A New Duration Modeling Approach for Mandarin Speech

Sin-Horng Chen, *Senior Member, IEEE*, Wen-Hsing Lai, and Yih-Ru Wang

***Abstract*—In this paper, a new duration modeling approach for Mandarin speech is proposed. It explicitly takes several major affecting factors as multiplicative companding factors (CFs) and estimates all model parameters by an EM algorithm. Besides, the three basic Tone 3 patterns (i.e., full tone, half tone and sandhi tone) are also properly considered via using three different CFs to separate their affections on syllable duration. Experimental results showed that the variance of the syllable duration was greatly reduced from 180.17 to 2.52 frame^2 (1 frame $=$5 ms) by the syllable duration modeling to eliminate effects from those affecting factors. Moreover, the estimated CFs of those affecting factors agreed well to our prior linguistic knowledge. Two extensions of the duration modeling method are also performed. One is the use of the same technique to model initial and final durations. The other is to replace the multiplicative model with an additive one. Lastly, a preliminary study of applying the proposed model to predict syllable duration for TTS is also performed. Experimental results showed that it outperformed the conventional regressive prediction method.**

***Index Terms*—Duration modeling, Mandarin, text-to-speech.**

I. Introduction

PROSODY refers to aspects of the speech signal other than the actual words spoken, such as timing and fundamental frequency pattern, and plays an important role in the disambiguation of discourse structure. Speakers use prosody to convey emphasis, intent, attitude, and to provide cues to aid listeners in the interpretation of their speech. Researchers have noticed that fluent spoken speech is not produced in a smooth, unvarying stream. Rather, speech has perceptible breaks, relatively stronger and weaker, as well as longer and shorter syllables. Without prosody, speech would be flat and toneless and would sound tedious, unpleasant, or even barely intelligible. Although it is known that prosody can be affected by many factors such as sentence type, syntactical structure, semantics and emotional state of speaker, the relationships between prosodic features and those affecting factors are not totally understood. Indeed, the lack of a general consensus in these areas is the main reason why prosody was still under-utilized in spoken language processing nowadays. Prosodic modeling is therefore important and urgent in speech processing.

In this paper, we concentrate our study on one important issue of prosodic modeling—duration modeling. Duration modeling is important in both automatic speech recognition (ASR) [1]–[4] and text-to-speech (TTS) [5]–[10]. In ASR, state duration models are usually constructed to assist in the HMM-based speech recognition. In TTS, synthesis of proper duration information is essential for generating a highly natural synthetic speech. A precise duration model is surely helpful for improving the performance of ASR as well as for making synthesized speech more natural in TTS.

In the past, durational characteristics of speech in various languages have been the subject of many recent researches. Many factors have been shown to have major effects on influencing segmental duration [11], [12]. The general goal of duration modeling is to find a computational relation between a set of affecting factors and the segment duration. Related literatures concerned with the finding of perceptual cues, with the development of duration-generating rules for synthesizing intelligible and natural-sounding speech, and with the automatic duration analysis for speech recognition, speech understanding and word finding. They can be categorized into two approaches: rule-based and data-driven. The former is a conventional one which uses linguistic expertise to manually infer some phonologic rules of duration generation based on observations on a large set of utterances. A prevalent method of the approach for TTS uses sequential rules to initially assign the duration of a segment with an intrinsic value, and then successively applies rules to modify it [5], [6]. Three main disadvantages of the approach can be found. First, manually exploring the effect of mutual interactions among linguistic features of different levels is highly complex. Second, the rule-inference process usually involves a controlled experiment, in which only a limited number of contextual factors is examined. The resulting inferred rules may therefore not be general enough for unlimited texts. Third, the rule-inference process is cumbersome. As a result, it is generally very difficult to collect enough rules without long-term devotion to the task.

The data-driven approach tries to construct a duration model from a large speech corpus, usually with the aid of statistical methods or artificial neural network (ANN) techniques. It first designs a computational model to describe the relationship between the segment duration and some affecting factors, and then trains the model on the speech corpus. The training goal is to automatically deduct phonologic rules from the speech corpus and implicitly memorize them in the model's parameters or ANNs weights. The primary advantage of this approach is that the rules can be automatically established from the training data set during the training process without the help of linguistic experts. Two popular methods of the approach are the decision tree-based [13] and multilayer perceptrons (MLP)-based methods [10]. The former uses a decision tree to

Manuscript received December 14, 2001; revised August 16, 2002. This work was supported by the NSC of Taiwan under Contract NSC89-2213-E-009-187. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng.

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: lwh@cht.com.tw).

Digital Object Identifier 10.1109/TSA.2003.814377

classify the segmental durations into some clusters according to their relationships with some linguistic features. The latter uses an MLP to learn the mapping between the segmental duration and some linguistic features. Combining these two methods by cascading regression tree and neural network was also proposed [14]. Criticisms raised against these two methods are the insufficient accuracy of duration prediction in the case of the decision tree-based method, and the difficulty in interpreting the hidden structure of the model learned in the MLP-based method.

Another popular method of the data-driven approach uses regression analysis to model the relationship between the segmental duration and some linguistic features. Regression methods in the linear or logarithmic domain [15], [16], or based on a sigmoid transformation function [17] can be used. The sums-of-products (SOP) method is based on multiple linear regression and is supervised on the basis of linguistic knowledge [18]–[20]. A piecewise linear transformation in the SOP method was proposed to expand the durations at two ends of duration range [21]. In [22], a computation-intensive algorithm, called "Multivariate Adaptive Regression Splines" (MARS), was used to estimate general functions of high-dimensional arguments given sparse data.

For the duration modeling in Mandarin speech, there are still quite few studies. The lack of an appropriate prosodic model is the major problem encountered. In this paper, we propose a statistical model for Mandarin syllable duration via considering some major affecting factors. Our goal is to deconfound the effects of these affecting factors so as to better understanding the mechanism of syllable duration generation in Mandarin speech. The affecting factors we considered include speaker-level speaking rate, utterance-level speaking rate, syllabic tone, base-syllable, and prosodic state. Here prosodic state is conceptually defined as the state in a prosodic phrase and is introduced to account for all other affecting factors, that are not covered by the former four factors, including word-level and syntactic-level linguistic cues.

We will also consider the affecting factor of Tone 3 in more detail. In natural Mandarin speech, Tone 3 can be pronounced in three basic patterns: falling-rising, low-falling, and middle-rising. The first one is a full tone and usually appears at the end of a word or a prosodic phrase. It usually pronounced slightly longer than the other three regular tones (i.e., Tone 1, Tone 2, and Tone 4) [23]. The second one is a half tone and is always pronounced shorter [23]. The third one sounds like Tone 2 and is resulted from the well-known sandhi rules of changing a Tone 3 to Tone 2 when it precedes another Tone 3. Due to the fact that these three Tone 3 patterns are different in their durations, we will consider them as three different tone affecting factors.

The paper is organized as follows. Section II discusses the proposed multiplicative syllable duration model of Mandarin speech in details. Section III describes the experimental results of the syllable duration modeling study on two databases. An extension to include the modelings of initial and final durations is also made. Detailed analyses of the influences of these five affecting factors are given in Section IV. In Section V, an additive duration model is introduced and compared with the multiplicative model. An application of using these two models in syllable duration prediction for Mandarin TTS is presented in Section VI. Conclusions and future works are given in Section VII.

II. Proposed Syllable Duration Model

In duration modeling, the desired modeling units can be speech segments like HMM states, phones, initials, finals, syllables or even words. Since Mandarin is a tonal and syllable-based language, syllable is the basic pronunciation unit. We therefore choose syllable duration as the modeling unit to start our study. An extension to additionally model initial and final durations is included in Section III. The proposed syllable duration model is designed based on the idea of taking each affecting factor as a multiplicative companding (compressing-expanding) factor (CF) to control the compression and stretch of the syllable duration. A parallel modeling approach using additive CF will also be discussed in Section V. In the following, we discuss the affecting factors used, the multiplicative syllable duration model, and the method of estimating the model parameters in detail.

A. Affecting Factors

In naturally spoken Mandarin Chinese, syllable duration varies considerably depending on various linguistic and non-linguistic factors. In this study, five major affecting factors including tone, base-syllable, speaker-level speaking rate, utterance-level speaking rate, and prosodic state are considered. In the following, we discuss their effects on influencing syllable duration.

Mandarin Chinese is a tonal and syllable-based language. Syllable is the basic pronunciation unit. Each character is pronounced as a syllable. There exist only about 1300 phonetically distinguishable syllables comprising the set of all legal combinations of 411 base-syllables and five tones. Tonality of a syllable is characterized by its pitch contour shape, loudness and duration. This means the tone of a syllable will affect its duration. We therefore consider tone as an affecting factor. One obvious phenomenon to show the affection of tone on syllable duration is that syllables with Tone 5 are always pronounced much shorter. Besides, Tone 3 can be pronounced in three basic patterns of falling-rising (full tone), low-falling (half tone), and middle-rising (sandhi tone) which are of different durations.

