Using Duration Information in HMM-based Automatic Speech Recognition

ZHU Yu

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy
in
Electronic Engineering

June 2005

Acknowledgment

I am greatly obliged to my supervisor Prof. Tan Lee for his guidance and support throughout this research. I also wish to give my thanks to Dr. Frank Soong, Prof. Y.T. Chan, Prof. P.C. Ching for their insightful suggestions.

I would like to appreciate Ms. Y. Qian for her expert support. I would like to thank Mr. N. H. Zheng, C. Qin for part of the experimental speech data. I would like to thank people who accompanied me, encouraged me, appreciated me and helped me, especially the friends in DSP and my roommates. To name only some of them: Ms. C. Yang, Dr. W.K. Lo, Mr. W. Zhang, Ms. Yvonne Lee, Mr. M. Yuan, Mr. Herman Yau, Ms. L.Y. Ngan, Ms. Patgi Kam, Ms. Anthea Tam, Ms. Joyce Chan, Ms. Y. J. Li, Ms S. Y. Tao, Mr. Michael Zhang, Mr. Arthur Luk, and Ms. X. L. Zhuang, Ms. X. Y. Zhang, Ms. M. M. Yang and Ms. Canny Zou.

Finally, I would like to express sincere gratitude to my family, Mr. Jianming Zhu, Ms. Chenfang Sun, Ms Dandan Zhu, for their continuous encouragement and discipline.

Thanks again to all people. Without your help I couldn't have this harvest during these two years.

Abstract of thesis entitled:

Using Duration Information in HMM-based Automatic Speech Recognition

Submitted by **Zhu Yu**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **Feb 2005**.

Hidden Markov Model (HMM) is the most predominant technique for automatic speech recognition (ASR). However, HMM is inadequate in representing the temporal structure of speech signal. An HMM does not give effective control to the duration properties of speech segments being modeled. In many applications, HMM-based systems frequently make errors. A significant portion of these recognition errors exhibit unreasonable absolute duration or relative duration. To alleviate this problem, explicit modeling of duration information has been proposed to confine the duration of recognized segments.

In this research, we focus on explicit duration modeling for Cantonese connected-digit recognition. Connected-digit recognition has many practical applications that often require high accuracy. Despite its limited vocabulary size, it is not straightforward to attain the desired performance level mainly because the combination of digits is unrestricted. The syllable compositions of Cantonese digits are generally very simple. In particular, the digit "5" and "2" can be regarded as of single phonemes. In our baseline recognition system without duration modeling, it is observed that a significant portion of recognition errors are due to the insertion of

digit "5" and "2" with very short durations. We propose to use both absolute duration models and relative duration models to constrain the duration of recognized digits. In particular, the relative duration of the tail part of a Cantonese digit is proposed as a kind of useful duration information. Considering that speaking rate variation may weaken the effectiveness of the duration model, speak-rate-dependent duration models are also investigated.

The duration models are integrated into the dynamic programming search algorithms for speech recognition. For each decision on path extension, the duration models are used to contribute an additional probabilistic score to the acoustic path score. Algorithms are developed for incorporating state-level duration score and word-level duration score to HMM-based recognition respectively. In the decoding algorithm, a weighting factor is used to balance the contribution from the acoustic model and the duration model.

Experiments have been carried out with the CUDIGIT corpus to evaluate the effectiveness of the models for various types of duration information. Empirical weights for different duration information are developed by many trials on male speech data. With these weights, the use of different duration information shows performance improvement to various degrees. For male speech, the recognition accuracy is improved by up to 1.06%. For female data, the recognition accuracy is improved by up to 0.51%. Speaking-rate-dependent duration models have also been evaluated. It shows further performance improvement over the speaking-rate-independent models.

Lastly, experiments have been carried out on a separate set of speech data. The results show that the empirical weights trained from CUDIGIT data are equally effective. The improvement of recognition accuracy is up to 2.36%.

