EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A study on the consistency analysis of energy parameter for Mandarin speech

Li-Te Shen[1], Cheng-Yu Yeh[2*] and Shaw-Hwa Hwang[1]

Abstract

In this study, a consistency analysis of energy parameter for Mandarin speech is presented. Identified as a result of inspection of the human pronunciation process, 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 hidden Markov model (HMM) algorithm is used first to decode HMM-state sequences within a syllable at the same time as to divide them into three segments. Second, based on a designated syllable, the vector quantization (VQ) with the Linde–Buzo–Gray algorithm is used to train the VQ codebooks of each segment. Third, the energy 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 energy warping process intra a syllable must be considered in a text-to-speech 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 [1-6] 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 [7-10]. In an attempt to implement TTS technology on an embedded system, there are two additional requirements imposed on such integrated system, namely a low-memory requirement and a low-computational complexity.

Reviewing the history of TTS technology development, the waveform-based synthesis units approach [11-18] 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 [10-13] and the other is the small footprint synthesis units approaches [4,14-18]. 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, i.e., a large deal of storage space is needed to reach a superior speech quality.

In contrast, a footprint TTS adopts a small size synthesis unit, which treats a set of fundamental speech

* Correspondence: cy.yeh@ncut.edu.tw
[2]Department of Electrical Engineering, National Chin-Yi University of Technology, 57, Sec. 2, Zhongshan Rd., Taiping Dist, Taichung 41170, Taiwan
Full list of author information is available at the end of the article

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 algorithms [14,15]. 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 that the consistency between the prosody and the spectrum embedded in the running speech is existent.

In the case of verifying the consistency property, the definition of consistency will first be explained in this study. Subsequently, the consistency analysis between the energy 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 article.

The rest of this article is outlined as follows. The modeling of the consistency analysis in Mandarin speech is described in Section 2. Procedure of consistency analysis between the energy parameter of prosody and the spectrum is presented in Section 3. Experimental results are demonstrated and discussed in Section 4. Finally, this study ends with conclusion section.

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, is coupled 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 study, the consistency property between the energy 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. 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 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 energy intra a syllable is the one of interests. The warping curve within a syllable can be obtained by exploring the energy 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 [19]. For evaluation of the MFCCs, the discrete Fourier transform is first performed to obtain its spectrum

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

then, a filter bank with M filters according to Mel-scale is defined as

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

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

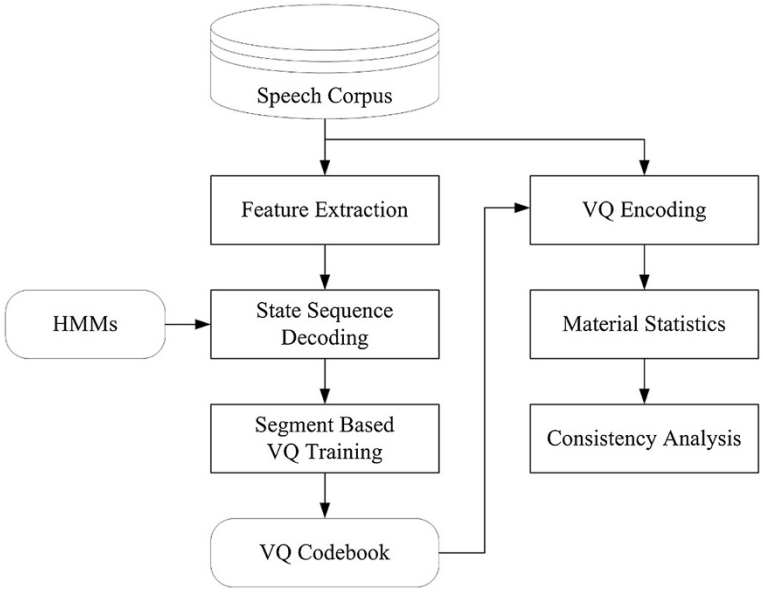

Figure 1 A flowchart of the procedure of consistency analysis.

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

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

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

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

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

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

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

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

3. Procedure of the consistency analysis

Figure 1 shows a flowchart of the procedure of consistency analysis. There are three steps required in the procedure. First, the feature extraction such as MFCCs, energy parameter, etc., are computed from a large speech database. Hence, the consistency analyses are made in the warping curve between the spectrum and the energy intra a syllable. Then, dividing them into three segments, the HMM decoding algorithm [19-23] 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.