Syllable duration is also seriously affected by the phonetic structure of base-syllables. Mandarin base-syllables have very regular phonetic structure. Each base-syllable is composed of an optional consonant initial and a final. The final can be further decomposed into an optional medial, a vowel nucleus, and an optional nasal ending. So base-syllables comprise one to four phonemes. Generally speaking, syllable duration increases as the number of constituent phonemes increases. Syllables with single vowels are shortest. Syllables with stop initials or no initials, and without nasal endings are pronounced shorter. Syllables with fricative initials and with nasal endings are longer.

Aside from syllable-based features, other high-level linguistic features, such as word-level and syntactic-level features, can also affect the syllable duration seriously. In this study, we use prosodic state to account for the influences of high-level linguistic features. Here, prosodic state is conceptually defined as the state in a prosodic phrase. Syllable duration varies in

different part of a prosodic phrase. The lengthening effect for the last syllable of a prosodic phrase is a well-known example. The reasons of using prosodic state to replace high-level linguistic features are three-folded. Firstly, durational information is a kind of prosodic feature so that syllable duration should match better to the prosodic phrase structure than to the syntactic phrase structure. Secondly, it is still difficult to automatically extract syntactic-level linguistic cues from unlimited natural Chinese texts. Thirdly, by this approach we can isolate the duration modeling study from the difficult task of syntactic analysis. The main disadvantage of using prosodic state is the lack of large speech corpora with prosodic tags being properly labeled. So we must treat the prosodic state of a syllable as hidden or unknown. Fortunately, we can solve the problem by the expectation-maximization (EM) algorithm which is a general methodology for maximum likelihood (ML) or maximum a posteriori (MAP) estimation from incomplete data. Substantial literatures have been devoted to the study of the EM algorithm and found that it is generally a first-order or linearly convergent algorithm [24]–[26]. A by-product of the approach is the automatic determination of prosodic states for all syllables in the training set. This is an additional advantage because prosodic labeling has become an interesting research topic recently [27]–[34] and such kind of prosodic information provides cues for resolving syntactic ambiguity in automatic speech understanding [29]–[32] and for improving the naturalness of TTS [33], [34].

Syllable duration will also be affected by the speaking rate. Natural speech is not always spoken in a constant speed. A speaker usually speaks using his familiar speed. But he can change the speed from time to time. In order to account for this impact upon the statistical syllable duration model, we consider two types of speaking rates, speaker-level and utterance-level, in this study.

B. Syllable Duration Model

The model is constructed based on the assumption that the effects of all affecting factors are combined multiplicatively [19] and is expressed by

$$Z_n = X_n \gamma_{t_n} \gamma_{y_n} \gamma_{j_n} \gamma_{l_n} \gamma_{s_n} \tag{1}$$

where Z_n and X_n are, respectively, the observed and normalized durations of the nth syllable; γ_p is the CF of the affecting factor p; t_n, y_n, j_n, l_n and s_n represent respectively the lexical tone, prosodic state, base-syllable, utterance, and speaker of the nth syllable; and X_n is modeled as a normal distribution with mean μ and variance v. Notice that prosodic state represents the state in a prosodic phrase and is treated as hidden. In Mandarin, there are 5 lexical tones and 411 base-syllables. If considering the three Tone 3 patterns of falling-rising, middle-rising and low-falling, we increase the number of tones to 7.

C. Training of the Model

To estimate the parameters of the model, an EM algorithm [35] based on the ML criterion is adopted. The EM algorithm is derived based on incomplete training data with prosodic state being treated as hidden or unknown. In the following, we discuss it in more detail.

To illustrate the EM algorithm, an auxiliary function is firstly defined in the expectation step (E-step) as

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

where N is the total number of training samples, Y is the total number of prosodic states, $p(y_n|Z_n, \bar{\lambda})$ and $p(Z_n, y_n|\lambda)$ are conditional probabilities, $\lambda = \{\mu, v, \gamma_t, \gamma_y, \gamma_j, \gamma_l, \gamma_s\}$ is the set of parameters to be estimated, and λ and $\bar{\lambda}$ are, respectively, the new and old parameter sets. Based on the assumption that the normalized duration X_n is normally distributed, $p(Z_n, y_n|\lambda)$ can be derived from the assumed model given in (1) and expressed by

$$p(Z_n, y_n|\lambda) = N\left(Z_n; \mu\gamma_{t_n}\gamma_{y_n}\gamma_{j_n}\gamma_{l_n}\gamma_{s_n}, v\gamma_{t_n}^2\gamma_{y_n}^2\gamma_{j_n}^2\gamma_{l_n}^2\gamma_{s_n}^2\right) \tag{3}$$

where $N(Z; \mu, v)$ denotes a normal distribution of Z with mean μ and variance v. Similarly, $p(y_n|Z_n, \bar{\lambda})$ can be expressed by

$$p(y_n|Z_n, \bar{\lambda}) = \frac{p(Z_n, y_n|\bar{\lambda})}{\sum_{y_n'=1}^{Y} p(Z_n, y_n'|\bar{\lambda})} \tag{4}$$

Then, sequential optimizations of these parameters can be performed in the maximization step (M-step).

A drawback of the above EM algorithm is that it may result in a nonunique solution because of the use of multiplicative affecting factors. This is obvious because, if we scale up an affecting factor and scale down another by the same value, the same objective value will be reached. To cure the drawback, we modify each optimization procedure in the M-step to a constrained optimization one via introducing a global duration constraint. The auxiliary function then changes to

$$Q(\bar{\lambda}, \lambda) = \sum_{n=1}^{N} \sum_{y_n=1}^{Y} p(y_n|Z_n, \bar{\lambda}) \log p(Z_n, y_n|\lambda) + \eta\left(\sum_{n=1}^{N} \mu\gamma_{t_n}\gamma_{y_n}\gamma_{j_n}\gamma_{l_n}\gamma_{s_n} - N\mu_z\right) \tag{5}$$

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

To execute the EM algorithm, initializations of the parameter set λ are needed. This can be done by estimating each individual parameter independently. Specifically, the initial CF for a specific value of an affecting factor is assigned to be the ratio of the duration mean of syllables with the affecting factor equaling the value to the duration mean of all syllables. Notice that, in the initializations of CFs for prosodic state, each syllable is pre-assigned a prosodic state by quantizing its duration. After initialization, all parameters are sequentially updated in each iterative step. Iterations are continued until a convergence is reached. The prosodic state can finally be assigned by

$$y_n^* = \max_{y_n} p(y_n|Z_n, \lambda). \tag{6}$$

The procedure of the EM algorithm is summarized below.

1) Compute initial values of λ by independently estimate each individual parameter from the training set.
2) Do for each iteration k:
   a) Update $\bar{\lambda} = \lambda$.
   b) E-step: Use **(3)–(5)** to calculate $Q(\bar{\lambda}, \lambda)$.
   c) M-step: Find optimal λ by
$$\lambda = \max_{\lambda} Q(\bar{\lambda}, \lambda). \tag{7}$$
   d) Termination test: If $L(k) - L(k-1) < \varepsilon$ or $k \geq K$ stop, where
$$L(k) = \sum_{n=1}^{N} \log p(Z_n|\lambda) \tag{8}$$
   is the total log-likelihood for iteration k and K is the maximum number of iterations.
3) Assign prosodic states by using **(6)**.

D. Modeling of Tone 3

We now extend the model to additionally consider the affections of the three Tone 3 patterns of falling-rising, middle-rising and low-falling. In this study, these three patterns are simply denoted as Tone 3_f, Tone 3_s, and Tone 3_h, and three separate CFs are used to account their affections on the syllable duration. The EM algorithm is then modified for parameter estimation. In initialization, we first assign all lexical Tone 3 to Tone 3_s when they precedes other lexical Tone 3, and then use VQ to divide all others lexical Tone 3 into two clusters of Tone 3_f and Tone 3_h. Besides, at the end of each iteration, syllables with lexical Tone 3 are re-assigned to one of these three patterns by

$$t_n^* = \arg\max_{t_n} p(t_n|Z_n, \lambda) \tag{9}$$

for $t_n \in \{3_f, 3_s, 3_h\}$, where $p(t_n|Z_n, \lambda)$ is the conditional probability of a Tone 3 pattern.

