintained consistency property. This experiment is illustrated with an example of the syllable "tʂ-ə-4". Firstly, all the pronunciations "tʂ-ə" are used to compute the mean and the standard deviation (STD) of each prosodic parameter, as tabulated in Table 13. As well, the trained VQ codebooks are presented in tabular form as Table 14 in the case of the syllable "tʂ-ə".

To analyze the *final* part, listed in Table 15 are the path probabilities of a voiced segment concerning the prosodic unit in the syllable "tʂ-ə-4", the word "這件 (tʂ-ə-4, tɕ-ian-4)", and the word "這種 (tʂ-ə-4, tʂ-uəŋ-3)". In addition, Fig. 5 shows a state diagram of the best path in relation to the prosodic unit distributed. There is a 0.542 probability that the best path of the syllable "tʂ-ə-4" is found within the word "這件 (tʂ-ə-4, tɕ-ian-4)", while a 0.5 probability that the best path of the syllable "tʂ-ə-4" is within the word "這種 (tʂ-ə-4, tʂ-uəŋ-3)", and a 0.09 probability that the best path is in the whole syllable "tʂ-ə-4". A strong consistency of the prosodic unit is again verified.

Finally, tabulated in Table 16 are the de-normalized codewords of the best path in the above two words according to Eqs. (13)–(15), extracted out of Tables 13 and 14,

Table 13
The mean and the standard deviation of prosodic parameters in the syllable "tʂ-ə" (Number for statistic: 834).

Mean and standard deviation of pitch mean in each state (Unit: 125μs)								
Mean	[0.000000	0.000000	0.000000	51.382175	52.926708	55.920696	58.573143	61.066414]
STD	[0.000000	0.000000	0.000000	17.857290	19.299454	19.694798	20.135628	20.304157]
Mean and standard deviation of duration in each state (Unit: 10ms)								
Mean	[1.299760	1.482014	1.158273	2.364508	3.366906	2.676259	1.000000	2.189448]
STD	[0.478632	0.692743	0.364997	3.568647	2.480071	2.385656	0.000000	2.055602]
Mean and standard deviation of energy in each state (Unit: dB)								
Mean	[54.896751	63.965672	59.421242	70.084328	74.249954	72.205971	68.464668	63.169605]
STD	[7.106410	2.874451	4.300119	3.443483	4.690422	4.909353	5.836747	6.746165]

Table 14
Codebooks of a normalized prosodic vector in the syllable "tʂ-ə".

	Codewords in each codebook
Segment #1	[0.000000 0.000000 0.000000 1.538520 −0.400141 −0.433628 −0.058085 0.123182 −0.141406]
	[0.000000 0.000000 0.000000 0.418359 1.172301 −0.272466 −0.243677 −1.467166 −1.827192]
	[0.000000 0.000000 0.000000 −0.062509 −0.145886 2.306121 −0.115716 0.041156 0.289179]
	[0.000000 0.000000 0.000000 −0.626285 0.386848 −0.262394 −3.166461 −0.444208 0.061068]
	[0.000000 0.000000 0.000000 −0.550311 −0.485836 −0.433628 0.711330 1.296548 0.950531]
	[0.000000 0.000000 0.000000 −0.626285 −0.695805 −0.433628 0.514520 0.108994 −0.318271]
	[0.000000 0.000000 0.000000 −0.626285 −0.610891 −0.433628 −0.119011 −0.765980 0.889096]
	[0.000000 0.000000 0.000000 −0.539232 1.168763 −0.433628 0.220570 0.100607 −0.406527]
Segment #2	[1.497943 1.415676 −0.081762 −0.932376 −1.097859 −1.293700]
	[−0.173258 −0.210116 −0.239954 −0.789118 0.953364 0.343848]
	[0.271343 0.233706 −0.214229 −0.690936 −0.225886 −0.311445]
	[−0.873528 −0.828932 −0.373020 1.787487 0.086498 0.630363]
	[−0.849530 −0.819414 −0.368112 0.815672 1.242554 0.970202]
	[−0.818658 −0.804049 −0.374986 1.072312 −1.274835 0.539879]
	[1.776961 1.914288 3.020290 −0.919809 −0.648105 −1.833852]
	[−0.733208 −0.715715 −0.364847 0.639586 0.224896 0.641495]
Segment #3	[−0.778307 −0.741523 −0.672463 −0.309943 0.000000 −0.358443 −0.098302 −0.746812 −0.713775]
	[−0.887484 −0.953115 −1.004276 −0.597848 0.000000 −0.565835 1.025150 1.115948 1.206885]
	[1.953215 2.039181 1.984922 1.155688 0.000000 0.021349 −1.854792 −1.675594 −1.452381]
	[−0.561095 −0.445448 −0.434604 1.021500 0.000000 −0.266559 0.658215 0.572323 0.529459]
	[1.652502 1.663007 1.848112 −0.061554 0.000000 3.284551 −1.412478 −0.699854 −0.591626]
	[0.218668 0.189728 0.152287 −0.496491 0.000000 −0.187862 0.112070 0.420929 0.327146]
	[0.294115 0.209627 0.291964 −0.493055 0.000000 1.538962 0.190995 0.771639 0.520724]
	[1.198005 1.143977 1.040416 0.669942 0.000000 −0.120778 −0.728086 −0.665575 −0.706018]