摘要

隱馬爾可夫模型(HMM)技術是目前自動語音識別領域的最主要的技術。然而當使用 HMM 刻劃語音信號的時候，它不能充分刻劃語音信號的時間結構。這表現在它不能對它所刻劃的語音段的段長和相對段長加以有效的控制。這裡所指的語音段包括 HMM 狀態和詞。在許多採用 HMM 技術的自動語音識別系統中，識別錯誤常常發生。有一大部分錯誤呈現出不合理的段長或相對段長特徵。爲了減少這些識別錯誤，顯式段長信息模型可以用來約束被識別出的語音段段長。

在這項研究中我們致力於為廣東話數字串識別建立顯式段長信息模型以提高識別性能。數字串識別有廣泛的應用，這些應用常常要求很高的識別率。儘管它的詞彙量很小，但是數字組合的任意性使得期望的識別率不容易達到。廣東話數字的音節十分簡單。數字“5”和數字“2”可以看成是只有一個音素的音節。在基綫系統中，我們發現有大量的識別錯誤是因爲非常短的“5”和“2”的插入。我們建議為段長和相對段長建立模型來約束段長。特別的，我們提出為廣東話數字的相對尾長建立模型。考慮到語速變化會影響段長模型的有效使用，我們進一步探討了語速相關的段長信息模型。

段長信息模型被使用到進行語音識別的動態規劃算法中。在每次作出路徑延長的決策時，由段長信息模型計算出來的概率得分被加到傳統的路徑概率得分裏面。我們分別實現了採用 HMM 技術的自動語音識別系統使用狀態段長和詞段長模型的語音識別算法.在識別算法中我們使用權重系數來平衡聲學模型和段長信息模型的相對貢獻。

我們在廣東話數字語音資料庫 CUDIGIT 上進行了使用不同段長信息的實驗。通過在男性語音數據的多次探測性試驗，我們得到了對不同段長信息的經驗權重系數。使用這些系數，識別性能得到了不同程度的提高。在男性語音識別試驗中，識別率最多可提高 1.06%。在女性語音識別試驗中，識別率最多可

提高 0.51%。我們也進行了使用語速相關的段長信息模型的實驗，這為識別性能帶來了進一步的提高。使用之前得到的經驗系數，我們又在其他語音資料庫進行了實驗來檢驗經驗系數的有效性，發現識別率最多可提高 2.36%。

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Speech and its temporal structure

Acoustically, speech is characterized by its spectral properties and temporal structure. By spectral properties, we refer to the short-time frequency spectrum. From the spectrographic representation, we can observe the energy of the frequency content over time. The temporal structure includes the duration of sound segments, intonation, loudness of sounds. It is also commonly referred to as the prosodic pattern.

Duration of sound segments provides important acoustic cues for human speech recognition. In many languages, there are cases that differentiating between some certain word(s) pair depends heavily on the duration. For instance, in English, there are confusing word pairs, "cheap" and "chip", "beat" and "bit", in which the vowel durations are contrastively different. In Finish [1], phone durations can be the clue in discriminating between certain words. Temporal structure modeling has attracted considerable attention in the ASR and relevant speech research areas.

This thesis concerns the use of duration for automatic speech recognition (ASR). In particular, we focus on the word-level and subword-level duration for the recognition of Cantonese connected digits.

1.2. Previous work on the modeling of temporal structure

The research on exploiting temporal features started in the early 1970s, when non-linear time normalization using dynamic programming was established as the most

effective speech recognition technique. Endpoint constraints, monotonicity constraints, various local continuity constraint, global path constraints, and slope weights [2]-[4] were adopted to impose control in the search for optimal warping path for, acoustically as well as linguistically meaningful, time normalization to account for the inherent temporal variability in speech utterances.