E. Testing of the Model

Although we have gotten CFs of all affecting factors from the above training procedure, some information is still not known in the testing phase. It includes the CFs of both speaker- and utterance-level speaking rates, the prosodic state of each syllable, and the tone pattern of a syllable with Tone 3. The following testing procedure is therefore set to estimate these unknown parameters:

1) Initialization:
   a) Fix the CFs for tone, base-syllable and prosodic state, and the mean and variance of the normalized syllable duration to the trained values and form a parameter set $\bar{\lambda}_1 \{\bar{\mu}, \bar{\nu}, \bar{\gamma}_t, \bar{\gamma}_y, \bar{\gamma}_j\}$
   b) Compute the initial values of the parameter set, $\lambda_2 = \{\gamma_l, \gamma_s\}$, needed to be determined in Step 2.
2) Do for each iteration k:
   a) Update $\bar{\lambda}_2 = \lambda_2$
   b) E-step: calculate
$$Q(\bar{\lambda}_2, \lambda_2) = \sum_{n=1}^{N} \sum_{y_n=1}^{Y} p(y_n|Z_n, \bar{\lambda}_1, \bar{\lambda}_2) \log p(Z_n, y_n|\bar{\lambda}_1, \lambda_2) \tag{10}$$
   c) M-step: Find optimal λ_2 by
$$\lambda_2 = \max_{\lambda_2} Q(\bar{\lambda}_2, \lambda_2) \tag{11}$$
   d) Termination test: If $L(k) - L(k-1) < \varepsilon$ or $k \geq K$ stop, where
$$L(k) = \sum_{n=1}^{N} \log p(Z_n|\bar{\lambda}_1, \lambda_2) \tag{12}$$
   is the total log-likelihood for iteration k and K is the maximum number of iterations.
3) Assign prosodic state and Tone 3 pattern by
$$y_n^* = \max_{y_n} p\left(y_n|Z_n, \bar{\lambda}_1, \lambda_2\right) \tag{13}$$
$$t_n^* = \arg\max_{t_n} p\left(t_n|Z_n, \bar{\lambda}_1, \bar{\lambda}_2\right) \tag{14}$$
   for $t_n \in \{3_f, 3_s, 3_h\}$.

After performing the above procedure, we can get the CFs of each testing utterance and speaker, the prosodic state of each syllable, the tone pattern of each syllable with Tone 3, and the normalized duration of each syllable.

III. Experimental Results

A. Databases

Effectiveness of the proposed syllable duration modeling method was examined by simulations on two databases. The first one is a high-quality, reading style microphone-speech database recorded in a sound-proof booth. It is referred to as the MIC database. It was generated by five native Chinese speakers including two males and three females. Among these five speakers, two of them were professional radio announcers. The database consisted of two data sets. One (MIC-sent) contained sentential utterances with texts belonging to a well-designed, phonetic-balanced corpus of 455 sentences. The lengths of these sentences ranged from 3 to 75 syllables with an average of 13 syllables. The other (MIC-para) contained paragraphic utterances with texts belonging to a corpus of 300 paragraphs which covered a wide range of topics including news, primary school textbooks, literature, essays, etc. The lengths of these paragraphs ranged from 24 to 529 syllables with an average of 170 syllables. The MIC database was divided into two parts: a training set and a test set. Table I shows some statistics of the MIC database. The training set contained, in total, 98 620 syllables and the test set contained 20 717 syllables. The mean (unit in frame) and variance (unit in frame^2) of syllable duration for the training and test sets are shown in Table II(a). Here one frame equals 5 ms.

TABLE I
STATISTICS OF THE MIC SPEECH CORPUS

Data Set	Speaker	Sentence	Paragraph	Syllable
Training	Male A	1-455	1-200	33404
Training	Female B	1-455	1-50	12619
Training	Male C	1-455	1-100	19502
Training	Female D	1-455	1-200	33095
Testing	Female E	None	200-300	20717

TABLE II
STATISTICS OF (A) THE OBSERVED DURATIONS IN THE MIC DATABASE, AND THE NORMALIZED DURATIONS OBTAINED IN (B) THE MULTIPLICATIVE AND (C) ADDITIVE MODELS WITH 16 PROSODIC STATES. (UNITS: MEAN AND RMSE IN FRAME AND VARIANCE IN frame^2; 1 FRAME $=5$ ms)

	Training set		Testing set	
	mean	Variance	mean	variance
Syllable	44.31	180.17	41.08	136.26
Initial	17.21	62.28	13.83	40.02
Final	31.75	117.06	30.94	104.15

(a)

	Training set			Testing set		
	mean	variance	RMSE	mean	variance	RMSE
Syllable	42.34	2.52	1.93	44.77	4.44	3.41
Initial	16.63	0.74	0.97	18.36	5.92	2.27
Final	31.50	2.12	1.66	33.90	3.40	3.25

(b)

	Training set			Testing set		
	mean	variance	RMSE	mean	variance	RMSE
Syllable	43.89	2.53	1.59	43.77	3.97	1.99
Initial	17.20	0.78	0.89	17.05	1.73	1.32
Final	31.44	1.84	1.36	31.38	2.85	1.69

(c)

After recording, all speech signals of the MIC database were converted into 16-bit data at 20-kHz sampling rate. They were then manually segmented into sub-syllables of initial and final. The associated texts were transcribed automatically by a linguistic processor with an 80 000-word lexicon. All transcription errors were manually corrected.

The second database is a 500-speaker, telephone-speech data set which is a subset of MAT-2000 provided by the Association of Computational Linguistics and Chinese Language Processing. It is referred to as the TEL database. It was collected from calls through public telephone networks in Taiwan. All speech signals were digitally recorded with a form of 8-kHz, 16-bit data. It contained phonetically balanced, short sentential utterances in reading style. All speech signals were automatically segmented using a set of 100-initial and 39-final HMM models.

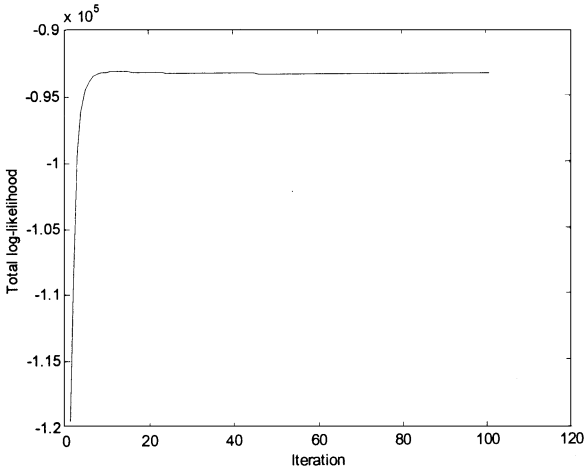
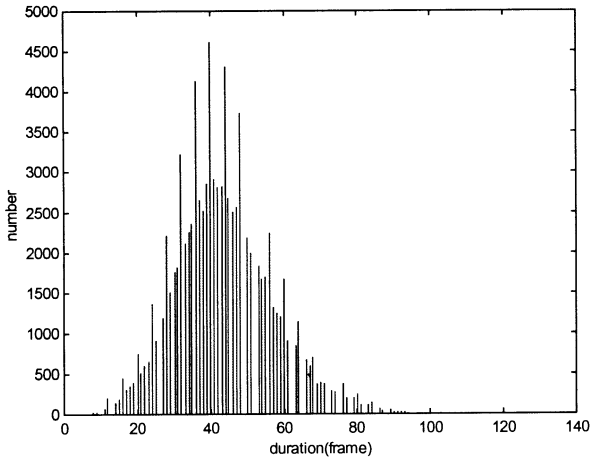


Fig. 1. Plot of total log-likelihood versus iteration number.

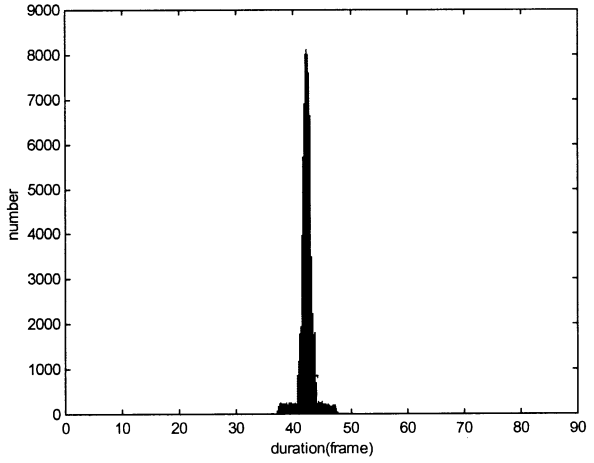
B. Experimental Results of Syllable Duration Modeling

We first examined the effect of syllable duration modeling using the MIC database with the number of prosodic states being set to 16. Table II(b) shows the experimental results. It can be seen from the third and sixth columns of Table II(b) that the variances of the normalized syllable duration were 2.52 and 4.44 frame^2 for the closed and open tests, respectively. Compared with the corresponding values shown in Table II(a), the variances of syllable duration shown in Table II(b) were greatly reduced after compensating the effects of these five affecting factors on the observed one. We also find from the fourth and seventh columns of Table II(b) that the root mean squared errors (RMSEs) between the observed and normalized syllable durations were 1.93 and 3.41 frames for the closed and open tests. Notice that the estimated syllable duration was obtained based on assigning the best prosodic state to each syllable using (6). Fig. 1 shows the plot of total log-likelihood $L(k)$ versus iteration number k. It can be found from Fig. 1 that the EM algorithm converges quickly in the first several iterations. The histograms of the observed and normalized syllable durations for the training set are plotted in Fig. 2. Can be seen from these two figures that the assumptions of Gaussian distribution for these two types of syllable duration are reasonable. From above discussions we can conclude that the proposed syllable duration modeling method is a promising one.