Table 1 Codebooks of an energy pattern in the syllable "u" (number of training data: 556; codeword unit: dB)

	Codewords in each codebook
Segment #1	[50.621429 49.954460 49.260521]
	[55.270870 54.926376 54.276802]
	[42.995041 43.893917 46.077393]
	[61.238552 61.125961 60.872932]
Segment #2	[54.152412 54.538475]
	[48.779587 48.376892]
	[42.987476 42.105495]
	[61.043819 61.857407]
Segment #3	[43.920788 42.156853 37.482262]
	[50.758591 49.041313 43.996132]
	[56.621819 56.363434 51.977314]
	[63.510189 64.381699 62.834660]

Table 2 Possible paths and corresponding probabilities for a segment sequence within the syllable "u-4" (number for statistic: 134)

		Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index1 = 1	Index2 = 1	0	0.089552	0	0
	Index2 = 2	0.089552	0.134328	0	0
	Index2 = 3	0.014925	0	0	0
	Index2 = 4	0	0	0	0
Index1 = 2	Index2 = 1	0.029851	0.126866	0.186567	0
	Index2 = 2	0.029851	0.014925	0	0
	Index2 = 3	0.014925	0	0	0
	Index2 = 4	0	0	0.014925	0
Index1 = 3	Index2 = 1	0	0.014925	0.014925	0
	Index2 = 2	0.007463	0	0	0
	Index2 = 3	0.044776	0	0	0
	Index2 = 4	0	0	0	0
Index1 = 4	Index2 = 1	0.014925	0.014925	0.014925	0
	Index2 = 2	0	0.007463	0	0
	Index2 = 3	0	0	0	0
	Index2 = 4	0	0	0.059701	0.059701

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 12 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 [24] is used to train the VQ codebooks of each spectral segment with respect to the energy

Table 3 Possible paths and corresponding probabilities for a segment sequence within the syllable "務 (u-4)" located in the word "服務 (f-u-2, u-4)" (number for statistic: 25)

		Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index1 = 1	Index2 = 1	0	0.080000	0	0
	Index2 = 2	0	0.520000	0	0
	Index2 = 3	0	0	0	0
	Index2 = 4	0	0	0	0
Index1 = 2	Index2 = 1	0	0.120000	0.240000	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
	Index2 = 3	0	0	0	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
	Index2 = 4	0	0	0	0.040000

Table 4 Possible paths and corresponding probabilities for a segment sequence within the syllable "務 (u-4)" located in the word "業務 (iɛ-4, u-4)" (number for statistic: 15)

		Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
Index1 = 1	Index2 = 1	0	0.066667	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	0.600000	0.20000	0
	Index2 = 2	0.133333	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
	Index2 = 3	0	0	0	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
	Index2 = 4	0	0	0	0

vector. Thus, there is a total of three codebooks constructed in each syllable. In this article, 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 $\mathbf{Eng}_{jk}$ representing the energy vector of the jth pattern in the kth syllabic cluster is defined as

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

where $e_{jk}(s_i)$, $1 \le i \le 8$, is the value of energy 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 the N_k and $1 \le j \le N_k$.

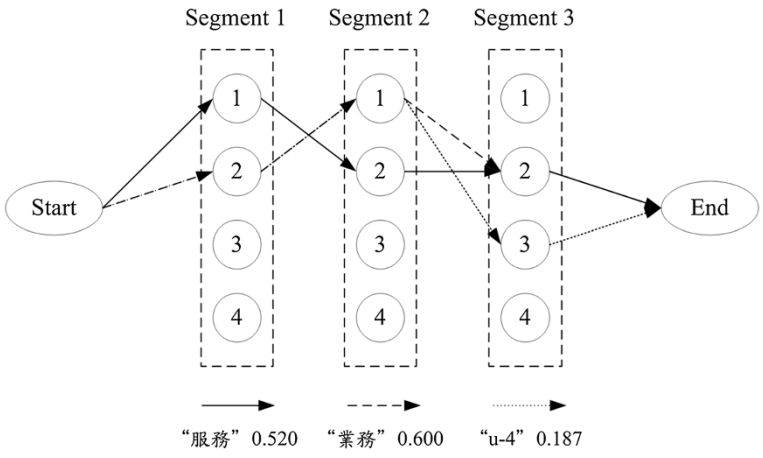

Figure 2 A state diagram of the best path in relation to the energy pattern distributed in the case of the syllable "u-4".