In recent years, data-driven statistical approaches have become predominant for ASR. In particular, Hidden Markov Model techniques provide a formal mathematical framework for modeling temporal and spectral variability in speech signals. This framework is amenable to a set of mathematically rigorous algorithms for automatic model parameter estimation and pattern classification [russel87]. For speech recognition, an HMM has a number of states that are arranged in a left-to-right topology. The HMM states may be thought of as a sequence of acoustic targets that constitute an utterance. The conditional independent state output probability density functions (pdf) describe the spectral variability in the realization of these targets. The underlying assumption of first-order Markov model governs the temporal variability. With these assumptions, HMM is sufficiently simple to enable mathematically rigorous optimization and search strategies such as Viterbi algorithm to be employed [5]. The successful application of HMM techniques in ASR makes HMMs become the most prominent techniques for speech recognition today.

However, in many applications, HMM-based systems frequently make errors. A significant portion of these recognition errors exhibit unreasonable time durations or relative duration proportion. For instance in the connected-digit task of Cantonese, Korean, mandarin and English, a large portion of recognition errors are due to insertions of short sounds.

This is due to the over-simplified assumptions of HMM in representing the temporal structure of speech signal. Under the first order Markov assumption, transition probability depends on the immediately neighboring states. It is inadequate to control the time duration of speech segments that correspond to an HMM state or the entire model. It is inadequate to give control to the relative duration of speech segments to each other. It is also pointed out that implicit duration modeling of HMM is not appropriate [5]-[7]. With the probabilistic self-transition in the Markov model, state duration is implicitly modeled by a geometric distribution. However, the geometric distribution often mismatches with the measured data in practical cases.

In summary, HMM-based recognizer under-utilizes duration information in the incoming speech due to inadequate modeling or knowledge in duration. To alleviate the problem, we can develop duration knowledge explicitly to complement HMM. With the duration knowledge, duration information in the incoming speech signal would be better utilized. The duration knowledge source should be a quantitative model. Then it can be used to assess the duration feature[1] extracted and output duration probability. This technique is referred to as explicit duration modeling. Subsequently, the duration probability can be used to contribute to the overall probability for speech recognition.

1.3. Integrating explicit duration modeling in HMM-based ASR system

There are a number of issues concerned with the approaches to integrate explicit duration modeling into the ASR process. They are mainly on constructing of explicit duration model and using the explicit duration model for recognition. Firstly, appropriate duration features need to be identified for duration modeling. Secondly, a parametric distribution has to be assumed for statistical characterization the duration features. The subsequent problem is how to estimate the parameters of duration model. Lastly, the problem is how to incorporate of duration models to the recognition process.

1.4. Thesis outline

Previous works on integrating explicit duration information in ASR will be reviewed in Chapter 2. A baseline system will be described in Chapter 3. Its recognition performance will be analyzed. Suggested techniques on explicit duration modeling and using explicit duration modeling for Cantonese connected-digit recognition will be discussed in Chapter 4 and Chapter 5. Experimental results and analysis of the

[1] The measurement of duration information is referred as the duration feature.

experiments on using duration modeling are given in Chapter 6 and Chapter 7. Lastly, the conclusion will be given in Chapter 8.

Chapter 2

Background

The principles of automatic speech recognition and HMM-based ASR system are described as the basis of our study. Previous works on using explicit duration information in ASR are reviewed and various issues on explicit duration modeling are addressed.

2.1. Automatic speech recognition process

Automatic speech recognition is a computation process that generates the most likely word string associated with a given speech waveform. The basic process of ASR is illustrated in Figure 2-1.

The incoming speech is first transformed and represented by a sequence of feature vectors, which is regarded as the incoming patterns for recognition. Subsequently, the similarities between the incoming speech patterns with a set of reference patterns (model) are measured. Each of these reference patterns essentially corresponds to a symbol that represents a specific speech sound. This similarity is measured with pattern matching techniques. The symbol (or sequence of symbols) whose reference pattern has the highest degree of similarity to the incoming speech patterns would give the recognized word string.