We then examined the case when the number of prosodic states changes. The resulting variances of the normalized syllable duration are shown in Fig. 3. It can be found from the figure that the variance of the normalized syllable duration decreased as the number of prosodic state increased. This shows that the syllable duration modeling was more precise as the number of prosodic states increased. The improvement became saturated



(a)



(b)

Fig. 2. Histograms of (a) the observed syllable duration and (b) the normalized one obtained by the multiplicative model for the training set.

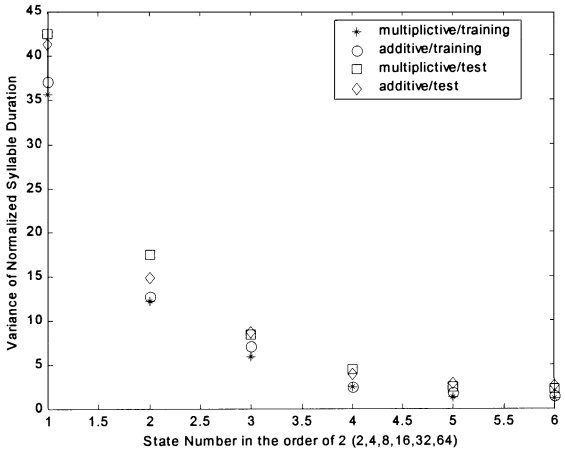


Fig. 3. Relations between the state number and the variance of the normalized syllable duration of multiplicative/additive model for the training and test sets.

when the number of prosodic states equaled 16. Similar findings can be observed for the corresponding RMSEs shown in Fig. 4.

We then inspected the appropriateness of the independence assumption on the affecting factors of the proposed model. This was performed in two ways. One was to examine the effect of the interaction of two affecting factors and another was to analyze large modeling errors. The former was to evaluate the gain of relaxing the independence assumption by considering the combination of two affecting factors while the latter was to check whether large modeling errors were resulted from the independence assumption. To consider the interaction of two affecting factors, we used one CF for each pair of all possible combinations of the two affecting factors. This will improve the accuracy of the model with the paid of increasing the number of model's parameters. Due to the facts that speaker and utterance speaking rates are global normalization factors and prosodic states are hidden and realized in a probabilistic form, we only considered the combination of the base-syllable and tone for simplicity. The syllable duration model was then modified to

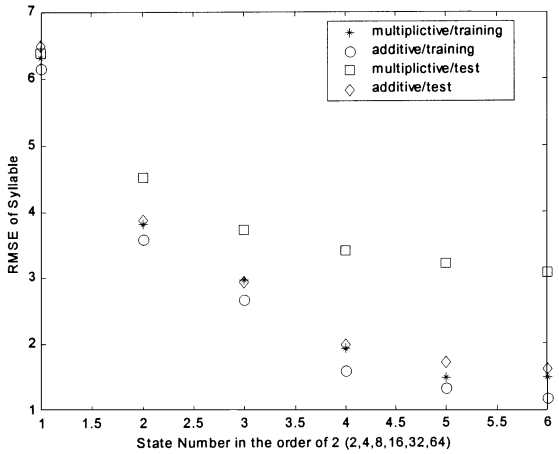


Fig. 4. Relations between the state number and the RMSE of the multiplicative/additive syllable duration models for the training and test sets.

TABLE III
THE ESTIMATED CFS FOR 5 LEXICAL TONES IN THE MULTIPLICATIVE MODEL

tone / γ	1	2	3	4	5
syllable	1.00	1.02	0.99	1.03	0.84

$$Z_n = X_n \gamma_{tj_n} \gamma_{y_n} \gamma_{l_n} \gamma_{s_n} \tag{15}$$

where γ_{tj} was the CF for the syllable with tone t and base-syllable j. Here, we only considered the case of 5 lexical tones and 8 prosodic states. The resulting variance and RMSE of the modified model were 4.67 frame^2 and 2.42 frame, respectively. Compared with the results of 4.73 frame^2 and 2.47 frame obtained by the original model, the improvements were negligible. So the two affecting factors of base-syllable and tone could be independently considered. To analyze large modeling errors, we first identified the syllables with absolute modeling errors located in extreme 5 percentile of error distribution, and then checked the associating affecting factors. Here we considered the case of 7 tones and 16 prosodic states. Some observations were found from the error analysis. Firstly, most large errors were occurred in States 15 and 14 and some were occurred in State 0. More precisely, 55.08%, 41.86% and 2.37% of large modeling errors were occurred in those three states, respectively. As shown in Table V (to be discussed later), States 15 and 14 have the largest

火 huo3$_s$ 腿 tui3$_f$*中 zhong1*亞 ya3$_f$ 硝 xiao1 酸 suan1 鹽 yan2 的 de5*添 tian1 加 jia1 量 liang4*是 shi4*十 shi2 斤 jin1 肉 rou4*放 fang4*一 yi1 分 fen1 一 yi1 厘 li2*以 yi3$_f$ 下 xia4*。

免 mian3$_h$ 費 fei4*試 shi4 吃 chi1*，買 mai3$_f$ 一 i1 送 song4 一 i1*，摸 mo1 彩 tsai3$_f$*贈 tzeng4 獎 jiang3$_f$*，花 hua1 招 jau1 百 bai3$_f$ 出 chu1*。

讓 rang4*孩 hai2 子 zi5*有 you3$_f$*成 cheng2 長 zhang3$_h$ 的 de5*自 zi4 由 you2*，不 bu4 一 yi2 定 ding4 要 yao4* 好 hao3$_h$ 風 feng1 好 hao3$_h$ 雨 yu3$_f$*，但 dan4 非 fei1* 一 yi2 味 wei4*狂 kuang2 風 feng1 暴 bao4 雨 yu3$_f$*；認 ren4 清 qing1*父 fu4 母 nu3$_s$ 與 yu3$_f$ 子 zi3$_s$ 女 niu3$_h$ 的 de5*距 gou4 離 li2*之 zhi1 必 bi4 要 yao4*，別 bie2 怕 pa4*代 dai4 溝 gou1*。

Fig. 5. Some examples of tone state labeling. Here $*$ denotes word boundary.

TABLE IV
THE ESTIMATED CFS FOR 7 TONES IN (A) THE MULTIPLICATIVE AND (B) ADDITIVE DURATION MODELS

tone / γ	1	2	3_f	4	5	3_s	3_h
syllable	1.01	1.02	1.03	1.03	0.85	0.95	0.92
initial	1.00	1.03	1.09	1.00	0.83	1.01	0.87
final	1.01	1.01	1.05	1.04	0.87	0.94	0.85

(a)

tone / γ	1	2	3_f	4	5	3_s	3_h
syllable	0.40	0.93	1.51	1.49	-5.50	-2.11	-3.68
initial	-0.03	0.29	1.70	0.09	-2.77	-0.90	-2.49
final	0.30	0.56	1.14	1.34	-3.83	-2.94	-3.62

(b)

CFs and State 0 has the smallest CF. Moreover, States 15 and 14 have much larger variances in their syllable duration distributions than all other states. Secondly, by more detailedly analyzing large errors occurred in States 15 and 14, we found that the first two most frequently occurred affecting factor combinations of (state, tone, base-syllable) are (15, 5, 43) and (14, 4, 3). But, as excluding the factor that they were mainly resulted from the two most frequently-used characters "(de)" ('s, of, -ly, an adjectival ending, a prepositional phrase, or a relative) and "(shi)" (is), we found that no preferences of base-syllables or tones were associated with those large errors. Based on these two observations, we can therefore conclude that most large modeling errors were occurred in prosodic states with extreme syllable duration and mainly resulted from the large variation in the original syllable duration instead of the independence assumption on the affecting factors of the proposed model. From above discussions, we recognized that it was proper to use the independence assumption in the current syllable duration model.

用 young 14*百 bai 14 子 zi 9 蓮 lian 15*、蕾 lei 11 絲 si 4 花 hua 15*、姬 ji 7 百 bai 11 合 he 15* 龍 long 13 膽 dan 15*、土 tu 5 耳 er 9 其 qi 11 桔 jie 10 梗 geng 13*和 he 14*蒜 suan 4 香 xiang 1 藤 teng 12*為 wei 4 材 cai 15*，以 yi 14*維 wei 4 納 na 6 斯 si 13*執 zhi 8 壺 hu 2 的 de 14*石 shi 10 膏 gao 13*花 hua 3 器 qi 14*烘 hong 10 托 tuo 15*，好 hao 4 一 yi 2 趟 tang 11*春 chun 4 雨 yu 14*濛 meng 3 濛 meng 10 的 de 15*郊 jiao 9 外 wai 14*田 tian 4 野 ye 9 風 feng 13 光 guang 15*。

Fig. 6. Example of prosodic state labeling. Here, $*$ denotes word boundary.