Table 15
Path probability of a voiced segment concerning the prosodic unit in (a) the syllable "tʂ-ə-4", (b) the word "這件 (tʂ-ə-4, tɕ-ian-4)", and (c) the word "這種 (tʂ-ə-4, tʂ-uəŋ-3)" (Numbers for statistic are 600, 48, and 34 respectively).

	Index3=1	Index3=2	Index3=3	Index3=4	Index3=5	Index3=6	Index3=7	Index3=8
Table 15(a)								
Index2=1	0	0	0	0	0	0	0	0.003333
Index2=2	0.010000	0.026667	0	0.076667	0	0.050000	0.026667	0
Index2=3	0.010000	0	0	0.026667	0	0.083333	0.043333	0.010000
Index2=4	0.060000	0.023333	0	0.006667	0	0.006667	0	0
Index2=5	0.080000	0.090000	0	0.013333	0	0.010000	0.003333	0
Index2=6	0.066667	0.033333	0	0.020000	0	0	0.003333	0.003333
Index2=7	0	0	0.003333	0	0	0	0	0
Index2=8	0.086667	0.073333	0	0.030000	0	0.020000	0	0
Table 15(b)								
Index2=1	0	0	0	0	0	0	0	0
Index2=2	0	0	0	0.020833	0	0	0	0
Index2=3	0	0	0	0	0	0.020833	0	0
Index2=4	0.020833	0	0	0	0	0	0	0
Index2=5	0.062500	0.208333	0	0	0	0	0	0
Index2=6	0.083333	0	0	0	0	0	0	0
Index2=7	0	0	0	0	0	0	0	0
Index2=8	0.541667	0.041667	0	0	0	0	0	0
Table 15(c)								
Index2=1	0	0	0	0	0	0	0	0
Index2=2	0	0	0	0.029412	0	0.088235	0	0
Index2=3	0.176471	0	0	0	0	0.029412	0	0
Index2=4	0	0	0	0	0	0	0	0
Index2=5	0.500000	0.117647	0	0	0	0	0	0
Index2=6	0	0	0	0	0	0	0	0
Index2=7	0	0	0	0	0	0	0	0
Index2=8	0.058824	0	0	0	0	0	0	0

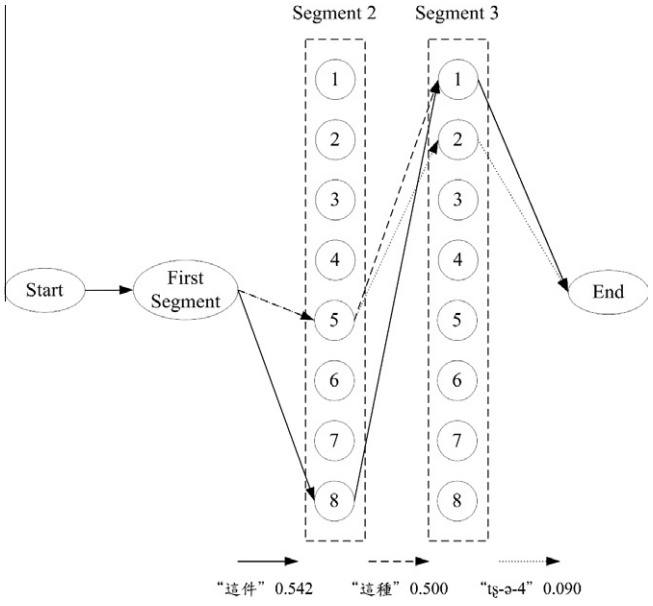

Fig. 5. A state diagram of the best path in relation to the prosodic unit distributed.

and then the prosodic unit can be recovered accordingly. In the second segment, the first path at Index2=8 and Index3=1 corresponds to a higher pitch mean and a lower energy level in comparison with the second path at Index2=5 and Index3=1, with a slight difference in the duration between both prosodic units.

5. System description of proposed Mandarin TTS and performance assessment

Presented in Fig. 6 is the system description of an acoustic module based Mandarin TTS. In the training part, the HMM decoding algorithm is firstly used to decode a state sequence within a syllable out of a speech database, which will be used in the synthesis unit generation and the prosody training mechanisms. The warping process between the spectrum and the prosody intra a syllable can be acquired in this phase.