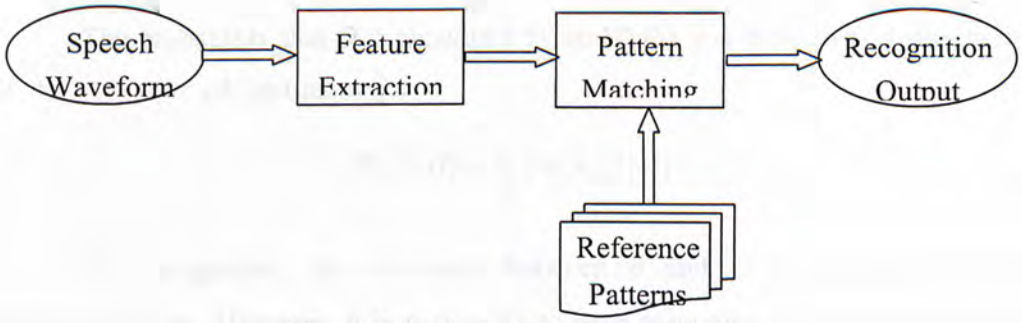


Figure 2-1: The basic process of ASR

In state-of-the-art ASR techniques, the reference patterns in the above process are represented by hidden Markov models.

2.2. HMM for ASR

2.2.1. HMM for ASR

HMM provides a good means to characterize two significant properties of speech signal, namely, temporal structure and spectral properties. Speech signal is time - varying as HMM states evolve from one to the other. On the other hand, speech signal is short-time quasi-stationary. Its short-time spectral properties are described by the state-specific probability density function.

In HMM-based acoustic modeling, it is assumed that the sequence of feature vector $O = \{o_1, o_2, ...o_T\}$ that associated with a speech symbol are generated by an HMM that represents the designated speech sound. At time t when state j is entered, a feature vector o_t is generated by $b_j(o_t)$, the probability density function. Furthermore, the transition from state i to state j is also probabilistic and is governed by the probability a_{ij}. Figure 2-2 illustrates, as an example, a six-state HMM generates the sequence o_1 to o_T with the corresponding state sequence $Q = \{q_1, q_2, ...q_T\}$.

The joint probability that O is generated by an HMM θ with the corresponding state sequence Q is computed as the product of the transition probabilities and the output probabilities, i.e.,

$$P(O,Q|\theta) = b_{q1}(o_1)\prod_{t=2}^{T} a_{t-1\,q_t} b_{q_t}(o_t) \tag{2-1}$$

The probability that O is generated by an HMM θ is then give by summation of $P(O,Q|\theta)$ over all legitimate Q:

$$P(O|\theta) = \sum_{Q} P(O,Q|\theta) \tag{2-2}$$

For recognition, the similarity between θ and O is measured by the likelihood $P(O|\theta)$. However, it is preferable to base recognition on the likelihood of the most likely state sequence. This likelihood generalizes easily to the continuous speech recognition case.

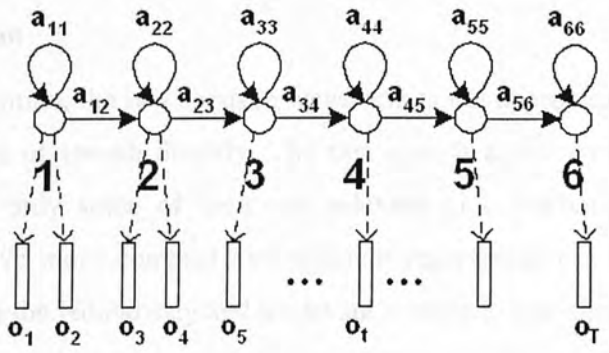


Figure 2-2: A hidden Markov model for speech recognition

2.2.2. HMM-based ASR system

An HMM-based ASR system consists of three key modules as shown in Figure 2-3. They are namely feature extraction, HMM decoding and HMM training. The feature extraction module transforms the speech waveform to feature vector sequence. The HMM decoding (recognition) is a computation process that generates word string associated with speech waveform. The HMMs needed for decoding is estimated by the HMM training module. The training process requires a large amount of speech data.