Lastly, we examined the effectiveness of the duration modeling on the TEL database. The database contained short sentential utterances of 500 speakers. The total number of syllables in the database was 42 958. The same training procedure used for the MIC database was applied. The number of states was set to 16. Due to the fact that each speaker only spoke 86 syllables in average, we neglected the affecting factor of utterance-level speaking rate by setting its CF $= 1$. The mean of the observed syllable duration was 25.13 frames (1 frame $=$10 ms) and the variance was 66.78 frame^2. After modeling, the mean and variance of the normalized syllable duration was 23.91 frames and 1.02 frame^2 and the resulting RMSE was 1.38 frames. These results were still quite promising even although the speaking style variation due to the large population of speakers was very high and the accuracy of the observed data due to the automatic segmentation by the HMM models was not as high as that of manual segmentation for the MIC database.

C. An Extension to Initial and Final Duration Modeling

We now extend the above syllable duration modeling to the duration modeling of two sub-syllable units: initial and final. As discussed previously, each Mandarin syllable is composed of an optional consonant initial and a final. The final comprises an optional medial, a vowel nucleus and an optional nasal ending. The goal of the current study is to exploit the relationship between the syllable duration and its component initial and final durations. In this study, both initial and final durations are modeled in the same way as the above syllable duration modeling.

An experiment using the MIC database was conducted to evaluate the performance of the initial and final duration modelings. The experiment was done without considering the null initial and the very short initials of {b, d, g} which are generally difficult to be segmented accurately. As shown in Table II(a), the variances of the observed initial (final) duration were 62.28 (117.06) and 40.02 (104.15) frame^2 for the training and test sets, respectively. Here one frame equals 5 ms. As shown in Table II(b), the variances of the normalized initial (final) dura-

TABLE V
ESTIMATED CFS FOR 16 PROSODIC STATES IN (A) THE MULTIPLICATIVE AND (B) ADDITIVE DURATION MODELS

state / γ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
syllable	0.56	0.72	0.79	0.84	0.89	0.91	0.95	0.98	1.00	1.02	1.05	1.09	1.14	1.22	1.33	1.69
initial	0.30	0.49	0.63	0.71	0.80	0.85	0.86	0.89	0.96	1.00	1.04	1.09	1.12	1.19	1.30	1.61
final	0.50	0.68	0.75	0.80	0.84	0.87	0.91	0.95	0.98	1.00	1.02	1.08	1.14	1.24	1.40	1.86

(a)

state / γ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
syllable	-16.07	-12.36	-9.69	-7.71	-5.79	-4.70	-3.14	-1.94	0	0.12	1.69	4.10	5.87	9.65	15.08	28.74
initial	-11.20	-6.82	-6.22	-4.98	-3.82	-3.60	-2.92	-2.49	-1.40	-0.41	0	0.89	1.39	3.56	6.03	12.69
final	-14.28	-10.24	-7.94	-6.45	-5.15	-4.24	-2.99	-1.73	-0.86	0	0.73	3.12	5.10	8.50	13.42	25.49

(b)

tions reduced to 0.74 (2.12) and 5.92 (3.40) frame^2 by the modeling for the closed and open tests, respectively. The RMSEs between the original and estimated initial (final) durations were 0.97 (1.66) and 2.27 (3.25) frames for the closed and open tests, respectively. This shows that good results of reducing the variance and RMSE were achieved in both modelings of initial and final durations. However, the relatively high variance of the normalized initial duration in the open test shows that the initial duration is more difficult to model. This may results from the intrinsic property of high variability in the durations of different consonant types.

For exploring the relation between syllable duration and initial/final duration, we conduct an experiment to set an additional constraint in the initial/final duration modeling to let the prosodic state of initial/final of a syllable share the same prosodic state of the syllable labeled by the syllable duration modeling. We could then modify the training algorithm of the initial/final model to an ML one with all prosodic states being predetermined by the training procedure of the syllable model. The objective function to be maximized in the ML training was

$$L_i(\lambda_i) = \sum_{n=1}^{N} p(Z_n^i|\lambda_i) + \eta_i \left(\sum_{n=1}^{N} \mu_i \gamma_{t_n}^i \gamma_{y_n}^i \gamma_{j_n}^i \gamma_{l_n}^i \gamma_{s_n}^i - N\mu_z^i \right) \tag{16}$$

for the initial model and

$$L_f(\lambda_f) = \sum_{n=1}^{N} p(Z_n^f|\lambda_f) + \eta_f \left(\sum_{n=1}^{N} \mu_f \gamma_{t_n}^f \gamma_{y_n}^f \gamma_{j_n}^f \gamma_{l_n}^f \gamma_{s_n}^f - N\mu_z^f \right) \tag{17}$$

for the final model. Here Z_n^i and Z_n^f were the observed initial and final durations of syllable n, λ_i and λ_f, η_i and η_f, μ_i and μ_f, γ_p^i, and γ_p^f were, respectively, the parameter sets, the Lagrange multiplier, the mean, and the CF of the affecting factor p for the initial and final duration models. μ_z^i and μ_z^f were the average of Z_n^i and Z_n^f and The ML training algorithm was realized by a sequential optimization procedure [36].

After training, we obtained the variances of the normalized initial (final) duration with shared prosodic states to be 15.39 (14.01) and 38.76 (29.82) frame^2 for the closed and open tests, respectively. The RMSEs between the observed and estimated initial (final) durations were 3.61 (3.57) and 4.85 (5.25) frame for the closed and open tests, respectively. Compared with the previous results shown in Table II(b), the results of shared prosodic state were inferior. This shows that the optimal prosodic states of both initial and final duration models were not matched with those of the syllable duration model. The mismatch may results from the inconsistency in the effect of linguistic features on the initial duration and on the final duration. A previous study [19] found that consonant-lengthening can occur at all initial positions especially at the beginning of a word, while vowel-lengthening can occur only at phrasal final.

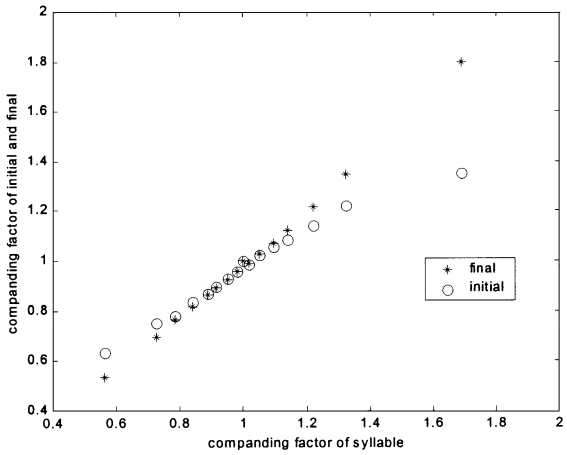


Fig. 7. Relations between the prosodic-state CFs of the initial and final duration models and those of the syllable duration model.

IV. ANALYSES OF CFS

For fully understanding the syllable/initial/final duration models, we analyzed the resulting CFs in detail. Table III shows the CFs of 5 lexical tones for the syllable duration modeling using the MIC database. Can be seen from Table III that Tone 5 has relatively smaller CF. This indicates that the associated syllable duration is much shorter than those of the other four regular tones. This agrees with our prior linguistic knowledge. As for the other four tones, their CFs are very close. Roughly

speaking, Tone 4 has slightly larger CF and Tone 3 has smaller one.

To further exploring the effect of the three Tone 3 patterns on the syllable/initial/final durations, we examined the experimental results of duration modeling using 7 tones shown in Table IV(a). It can be found from the table that the CFs of Tones 3_f, 3_s, and 3_h are quite different. Tone 3_f is the longest while Tone 3_h is the shortest. Some examples are displayed in Fig. 5. It can be found from Fig. 5 that Tone 3_f tends to appear at the end of a prosodic phrase and Tone 3_h tends to appear at the beginning of a word. This observation matches with the prior linguistic knowledge [23]. It is known that, in Taiwan, Tone 3_f appears only at sentence (or prosodic phrase) ending and at isolated syllable, while Tone 3_h may appear at all other places in continuous speech.

Table V(a) shows the CFs of 16 prosodic states for the MIC database. It can be found from Table V(a) that State 15 has the largest CF while State 0 has the smallest one for all three duration models. Fig. 6 shows an example of prosodic state labeling for a part of a Mandarin paragraphic utterance by the EM training algorithm. From Fig. 6, we find that State 15 usually associates with the ending syllables of sentences or phrases and State 0 always associates with intermediate syllables of polysyllabic words. Besides, prosodic states with larger CFs tend to appear at word boundaries while those with smaller CFs tend to appear at intermediate parts of words or prosodic phrases. The finding also complies with the prior knowledge of the lengthening effect for the last syllable of a prosodic phrase or sentence.

We then examined the relationship between the prosodic-states CFs of the syllable duration model and those of the initial and final duration models. Fig. 7 displays the prosodic-state CFs of the initial and final duration models versus those of the syllable duration model for the shared-prosodic-state case. Can be seen from Fig. 7 that the CFs matched well in the three models for all states except the extreme cases of States 0 and 1 which have the smallest CFs and of States 13, 14 and 15 which have the largest CFs. At these extreme cases, final (initial) duration was compressed or stretched more (less) serious than syllable duration.

We then analyzed the CFs of 411 base-syllables for the three duration models using a top-down decision tree method. The method used the following criterion to determine whether a node (cluster) was to be split into two son nodes (subclusters) based on a specific question.