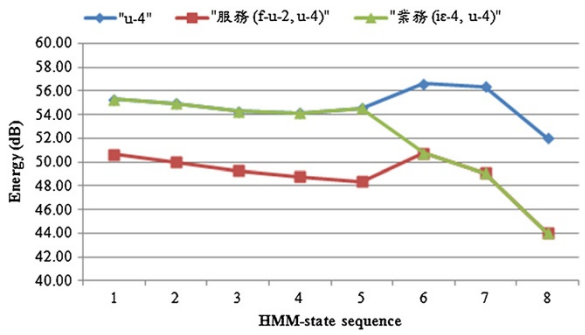

Figure 3 Energy warping curves for the best path within the whole syllable "u-4", the syllable "務 (u-4)" located in the word "服務 (f-u-2, u-4)", and the syllable "務 (u-4)" located in the word "業務 (iɛ-4, u-4)", respectively.

As the last step, the energy 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 article. The first consistency analysis is tested on a main-vowel syllable, i.e., the Mandarin syllable "ㄨ", which international phonetic alphabet (IPA) is labeled as "u". The second is tested on an *initial-final* syllable, i.e., the Mandarin syllable "ㄐ ㄧ", which IPA is labeled as "tɕ-i". All the experiments are conducted on a Chinese speech database with 8 kHz sampling frequency and 16-bit PCM format, containing 70,486 syllables out of 1,310 sentences by one male speaker, taking 297 MB of storage space, and a running time of 316 min.

Table 5 Codebooks of an energy pattern in the syllable "tɕ-i" (number of training data: 1116; codeword unit: dB)

	Codewords in each codebook
Segment #1	[41.460808 48.002613 46.519337]
	[42.038811 53.073967 53.203304]
	[50.116665 57.359455 55.549446]
	[47.017490 52.473804 49.475338]
Segment #2	[46.406963 46.657116]
	[57.987080 61.677597]
	[50.711330 52.214386]
	[53.727646 56.790955]
Segment #3	[47.587864 43.693424 38.622551]
	[56.583630 54.815712 50.843594]
	[62.393520 61.550232 58.198242]
	[54.124535 50.413986 43.915096]

Table 6 Path probability of a voiced segment concerning the energy pattern in (a) the syllable "tɕ-i-4", (b) the word "計畫 (tɕ-i-4, x-ua-4)", and (c) the word "技術 (tɕ-i-4, ʂ-u-4)" (numbers for statistic are 344, 28, and 22, respectively)

	Index3 = 1	Index3 = 2	Index3 = 3	Index3 = 4
(**a**)				
Index2 = 1	0.046512	0	0	0.005814
Index2 = 2	0.011628	0.075581	0.110465	0.081395
Index2 = 3	0.093023	0.081395	0	0.093023
Index2 = 4	0.075581	0.180233	0.046512	0.098837
(**b**)				
Index2 = 1	0	0	0	0
Index2 = 2	0	0	0	0
Index2 = 3	0	0.071429	0	0.535714
Index2 = 4	0	0.142857	0.214286	0.035714
(**c**)				
Index2 = 1	0	0	0	0
Index2 = 2	0	0	0.090909	0
Index2 = 3	0	0.090909	0	0.181818
Index2 = 4	0	0.636364	0	0