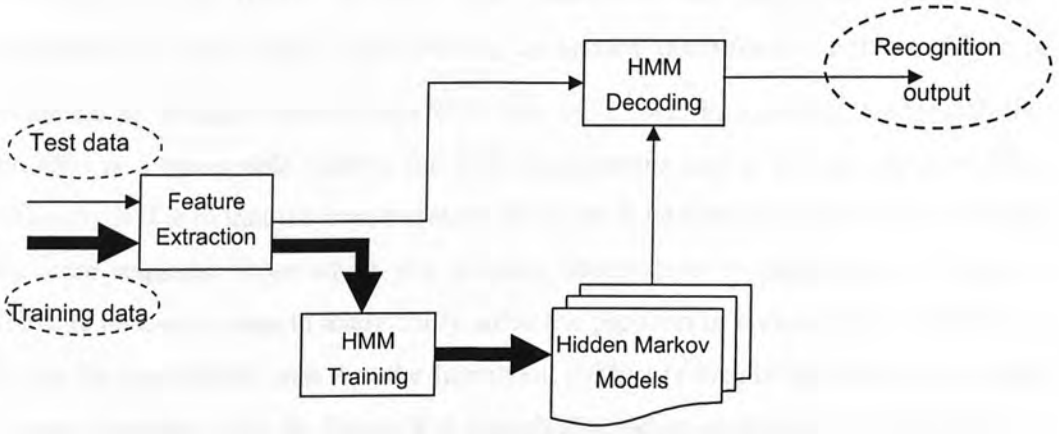


Figure 2-3: Block diagram of HMM-based ASR system

Feature extraction

For speech recognition, the raw acoustic waveform is not appropriate to represent the linguistic contents of speech directly. The raw speech signal contains a variety of information, and only some of them are relevant to a particular ASR task. It's necessary to derive more compact and efficient representations, termed as speech features, to reduce the redundancy and irrelevant content in raw signals. Ideal features for ASR would therefore be ones that can contain only the speech information of interest and remain immune to other sources of acoustic variation in the raw data [8].

The most commonly used speech features include Linear Prediction Cepstral Coefficients (LPC), Mel-frequency Cepstral Coefficients (MFCC) and their extensions. In particular, MFCCs are the choice in many practical applications. They give good discrimination and lend themselves to a number of manipulations [9].

MFCC is based on filter-bank analysis. It employs a number of Mel-scaled band-pass filters to achieve non-linear frequency resolution that resembles how human ear resolves frequencies across the audio spectrum. Since filter bank outputs are highly correlated and hence, a cepstral transformation is applied to de-correlate them. The features then can be modeled by multivariate distributions with diagonal covariance matrixes. Otherwise, full covariance matrixes are computationally very expensive.

HMM training

Training of an HMM involves the estimation and adjustment of model parameter $\hat{\theta}$ $:\{\pi_j, b_j(o_t), a_{ij}\}$according to certain optimization criteria. Given a sequence of acoustic observations $O = \{o_1, o_2, ...o_T\}$, maximizing the probability $P(O|\theta)$ is a reasonable criteria for ASR applications and is commonly used. The difficulty is due to that the maximization has to be done with incomplete data because the state sequence from which the acoustic observation is generated is unknown. There is no known way to analytically solve the problem in a closed form. However, $\hat{\theta}$ can be determined such that the likelihood $P(O|\theta)$ is locally maximized using an iterative procedure like the Baum-Welch method or using gradient techniques[10].

Baum-Welch algorithm is widely used for HMM training. It accomplishes likelihood maximization in a two step procedure, known as "re-estimation". Based on an existing model θ, the first step of the algorithm is to transform the objective function $P(O|\theta)$ into a new auxiliary function $Q(\hat{\theta},\theta)$ .It can be proved that $Q(\hat{\theta},\theta) \geq Q(\hat{\theta},\hat{\theta})$ implies $P(O|\theta) \geq P(O|\hat{\theta})$. The second step of the algorithm is to maximize $Q(\hat{\theta},\theta)$ as function of θ to obtain a higher likelihood $P(O|\theta)$.The two steps iterate until the likelihood $P(O|\theta)$ converges.