> Split based on the question with maximum $|\mu_1 - \mu_2|$, if ($|\mu_1 - \mu_2| >$ Threshold A) and ($v >$ Threshold B) and ($n_1 >$ Threshold C) and ($n_2 >$ Threshold C). Here (μ, v, n), (μ_1, v_1, n_1) and (μ_2, v_2, n_2) are, respectively, triples of means, variances and sample counts of the node and the two son nodes split based on a question.

There were in total 15 questions used in the construction of the decision trees for the three models. The question set was designed to consider: (1) the way of articulation, such as aspiration, voiced/unvoiced, stop, and fricative; (2) the phonetic structure of Mandarin base-syllables, such as single vowel, compound vowel, with nasal ending, and with medial; and (3) the category of vowel nucleus, such as open vowel. They are listed in Appendix.

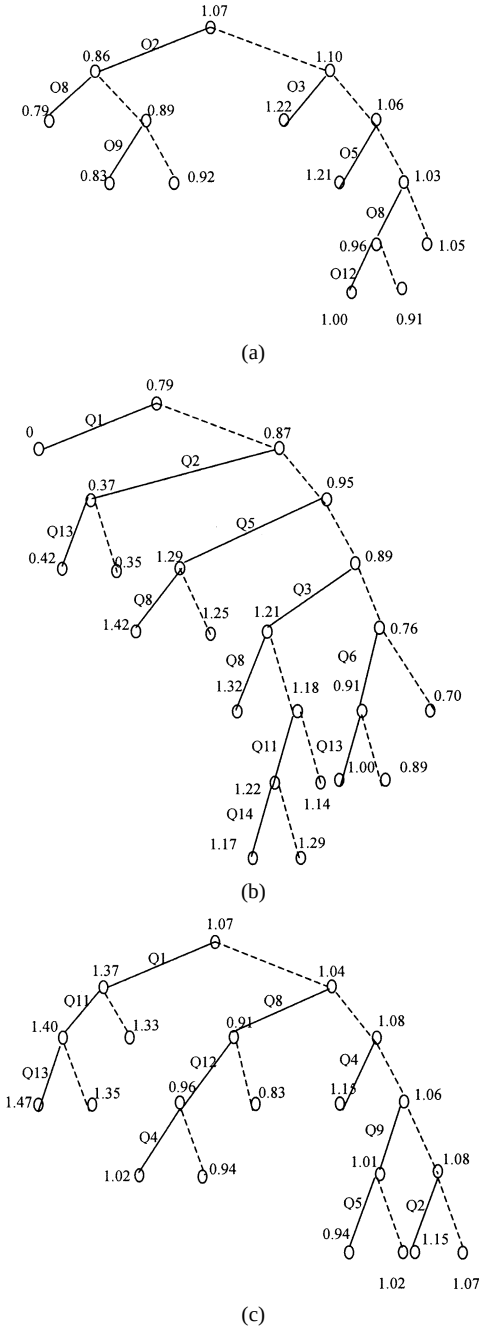


Fig. 8. Decision tree analyses of the base-syllable CFs for (a) syllable, (b) initial, and (c) final duration models. The number associated with a node is the mean of the CFs of the base-syllables belonging to the cluster. Solid line indicates positive answer to the question and dashed line indicates negative answer.

The three trees we constructed are displayed in Fig. 8. It can be found from the syllable-duration tree, shown in Fig. 8(a), that the syllables with initial belonging to {b, d, g} (based on Q2) are shorter (average $\mathrm{CF} = 0.86$) and syllables with initial belonging to {f, s, sh, shi, h, ts, ch, chi} (based on Q3 and Q5) are generally longer (average $\mathrm{CF} = 1.22$ and 1.21). Besides, syllables with final being single vowel (based on Q8) are much shorter (average $\mathrm{CF} = 0.79$ and 0.96). In Fig. 8(b), the initial-duration tree shows that an initial is shorter when it belongs to {b, d, g} (based on Q2) and is longer when it belongs to {f, s, sh, shi, h, ts, ch, chi} (based on Q3 and Q5). Moreover,

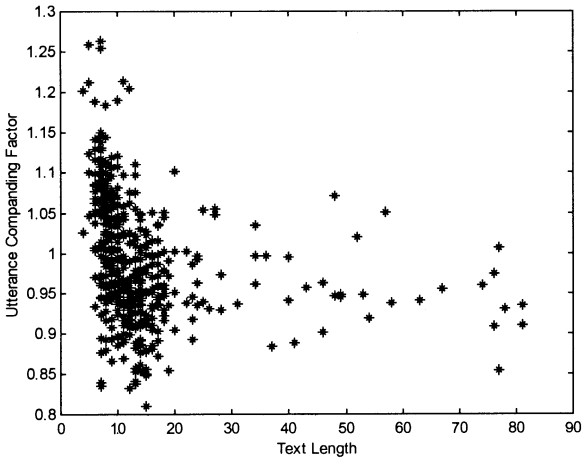


Fig. 9. Relation between the utterance length (in syllable) and utterance CF.

an initial becomes longer when it is followed by a final with single vowel (based on Q8). Lastly, as shown in Fig. 8(c), a final is shorter when it is a single vowel (based on Q8) or is preceded by an initial belonging to {ts, ch, chi} (based on Q5). A final is much longer when the preceding initial is a null one (based on Q1) and is longer when it contains a medial (based on Q11). All above observations match with the knowledge of the phonetic characteristics of Mandarin base-syllables. These trees can be used to help us making predictions of syllable/initial/final durations according to the base-syllable type.

We then examined the relationship between the utterance-level speaking speed and utterance length. Fig. 9 displays the scattering plot of utterance CF versus utterance length (in syllable). It can be found from the figure that the speaking speeds of utterances with length shorter than 15 syllables spread widely, while utterances with length longer than 15 syllables tend to be pronounced faster (i.e., $\mathrm{CF} < 1$).

We then compared a speaking rate estimate by the proposed model with a conventional one based on average syllable duration. The former is the product of utterance CF and speaker CF while the latter (referred to as Scheme A) is the average duration of all syllables in the utterance. Correlation coefficients of these two estimates for the MIC database and its two subsets, MIC-para and MIC-sent, were calculated and displays in the second column of Table VI. It can be found from the table that relatively high correlation coefficient of 0.92 was obtained for the MIC-para data set while a low value of 0.35 was obtained for the MIC-sent data set. This shows that the average syllable duration can be a good estimate of speaking rate only when the length of the utterance is long. This mainly results from the content-richness of long utterances to smooth out the influences of various affecting factors. To confirm this viewpoint, three other schemes of speaking rate estimation by averaging syllable duration with some affecting factor being compensated were also tested. They included (1) Scheme B—compensated by the CFs for tone; (2) Scheme C—compensated by the CFs for tone and base-syllable; and (3) Scheme D—compensated by the CFs for tone, base-syllable and prosodic state. The experimental results are displayed in the 3rd, 4th, and 5th columns of Table VI. We find from the table that the value of correlation coefficient increased significantly when more affecting factors were compensated. A correlation coefficient of 0.977 for the MIC-sent data set was obtained for Scheme D.

TABLE VI
CORRELATION COEFFICIENTS BETWEEN THE PRODUCT OF BASE-SYLLABLE CFS OF UTTERANCE AND SPEAKER AND THE AVERAGE SYLLABLE DURATION (ASD) FOR THE MIC DATABASE AND ITS TWO SUBSETS. A: ASD; B: ASD WITH TONE COMPENSATED; C: ASD WITH TONE AND BASE-SYLLABLE COMPENSATED; D: ASD WITH TONE, BASE-SYLLABLE, AND PROSODIC STATE COMPENSATED

Database	Correlation Coefficient			
	A	B	C	D
MIC	0.48	0.49	0.58	0.981
MIC-para	0.92	0.92	0.93	0.998
MIC-sent	0.35	0.38	0.5	0.977

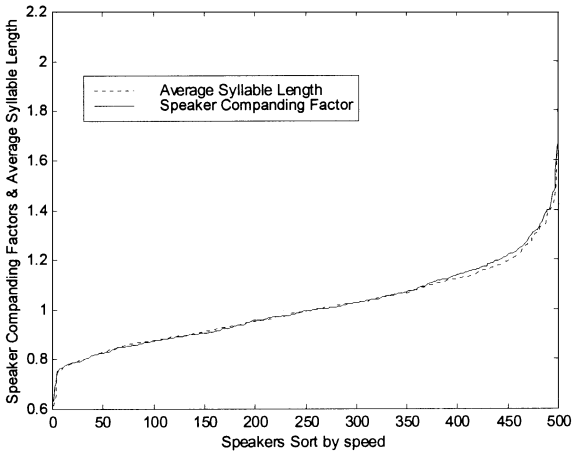


Fig. 10. Relation between the average syllable duration uttered by different speakers and their speaker CFs.