In the synthesis unit generation, 411 syllabic waveforms are employed as synthesis units, each comprising the information on HMM-states. Accordingly, an intra-syllable

Table 16
Two de-normalized codebooks and corresponding recovered prosodic units for a voiced segment in the syllable "tʂ-ə".

The de-normalized codeword and its recovered prosodic unit at Index2=8 and Index3=1	
Segment #2	[38.289067 39.113799 1.062498 4.953125 70.858754 77.258836]
Segment #3	[40.592097 43.642112 47.412620 1.936842 1.000000 1.452632 71.723372 64.105715 58.354361]
Pitch mean (125 μs)	[38.289067 39.113799 40.592097 43.642112 47.412620]
Duration (10 ms)	[1.062498 4.953125 1.936842 1.000000 1.452632]
Energy (dB)	[70.858754 77.258836 71.723372 64.105715 58.354361]
The de-normalized codeword and its recovered prosodic unit at Index2=5 and Index3=1	
Segment #2	[36.211871 37.112465 1.050846 5.389830 74.363042 78.800611]
Segment #3	[40.592097 43.642112 47.412620 1.936842 1.000000 1.452632 71.723372 64.105715 58.354361]
Pitch mean (125 μs)	[36.211871 37.112465 40.592097 43.642112 47.412620]
Duration (10 ms)	[1.050846 5.389830 1.936842 1.000000 1.452632]
Energy (dB)	[74.363042 78.800611 71.723372 64.105715 58.354361]

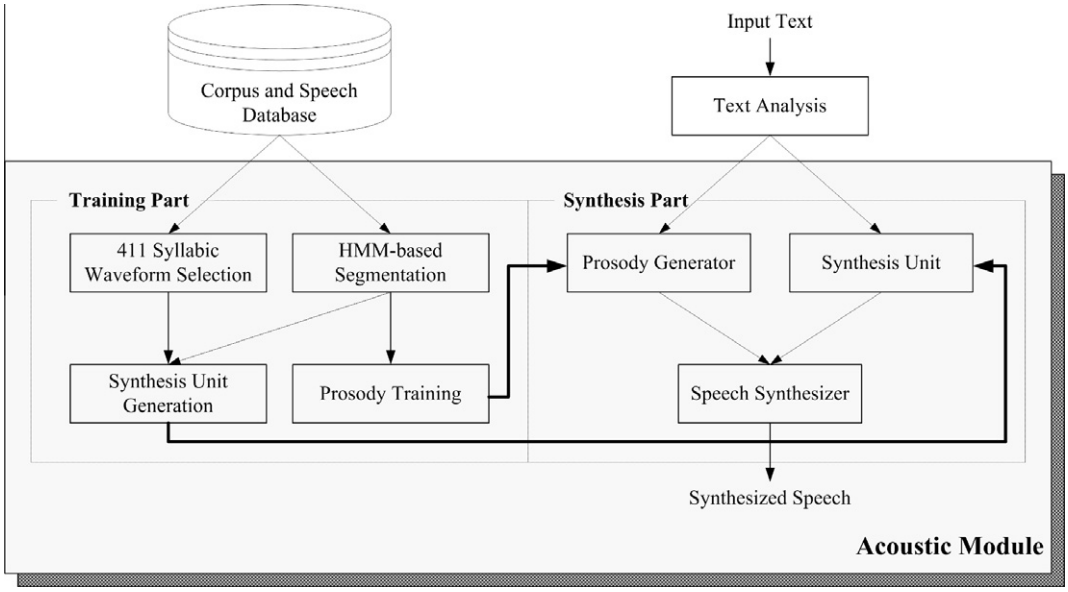

Fig. 6. System description of an acoustic module based Mandarin TTS system.

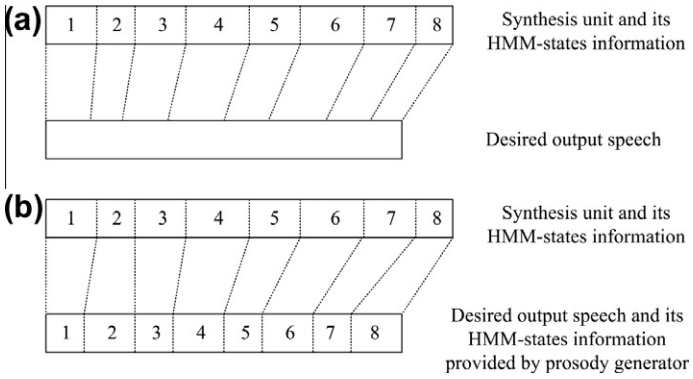

Fig. 7. Illustrations of (a) the regular PSOLA algorithm, and (b) a modified version with a warping process.

warping process is made in speech synthesis. In the prosody training mechanism, a recurrent neural network (RNN) based prosodic generator (Chen et al., 1998) and the error back propagation (EBP) algorithm are used to learn the relationship between the prosodic parameters and the linguistic features. As the target values of the prosodic generator, the output prosodic parameters include state durations, state energies and pitch contour of each syllable.