HMM decoding

For the expedience of explanation, we start the discussion from the decoding process in the task of isolated-word recognition. Suppose we have a vocabulary of V words to be recognized and each word is modeled by a distinct HMM. Let O represents the incoming speech pattern. The likelihood $P(O|\theta_i)$ is calculated for each HMM. The word whose model has the maximum likelihood to generated the incoming speech patterns are selected as the recognition result. As mentioned earlier, $P(O|\theta_i)$ refer to the likelihood associated with the most likely state sequence for recognition. The most likely state sequence can be efficiently determined by the Viterbi algorithm [11]. Each legitimate state sequence can be represented by a path in the search space formed by HMM states. The Viterbi algorithm essentially finds the optimal path in the search space.

The Viterbi algorithm utilizes the idea of dynamic programming [12]. It divides a complicated problem to sub-problems that can be sequentially solved. In the case of speech recognition, the problem of optimal path search can be divided into sub-problems at frame level, i.e. for each time index t. The sub-problems at frame t is

to find the best path extends to each possible state. We denote *(t,j)* as the token of the partial optimal path extends to state *j* at frame *t*. *L(t,j)* refers to the likelihood of the path. A sub-problem at frame *t* can be solved given the solutions of sub-problems at frame *t-1*. The algorithm is described below:

- Initialization: to solve the sub-problems at frame 1:

$$L(1,j) = \pi_j b_j(o_1) \qquad 1 \le j \le N \tag{2-3}$$

- Recursion (path extension): to solve the sub-problems at Frame *t* given the solutions of sub-problems at Frame *t-1*. The predecessor of *(t,j)* can be any state than can transit to it. Each possible predecessor *(t-1,i)* needs to be evaluated. A path extension decision is made to choose the best predecessor according to the following equation:

$$L(t,j) = \max_{1 \le i \le j} \left\{ L(t-1,i) \times a_{ij} \right\} \times b_j(o_t) \qquad 2 \le t \le T, 1 \le j \le N \tag{2-4}$$

  In each recursion, the path is extended by one frame.

- Termination, to terminate the recursion at frame *T*:

$$P(O \mid \theta_i) = L(T,N) \tag{2-5}$$

Subsequently, the best matching model is obtained by comparing $P(O|\theta_i)$ for all the HMMs:

$$\text{recognized word} = \arg\max_i P(O \mid \theta_i) \tag{2-6}$$

In this procedure, the choice of step size for path extension is important. The step size is also known as speech dynamics. Under assumption of HMM, we can choose a frame.

For the ease of computation, log likelihood is commonly used for calculating path likelihood. In the subsequent study, log likelihood is referred to as score.

For connected-word recognition, HMMs that represent different speech units are connected. In this case, path extension can be within or across models. There are a couple of realizations of dynamic programming algorithm for connected word recognition such as two-level dynamic programming approach [13], level-building approach [14] and one-stage approach [15]-[17]. In the two-level algorithms, the time

span of a word was chosen as the step size for path extension. In one-stage algorithm, a frame was chosen as the step size for path extension. One-stage algorithm is essentially the Viterbi algorithm for connected word recognition.

For large-vocabulary continuous speech recognition (LVCSR) or grammar-constrained connected word recognition, in addition to the HMM-based acoustic models, the recognition process needs to invoke the linguistic knowledge from the lexicon and the language model in order to produce a meaningful word sequence. These knowledge sources are different with HMM. The probability from these information sources can not obtained frame by frame. To use these knowledge sources, path extension need to cover a much longer time span. In this case, the search process becomes computationally very expensive and sometimes even unaffordable. To alleviate this problem, sub-optimal algorithms have been developed for efficient decoding at the cost of performance degradation. Actually, explicit duration of speech sounds is such kind of knowledge source. Similar problem will be encountered in using of duration information.