We then analyzed the speaker CFs of the syllable duration model trained using the TEL database to see whether they were matched with the conventional speaking rate estimate of average syllable duration. Fig. 10 shows the speaker CF of the model versus speaker-level average syllable duration. Notice that each speaker spoke about 87 syllables and the speaker-level average syllable durations have been sorted in an increasing order for easy observation. It is clearly shown in Fig. 10 that the two curves match quite well to each other, even in the extreme cases of very slow and very fast speeds. We can therefore conclude that the speaker CFs of the syllable-duration model were effective estimates of speakers' speaking rates.

From above discussions, we find that the proposed syllable/initial/final models agree well with our general linguistic knowledge of Mandarin speech in many aspects. We can therefore conclude that the duration modeling method is effective on separating confounding influences of several major affecting factors.

V. ADDITIVE DURATION MODEL

In [19], a number of analyses for segment durations of Mandarin speech were performed and used to built additive and multiplicative duration models via computing the estimated intrinsic

durations of segments and the coefficients of contextual factors. It reported that the multiplicative model performed, in general, better than the additive model. For performance comparison, we then extend our study to construct additive model for syllable/initial/final duration. By considering the same affecting factors, we express the model by

$$Z_n = X_n + \gamma_{t_n} + \gamma_{y_n} + \gamma_{j_n} + \gamma_{l_n} + \gamma_{s_n}. \quad (18)$$

The model can be trained by the same EM algorithm with (1) being replaced by (18). The auxiliary function is accordingly changed to

$$Q(\bar{\lambda}, \lambda) = \sum_{n=1}^{N} \sum_{y_n=1}^{Y} p(y_n|Z_n, \bar{\lambda}) \log p(Z_n, y_n|\lambda) + \eta\left(\sum_{n=1}^{N} (\mu + \gamma_{t_n} + \gamma_{y_n} + \gamma_{j_n} + \gamma_{l_n} + \gamma_{s_n}) - N\right. \quad (19)$$

With the number of prosodic states being set to 16, the experimental results using the MIC database are shown in Table II(c). It can be found from Table II(a) and (c) that the variances of the training (test) data set were greatly reduced from 180.17 (136.26) to 2.53 (3.97) for syllable duration, from 62.28 (40.02) to 0.78 (1.73) for initial duration, and from 117.06 (104.15) to 1.84 (2.85), respectively. The corresponding mean squared errors between the observed and estimated syllable durations were 1.59 (1.99), 0.89 (1.32), and 1.36 (1.69) for syllable, initial, and final duration modelings, respectively. Compared with the experimental results of multiplicative models shown in Table II(b), the performances of these additive models were slightly better.

We then examined the results when the number of prosodic states increased. As shown in Figs. 3 and 4, both the variance of the normalized syllable duration and the RMSE between the observed and estimated syllable durations decreased as the number of prosodic states increased. They became saturated when the prosodic state number equals 16.

The CFs for tone and prosodic states are listed in Tables IV(b) and V(b). By comparing them with those of the multiplicative models, we found that they are very consistent. A negative (positive) CF in additive model corresponded to a CF with value less (greater) than 1 in multiplicative model. Moreover, the distance of a CF in additive model to zero was approximately equal to the product of the mean of the observed duration and the distance of the corresponding CF in multiplicative model to one, i.e.,

$$CF_{\text{additive}} \approx (CF_{\text{multiplicative}} - 1)\mu_z. \quad (20)$$

To further examining the consistency of the two modeling approaches, we calculate the distribution of the pair of syllable prosodic states labeled, respectively, by the multiplicative and additive syllable duration models. Table VII shows the results of the MIC database for the case of 8 prosodic states. It can be found from the table that the distribution concentrates in the vicinity of main diagonal. This shows that the prosodic state labelings by these two models were highly consistent.

TABLE VII
DISTRIBUTION OF PROSODIC STATES LABELED BY THE MULTIPLICATIVE AND ADDITIVE MODELS

mul / add	0	1	2	3	4	5	6	7
0	2879	3093	122	1	0	0	0	0
1	891	5154	2506	131	5	0	0	0
2	345	2329	5093	2299	102	1	0	0
3	70	665	2621	6432	2171	65	0	0
4	5	157	514	2725	9459	1984	31	0
5	0	7	36	220	2080	11742	3025	1
6	0	0	0	2	15	695	16388	2777
7	0	0	0	0	0	0	188	9594

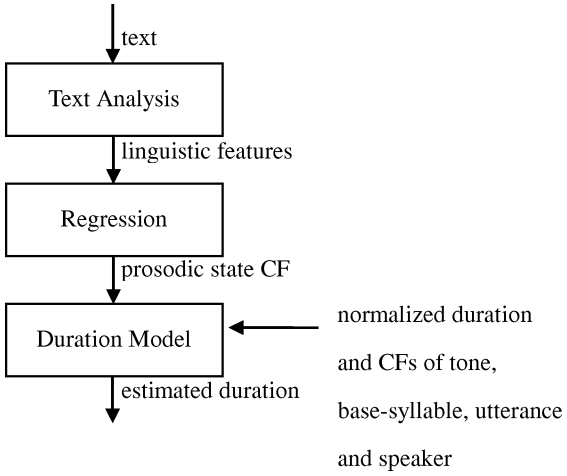


Fig. 11. Hybrid statistical/regression approach for syllable duration prediction.

VI. APPLICATION TO DURATION PREDICTION FOR TTS

A hybrid method incorporating the above duration model into a linear regression method to predict syllable duration for TTS is proposed. Fig. 11 shows a block diagram of the method. Instead of directly predicting syllable duration from the input linguistic features by the conventional linear regression method, the proposed method first estimates the prosodic-state CF from the linguistic features by the linear regression technique. Input linguistic features used include: 1) current word length: $\{1, 2, 3, > 3\}$; 2) current syllable position in word: {1st, intermediate, last}; 3) sentence length: $\{1, [2, 5], [6, 10], [11, 15], [16, 20], >20\}$; 4) current syllable position in sentence: {1st, 2nd 3rd, [4th, 5th], [6th, 7th], [8th, 11th],last, 2nd last, 3rd last, [5th last, 4th last], [7th last, 6th last], [11th last, 8th last], others}; Smaller count number from the beginning or ending wins and count from the ending breaks the tie; 5) punctuation mark after the current syllable (12 types + null); 6) part of speech (53 types) categorized by the Speech Lab of NCTU, Taiwan. Meanwhile, the CFs of base-syllable and tone are directly assigned based on the results of text analysis. The CFs of speaker and utterance are assigned to the values found by the EM training algorithm to disregard the effect of speaking rate. They can also be directly assigned to meet the required speaking speed control of TTS in practical

TABLE VIII
RMSES OF THE HYBRID METHOD USING MULTIPLICATIVE AND ADDITIVE MODELS WITH 8 PROSODIC STATES AND THE LINEAR REGRESSION METHOD

RMSEs	Closed Test	Open Test
Hybrid/mul	9.32	11.18
Hybrid/add	8.80	12.04
Regression	9.37	15.47

applications. The normalized syllable duration can be obtained by a linear regressive estimation like the prosodic-state CF. But due to the fact that the variances of the normalized syllable duration in both the multiplicative and additive models are very small, we simply set its value to be its mean. Lastly, all these parameters are combined and used in the syllable duration model to generate the syllable duration estimate. Notice that the linguistic features used here are extracted from the input text by an automatic word tokenization algorithm with an 80 000-word lexicon.

For performance comparison, the conventional linear regression method was also implemented. It used a linear combination of weighted input linguistic features to generate the syllable duration estimate. For fair comparison, input linguistic features used in the method comprised all above features and some other syllable-level features, including lexical tones (5 types) of the preceding, current and succeeding syllables, initials (21 types + null) of the current and succeeding syllables, medial 3 types + null) of the current syllable, and finals (14 types) of the preceding and current syllables.

Two schemes of the hybrid method using respectively the multiplicative and additive duration models were tested. Experimental results using the MIC database are shown in Table VIII. It can be found from the table that both schemes of the hybrid method outperformed the linear regression method. RMSEs of 9.32 (8.8) and 11.18 (12.04) frame were obtained for the hybrid method with multiplicative (additive) duration model in closed and open tests, respectively.

VII. CONCLUSIONS AND FUTURE WORKS

A new statistical-based duration modeling approach for Mandarin speech has been proposed in the paper. Experimental results have confirmed that it was very effective on separating several main factors that seriously affects the syllable duration of Mandarin speech. Aside from greatly reducing the variance of the modeled syllable duration, the estimated CFs conformed well to the prior linguistic knowledge of Mandarin speech. Besides, the prosodic-state labels produced by the EM training algorithm were linguistically meaningful. So it is a promising syllable duration modeling approach for Mandarin speech.

Some future works are worthwhile doing. Firstly, the syllable duration model can be further improved via considering more affecting factors. This needs the help of a sophisticated text analyzer. Secondly, the applications of the model to both ASR and TTS are worth further studying. Lastly, the approach may be extended to the modeling of other prosodic features such as pitch, energy, and inter-syllable pause duration.