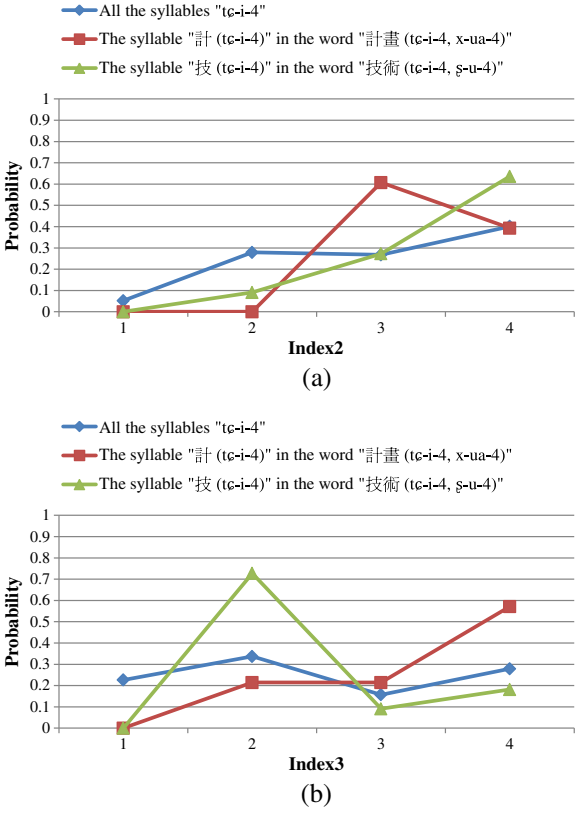

Figure 4 Probability distribution of the energy pattern. (**a**) The second segment. (**b**) Third segment within the syllable "tɕ-i-4".

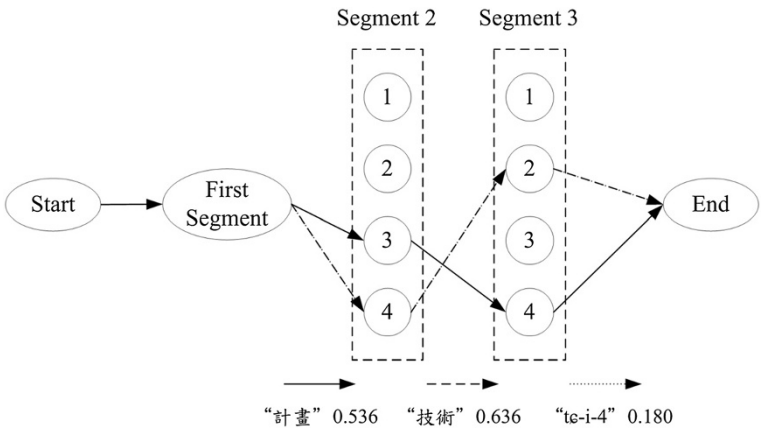

Figure 5 A state diagram of the best path in relation to the energy pattern distributed in the case of the syllable “tɕ-i-4”.

4.1. Consistency analysis for the case of Mandarin syllable “ㄨ”

Taking the Mandarin syllable “ㄨ (u)” as an example to analyze the consistency between the energy and the spectrum in this experiment, the trained VQ codebooks of energy for the syllable “ㄨ (u)” are tabulated in Table 1.

Taking a further step to analyze the whole pronunciation with “u-4”, meaning the syllable “u” with the fourth tone and a subset in the syllable “u”, the possible paths and their probabilities for the segment sequences within the syllable “u-4” 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, its value is set from 1 to 4, represents a corresponding codeword in the codebook. There are 134 of the whole pronunciations with “u-4” tested in Table 2, and there is 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 = 4, Index2 = 4, and Index3 = 4 as an example, it indicates that the energy vectors of all segments located in the second cluster, respectively, has 0.059701 of probability. It also means that all segments belong to the highest energies can be seen according to Table 1. Besides, the various path transitions within the syllable demonstrate the different energy contours in the same syllable.

Tabulated in Table 3 are the possible paths and associated probabilities for the segment sequence within the syllable “務 (u-4)” located in the word “服務 (f-u-2, u-4)”. A total of 25 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 0.52, meaning that the energy pattern for syllable “務 (u-4)” embedded in the word “服務 (f-u-2, u-4)” is consistent.

Moreover, tabulated in Table 4 are the possible paths and corresponding probabilities for the segment sequence within the syllable “務 (u-4)” embedded into the word “業務 (iɛ-4, u-4)”. As little as four paths are found with the largest probability of 0.6 among such paths. As before, it is also indicated that the energy pattern for the syllable “務 (u-4)” located in the word “業務 (iɛ-4, u-4)” is consistent.