In the synthesis phase, a modified PSOLA algorithm is used as a speech synthesizer to make the prosodic modification imposed on a synthesis unit. Presented in Fig. 7 is an illustration of the warping process in the speech synthesizer. Not taking into account the warping process, presented in Fig. 7(a) is a regular PSOLA algorithm, where the syllabic duration of synthesis unit is modified and scaled in proportion to the desired output speech. However, in Fig. 7(b), the synthesis unit together with its HMM-states information was mapped into the corresponding HMM-state segment, which is provided by the prosody generator, and then the modified PSOLA algorithm is used to make the segment-based prosodic modification on the output speech. Thus, the warping process between the spectrum and the prosody intra a syllable can be achieved successfully.

In the end, our TTS proposal requires less than 2.4 MB of storage memory and 25 KB of runtime memory in a sampling rate of 16 KHz, 16-bit PCM waveform format. In the speech quality assessment, a subjective and forced-choice listening test was conducted to compare performance between the baseline and the proposed TTS systems. A set of 10 arbitrary sentences was selected and synthesized by both systems in each test group. Each pair of synthetic sentences was evaluated by 20 listeners. The preference score is evaluated for the two systems by each single listener in each group. It is experimentally demonstrated that the group of the proposed TTS was preferred to that of the baseline TTS in 76% of the test cases, while the other way around in the remaining 24%. Hence, the proposed TTS system is confirmed to exhibit a remarkable improvement in speech quality.

6. Conclusions

This paper is proposed mainly with a focus on the consistency analysis of an acoustic module for Mandarin speech. It is validated experimentally that the warping curve between the prosody 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 warping process of the spectrum and the prosody intra a syllable must be taken into account in a TTS system as a way to improve the synthesized speech quality.

Furthermore, this study on consistency analysis will be applied to the determination of a target cost criterion of a unit-selection based TTS with an expected performance improvement.

Acknowledgments

This research was financially supported by the Ministry of Economic Affairs under Grant No. 100-EC-17-A-03-S1-123 and the National Science Council under Grant No. NSC 95-2221-E-027-090, Taiwan, Republic of China.

References

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.

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.

Chen, 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.

Chou, F.C., Tseng, C.Y., 1998. Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties. Proc. ICASSP, 893–896.

Chou, F.C., Tseng, C.Y., Lee, L.S., 1996. Automatic generation of prosodic structure for high quality Mandarin speech synthesis. Proc ICSLP, 1624–1627.

Chou, F.C., Tseng, C.Y., Chen, K.J., Lee, L.S., 1997. A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units. Proc. ICASSP, 923–926.

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

Dey, S., Kedia, M., Basu, A., 2007. Architectural optimizations for text to speech synthesis in embedded systems. Proc. ASP-DAC, 298–303.

Guo, Q., Wang, B., Katae, N., 2008. Speech database compacted for an embedded Mandarin TTS system. Proc. ISCSLP, 1–4.

Huang, X.D., Acero, A., Hon, H.W., 2001. Hidden Markov Models. Spoken Language Processing. Prentice Hall PTR, New Jersey, pp. 377–413.

Hwang, S.H., Chen, S.H., Wang, Y.R., 1996. A Mandarin text-to-speech system. Proc. ICSLP, 1421–1424.

Hwang, S.H., Yeh, C.Y., 2005. The synthesis unit generation algorithm for Mandarin text to speech. GESTS Int. Trans. Speech Sci. Eng. 2 (1), 91–102.

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.

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

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.

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

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

O'Malley, M.H., 1990. Text-to-speech conversion technology. Computer 23 (8), 17–23.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Spelta, C., Manzoni, V., Corti, A., Goggi, A., Savaresi, S.M., 2010. Smartphone-based vehicle-to-driver/environment interaction system for motorcycles. IEEE Embed. Syst. Lett. 2 (2), 39–42.

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.

Yeh, C.Y., Chen, K.L., 2010. The research and implementation of acoustic module based Mandarin TTS. Proc ISCCSP, 1–4.

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

Ying, Z., Shi, X., 2001. An RNN-based algorithm to detect prosodic phrase for Chinese TTS. Proc ICASSP, 809–812.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proc ICASSP, 1315–1318.

Yue, D.J., 2010. Two stage concatenation speech synthesis for embedded devices. Proc. ICALIP, 1652–1656.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.

Zhu, Y., Zhao, L., Xu, Y., Niimi, Y., 2002. A Chinese text-to-speech system based on TD-PSOLA. Proc TENCON, 204–207.