APPENDIX

The question set used to construct the decision trees for the base-syllable CFs of the syllable/initial/final duration models:

Q1. Null initial?
Q2. Initial in {b, d, g}?
Q3. Initial in {f, s, sh, shi, h}?
Q4. Initial in {m, n, l, r}?
Q5. Initial in {ts, ch, chi}?
Q6. Initial in {p, t, k}?
Q7. Initial in {tz, j, ji}?
Q8. Single vowel?
Q9. Compound vowel?
Q10. Nasal ending?
Q11. With medial?
Q12. Open vowel?
Q13. Vowel begins with {i}?
Q14. Vowel begins with {u}?
Q15. Vowel begins with {iu}?

ACKNOWLEDGMENT

The authors thank the Chunghwa Telecommunication Labs and the Association of Computational Linguistics and Chinese Language Processing for supplying the speech databases.

REFERENCES

[1] C. Mitchell, M. Harper, L. Jamieson, and R. Helzermam, "A parallel implementation of a hidden markov model with duration modeling for speech recognition," in *Digital Signal Process.*, vol. 5, 1995, pp. 43–57.
[2] X. Huang, H. Hon, M. Huang, and K. Lee, "A comparative study of discrete, semicontinuous, and continuous hidden markov models," *Comput., Speech, Lang.*, vol. 7, pp. 359–368, 1993.
[3] S. Levinson, "Continuously variable duration hidden markov models for speech analysis," in *Proc. IEEE ICASSP*, 1986, pp. 1241–1244.
[4] A. Anastasakos, R. Schwartz, and H. Shu, "Review of text-to-speech conversion for English," in *Proc. ICASSP*, vol. 1, 1995, pp. 628–631.
[5] D. H. Klatt, "Review of text-to-speech conversion for english," *J Acoust. Soc. Amer*, vol. 82, pp. 137–181, 1987.
[6] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1309–1320, 1989.
[7] N.-H. Pan, W.-T. Jen, S.-S. Yu, M.-S. Yu, S.-Y. Huang, and M.-J. Wu, "Prosody model in a Mandarin text-to-speech system based on a hierarchical approach," in *IEEE Int. Conf. Multimedia and Expo*, vol. 1, 2000, pp. 448–451.
[8] B. Ao, C. Shih, and R. Sproat, "A corpus-based Mandarin text-to-speech synthesizer," in *Proc. ICSLP*, vol. S29, 1994, pp. 1771–1774.
[9] B. Mobius and J. van Santen, "Modeling segmental duration in german text-to-Speech synthesis," in *Proc. ICSLP*, vol. 4, 1996, pp. 2395–2398.
[10] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 226–239, 1998.
[11] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1553–1573, Apr. 1988.
[12] ——, "Segmental durations in connected-speech signals: Syllabic stress," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1574–1585, Apr. 1988.
[13] M. D. Monkowski, M. A. Picheny, and P. S. Rao, "Context dependent phonetic duration models for decoding conversational speech," in *Proc. ICASSP*, vol. 1, 1995, pp. 528–531.
[14] J. W. A. Fackrell, H. Vereecken, J.-P. Martens, and B. V. Coile, "Multilingual prosody modeling using cascades of regression trees and neural networks," in *Proc. EUROSPEECH*, 1999.
[15] K. Takeda, Y. Sagisaka, and H. Kuwabara, "On sentence-level factors governing segmental duration in japanese," *J. Acoust. Soc. Amer.*, vol. 86, no. 6, pp. 2081–2087, Dec. 1989.

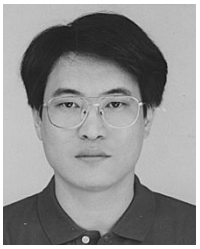
[16] N. Kaiki, K. Takeda, and Y. Sagisaka, "Statistical analysis for segmental duration rules in japanese speech synthesis," in *Proc. ICSLP*, vol. 1.5, 1990, pp. 17–20.
[17] K. E. A. Silverman and J. R. Bellegarda, "Using a sigmoid transformation for improved modeling of phoneme duration," in *Proc. ICASSP*, 1999.
[18] J. P. H. van Santen, "Contextual effects on vowel duration," *Speech Commun.*, vol. 11, pp. 513–546, 1992.
[19] C. Shih and B. Ao, "Duration study for the bell laboratories Mandarin text-to-speech system," in *Progress in Speech Synthesis*. New York: Springer, 1997, pp. 383–399.
[20] J. van Santen, C. Shih, B. Mobius, E. Tzoukermann, and M. Tanenblatt, "Multi-lingual duration modeling," in *Proc. EUROSPEECH*, 1997.
[21] J. R. Bellegarda, K. E. A. Silverman, K. Lenzo, and V. Anderson, "Statistical prosodic modeling: From corpus design to parameter estimation," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 52–66, Jan. 2001.
[22] M. Riedi, "Modeling segmental duration with multivariate adaptive regression splines," in *Proc. EUROSPEECH*, 1997.
[23] C.-L. Shih, "Tone and intonation in Mandarin," in *Working Papers of the Cornell Phonetics Laboratory*, June 1988, pp. 83–109.
[24] J. Ma, L. Xu, and M. I. Jordan, "Asymptotic convergence rate of the EM algorithm for gaussian mixtures," *Neural Comput.*, vol. 12, pp. 2881–2907, 2000.
[25] L. Xu, "Comparative analysis on convergence rates of the em algorithm and its two modifications for gaussian mixtures," *Neural Process. Lett.*, vol. 6, pp. 69–76, 1997.
[26] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for gaussian mixtures," *Neural Comput.*, vol. 8, pp. 129–151, 1996.
[27] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 469–481, Oct. 1994.
[28] A. Batliner, R. Kompe, A. Kiebling, H. Niemann, and E. Noth, "Syntactic-prosodic labeling of large spontaneous speech data-bases," in *Proc. ICSLP*, 1996, pp. 1720–1723.
[29] W.-J. Wang, Y.-F. Liao, and S.-H. Chen, "Prosodic modeling of Mandarin speech and its application to lexical decoding," in *Proc. EUROSPEECH 99*, vol. 2, pp. 743–746.
[30] K. Iwano and K. Hirose, "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition," in *Proc. ICASSP*, 1999, pp. 133–136.
[31] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information," in *Proc. ICASSP*, 1993, pp. II-51–II-54.
[32] H.-Y. Hsieh, R.-Y. Lyu, and L.-S. Lee, "Use of prosodic information to integrate acoustic and linguistic knowledge in continuous Mandarin speech recognition with very large vocabulary," in *Proc. ICSLP*, vol. 2, 1996, pp. 809–812.
[33] S. de Tournemire, "Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in french," in *Proc. EUROSPEECH*, 1997.
[34] F.-C. Chou, C.-Y. Tseng, K.-J. Chen, and L.-S. Lee, "A chinese text-to-Speech based on part-of-Speech analysis, prosodic modeling and nonuniform units," in *Proc. ICASSP*, 1997, pp. 923–926.
[35] S. Young and G. Bloothooft, *Corpus-Based Methods in Language and Speech Processing*. Norwell, MA: Kluwer, 1997, pp. 1–26.
[36] C. C. Ho and S. H. Chen, "A maximum likelihood estimation of duration models for taiwanese speech," in *Proc. ISAS/SCI'2000*, Orlando, FL, July 2000.
[37] ——, "A hybrid statistical/RNN approach to prosody synthesis for taiwanese TTS," in *Proc. ICSLP'2000*, Beijing, China, Oct. 2000.
[38] L. Wen-hsing and C. Sin-horng, "A novel syllable duration modeling approach for Mandarin speech," in *Proc. ICASSP*, vol. 1, 2001, pp. 93–96.
[39] W.-H. Lai and S.-H. Chen, "Analysis of syllable duration models for Mandarin speech," in *Proc. ICASSP*, vol. 1, 2002, pp. 497–500.

Sin-Horng Chen (S'81–M'83–SM'94) received the B.S. degree in communication engineering and the M.S. degree in electronics engineering from National Chiao Tung University, Taiwan, R.O.C., in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University in 1983.

From 1978 to 1980, he was an Assistant Engineer for Telecommunication Labs, Taiwan. He became an Associate Professor and a Professor at the Department of Communication Engineering, National Chiao Tung University, in 1983 and 1990, respectively. He also became the Department Chairman (1985–1988, 1991–1993). His major research area is speech processing, especially in Mandarin speech recognition and text-to-speech.

Wen-Hsing Lai received the B.S. and M.S. degrees in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1988 and 1990, respectively, and is currently pursuing the Ph.D. degree.

She is an Associate Researcher at Telecommunication Labs, Chunghwa Telecom Co., Taiwan. Her major research area is speech processing, especially in text-to-speech.



Yih-Ru Wang received the B.S. and M.S. degrees from the Department of Communication Engineering in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, National Chaio Tung University, Taiwan, R.O.C., in 1995.

He was an Instructor of the Department of Communication Engineering, National Chaio Tung University, from 1987 to 1995. In 1995, he became an Associate Professor. His general research interests are Mandarin spontaneous speech recognition and the application of neural network in speech processing.