In addition, a state diagram of the best path in relation to an energy pattern distributed is made in Figure 2. There is a 0.520 probability that the best path of the syllable “務 (u-4)” is found within the word “服務 (f-u-2, u-4)”, while a 0.600 probability that the best path of the syllable “務 (u-4)” is within the word “業務 (iɛ-4, u-4)”, and a 0.187 probability for the best path in the whole syllable “u-4”. There is a much higher probabilities that the best path lies in the words “服務 (f-u-2, u-4)” and “業務 (iɛ-4, u-4)” than there is for the whole syllable. A

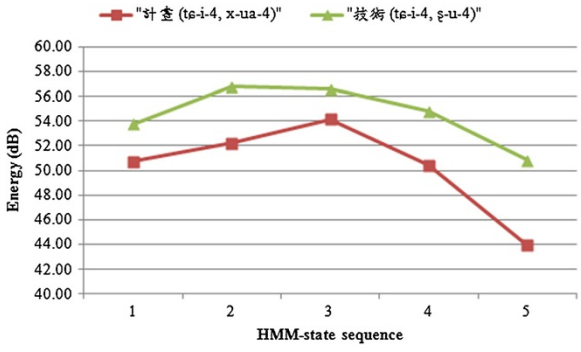

Figure 6 Energy warping curves for the best path within the syllable “計 (tɕ-i-4)” located in the word “計畫 (tɕ-i-4, x-ua-4)” and the syllable “技 (tɕ-i-4)” located in the word “技術 (tɕ-i-4, ʂ-u-4)”, respectively.

strong consistency between the energy pattern and the spectrum is validated by these experimental results.

Finally, presented in Figure 3 are energy warping curves for the best path within the whole syllable "u-4", the syllable "務 (u-4)" located in the word "服務 (f-u-2, u-4)", and the syllable "務 (u-4)" located in the word "業務 (iɛ-4, u-4)", respectively. Each warping curve in Figure 3 is obtained by an observation of Figure 2 and Table 1. It is evident that the same syllable in different word acquires a distinct energy warping curve. These results demonstrate that the influence of energy warping curve is not only on the global sentence, but also on the intra syllable.

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

Taking the Mandarin syllable "ㄐㄧ(tɕ-i)" as the second example to analyze the consistency between the energy and the spectrum in this experiment. The trained VQ codebooks of energy for the syllable "tɕ-i" are presented in Table 5. In this case, the waveform of syllable "ㄐㄧ (ɕ-i)" 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, listed in Table 6 are the path probabilities of a voiced segment concerning the energy pattern in (a) the whole syllable "tɕ-i-4", (b) the syllable "計 (tɕ-i-4)" located in the word "計畫 (tɕ-i-4, x-ua-4)", and (c) the syllable "技 (tɕ-i-4)" located in the word "技術 (tɕ-i-4, ʂ-u-4)". 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 4a is the probability distribution of the energy pattern in the second segment. Thus, each value in Figure 4a is obtained by the summation of the probabilities of each row in Table 6, i.e., the summation of the probabilities from Index3 = 1 to Index3 = 4. Similarly, each value in Figure 4b, meaning the probability distribution of the energy pattern in the third segment, is obtained by the summation of the probabilities from Index2 = 1 to Index2 = 4. As such, a difference in consistency is seen as before between the words "計畫 (tɕ-i-4, x-ua-4)" and "技術 (tɕ-i-4, ʂ-u-4)".

In addition, Figure 5 shows a state diagram of the best path in relation to the energy pattern distributed. There is a 0.536 probability that the best path of the syllable "tɕ-i-4" is found within the word "計畫 (tɕ-i-4, x-ua-4)", while a 0.636 probability that the best path of the syllable "tɕ-i-4" is within the word "技術 (tɕ-i-4, ʂ-u-4)", and a 0.180 probability that the best path is in the whole syllable "tɕ-i-4". A strong consistency of the energy pattern is verified by these experimental results.

Finally, presented in Figure 6 are energy warping curves for the best path within the syllable "計 (tɕ-i-4)" located in the word "計畫 (tɕ-i-4, x-ua-4)" and the syllable "技 (tɕ-i-4)" located in the word "技術 (tɕ-i-4, ʂ-u-4)", respectively. Each warping curve in Figure 6 is obtained by the observation of Figure 5 and Table 5. The warping curve for the best path within the whole syllable "tɕ-i-4" is identical to which within the word "技術 (tɕ-i-4, ʂ-u-4)". The same syllable in different word acquires a distinct energy warping curve is again verified.