2.3. General approaches to explicit duration modeling

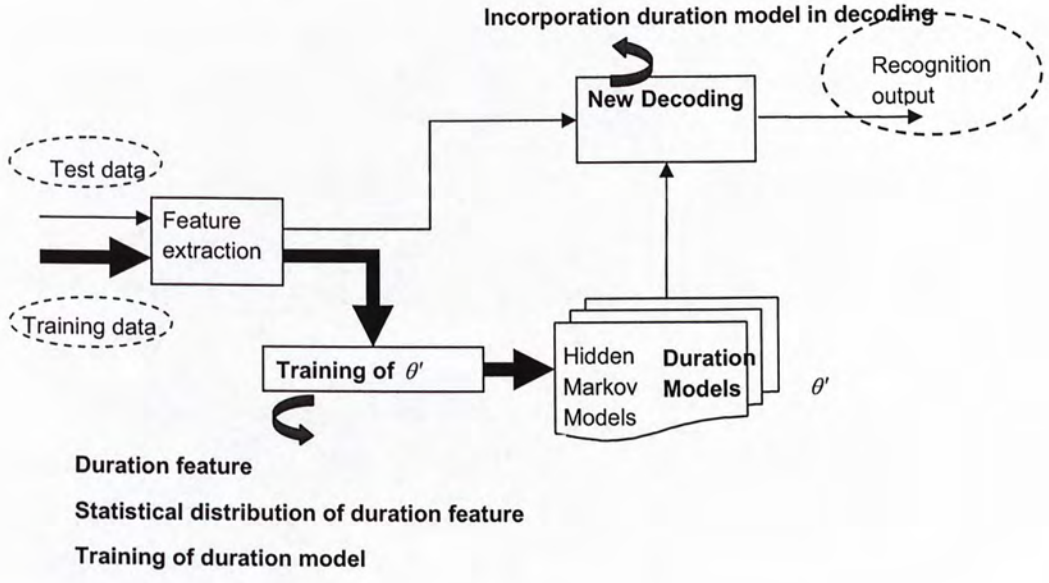


Figure 2-4: Integrating explicit duration modeling into ASR

In general, the incorporation of explicit duration knowledge into to the ASR involves two steps:

1) Development of duration models

In the construction of the model θ' with duration models included, there are three major issues. Firstly, appropriate duration feature need to be identified for duration modeling. Secondly, a reasonable distribution has to be assumed for statistical characterization the duration feature. Lastly, an algorithm of estimating parameters of θ' needs to be developed.

2) Incorporate the duration model into HMM-based recognition

For the use of duration models, a new decoding algorithm is needed.

2.3.1. Explicit duration modeling

2.3.1.1 Different duration features

Duration can be measured and modeled at segments of various lengths. The measurement of duration information is referred to as duration feature. In an HMM-based system, HMM is used to model and segment speech signals. Basically, measurement of state-level time duration and model-level time duration can be obtained. Model-level duration essentially gives the duration of a linguistic unit. State-level duration contains information about the length of sub-segments of a linguistic unit.

Modeling of duration information at state and model level has been intensively investigated. At state level, the time duration of state was explicitly modeled in [5]-[7] [18] [19]. It was aimed to complement the problematic implicit state duration modeling by HMM. At model level, the duration information are expected to be meaningful since an HMM corresponds to a linguistic unit. The time duration of a model was explicitly modeled in [20]-[22]. In contrast to absolute duration, there is another kind of duration feature, namely relative duration. Relative duration of a speech segment reflect possible internal adjustment between the sub-components of this segment. It is expected to be more stable and robust against speaking rate variation. In [10] [22], relative state duration of an HMM was explicitly modeled. In Power's experiments, recognition performance of using relative duration model is better than that of using duration model for state duration and model duration.