5. Conclusions

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

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

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

Author details

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

Received: 20 September 2012 Accepted: 26 November 2012
Published: 17 December 2012

References

1. DH Klatt, Review of text-to-speech conversion for English. J. Acoust. Soc. Am. **82**, 737–793 (1987)
2. LS Lee, CY Tseng, OY Ming, The synthesis rules in a Chinese text-to-speech system. IEEE T. Acoust. Speech **37**, 1309–1320 (1989)
3. MH O'Malley, Text-to-speech conversion technology. Computer **23**, 17–23 (1990)
4. SH Hwang, SH Chen, YR Wang, A Mandarin text-to-speech system, in *Proceedings of the ICSLP*. Philadelphia, USA Vol. 3, 1421–1424 (1996)
5. W Mattheyses, L Latacz, W Verhelst, On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EJASMP, (2009). doi:10.1155/2009/169819
6. JD Edge, A Hilton, P Jackson, Model-based synthesis of visual speech movements from 3D video. EJASMP, (2009)
7. S Karabetsos, P Tsiakoulis, A Chalamandaris, S Raptis, Embedded unit selection text-to-speech synthesis for mobile devices. IEEE Trans. Consum. Electron. **55**, 613–621 (2009)
8. C Spelta, V Manzoni, A Corti, A Goggi, SM Savaresi, Smartphone-based vehicle-to-driver/environment interaction system for motorcycles. IEEE Embed. Syst. Lett. **2**, 39–42 (2010)
9. DJ Yue, Two stage concatenation speech synthesis for embedded devices, in *Proceedings of the ICALIP*. Shanghai, China Vol. 1, 1652–1656 (2010)
10. A Chalamandaris, S Karabetsos, P Tsiakoulis, S Raptis, A unit selection text-to-speech synthesis system optimized for use with screen readers. IEEE Trans. Consum. Electron. **56**, 1890–1897 (2010)

11. CH Wu, JH Chen, Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis. Speech Commun. **35**, 219–237 (2001)
12. FC Chou, CY Tseng, LS Lee, A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese. IEEE Trans. Speech Audio Process. **10**, 481–494 (2002)
13. JR Bellegarda, A Dynamic, Cost weighting framework for unit selection text-to-speech synthesis. IEEE Trans. Audio Speech Lang. Process. **18**, 1455–1463 (2010)
14. E Moulines, F Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. **9**, 453–467 (1990)
15. Y Zhu, L Zhao, Y Xu, Y Niimi, A Chinese text-to-speech system based on TD-PSOLA, in *Proceedings of the TENCON*. Beijing, China Vol. 1, 204–207 (2002)
16. SH Chen, SH Hwang, YR Wang, An RNN-based prosodic information synthesizer for Mandarin text-to-speech. IEEE Trans. Speech Audio Process. **6**, 226–239 (1998)
17. Z Ying, X Shi, An RNN-based algorithm to detect prosodic phrase for Chinese TTS, in *Proceedings of the ICASSP*. Salt Lake City, Utah, USA Vol. 2, 809–812 (2001)
18. CY Yeh, SH Hwang, Efficient text analyzer with prosody generator-driven approach for Mandarin text-to-speech. IEE Proc. Vis. Image Signal Process. **152**, 793–793 (2005)
19. XD Huang, A Acero, HW Hon, Hidden Markov models, in *Spoken Language Processing* (Prentice Hall PTR, NJ, 2001), pp. 377–413
20. LR Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
21. U Simsekli, A Jylha, C Erkut, T Cemgil, Real-time recognition of percussive sounds by a model-based method. EURASIP J. Adv. Signal Process (2011). doi:10.1155/2011/291860
22. S Winters-Hilt, Z Jiang, C Baribault, Hidden Markov model with duration side information for novel HMMD derivation, with application to eukaryotic gene finding. EURASIP J. Adv. Signal Process (2010). doi:10.1155/2010/761360
23. H Zen, K Tokuda, AW Black, Statistical parametric speech synthesis. Speech Commun. **51**, 1039–1064 (2009)
24. Y Linde, A Buzo, R Gray, An algorithm for vector quantizer design. IEEE Trans. Commun. **28**, 84–95 (1980)

doi:10.1186/1687-4722-2012-28
Cite this article as: Shen *et al.*: A study on the consistency analysis of energy parameter for Mandarin speech. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:28.