For accurate modeling of absolute state duration, hidden semi-Markov model (HSMM), or variable duration HMM was proposed. HSMM is obtained by replacing the underlying Markov model in a conventional HMM by a semi-Markov model [5]-[7]. In other words, the self-transition probabilities are ignored and the state occupancy is defined by a state duration distribution.

Another two frameworks, ESHMM and IHMM were also for accurate state duration modeling. They implicitly model the state duration other than explicit modeling. ESHMM [24] was obtained by replacing each state in a conventional HMM with another HMM, referred to as a sub-HMM. The output pdfs of each sub-

HMM are identical to the output pdf of the state in the original model which they replace. The resultant ESHMM is said to be functionally equivalent to a HSMM in which the duration pdf for a given state is the overall duration pdf of the associated sub-HMM. In [25], IHMM was proposed. It uses time-dependent state transitions rather than the time-independent one in HMM.

Generally speaking, by using duration models with appropriate duration features, speech recognition performance is always improved to some extent. In particular, the normalized duration features have been shown very effective. However, there has not been a widely accepted conclusion on which duration feature is necessary more useful than others. The effectiveness of duration model depends on many factors including the task specification, the quality of acoustic data, the way how duration feature is modeled and the way how the duration model is used.

Computation expensiveness cost of using different duration feature is different. This will be elaborated in details in section 2.3.3.

2.3.1.2 Statistical distribution of duration features

Duration model is typically a statistical model that describes the distribution of a specific duration feature. Empirical distribution, i.e., the actual distribution duration measured from real speech data, is the most accurate estimation that can be achieved. It was used in [6] [22]. However, it is a non-parametric model. Each probability should be estimated. A huge amount of training data is necessary for accurate modeling.

Parametric distribution can be used. Only a few free parameters need to be estimated for parametric model. Poisson assumption was suggested to modeling state duration in [5].

$$P(d=k)=e^{-\lambda_i}\lambda_i^{\,k}/k! \tag{2-7}$$

with parameter λ, which is the expected duration of ith state. It has only 1 free parameter. This may limit its flexibility to model distribution of various shapes. In particular, its variance is always equal to the mean. This is inappropriate in most cases and people resort to other parametric distributions. Some continuous distribution assumptions were adopted. Although it is seems inappropriate to model discrete

values, it does not affect the effectiveness of duration models. Proposal included the Gaussian family

$$p_i(d) = \mathrm{N}(d, u_i, \sigma_i^2) \tag{2-8}$$

with parameters μ_i and σ_i, or the Gamma family

$$p_i(d) = \frac{x^{a-1}e^{-x/b}}{b^a \Gamma(a)} \tag{2-9}$$

with parameters a and b, and with mean ab and variance ab^2 [7].

The Gamma distribution was further used for model-level duration modeling [20] [26]. The Gaussian distribution was proposed for model-level duration modeling [21][27].

The Gamma distribution has a very flexible shape that can be from very skew to very symmetric. In addition, the Gamma distribution assigns zero probability to negative duration. David Burshtein compared the Gamma distribution and the Gaussian distribution for duration modeling. In his experiment, he showed that the Gamma distribution is almost always closer to the empirical distribution than the Gaussian distribution for the modeling of both state duration and model duration. Rong Dong found that the Gamma distribution is superior to the Gaussian distribution and the Poisson distribution in terms of recognition accuracy.

The mixture Gaussian distribution is able to model arbitrary distribution if there are sufficient mixture components. Therefore, it is capable to model any kind of duration features. Sufficient data is necessary because more free parameters need to be estimated. In [28], word duration feature, which is a vector comprising of the durations of the individual phones in a word, was modeled by the mixture Gaussian distribution.

2.3.1.3 Irrelevant factors that affect duration properties

For durations to be useful information source in ASR, it is desirable if they only contain only speech information of interest and immune to other source of acoustic variation. However, many irrelevant factors may affect the absolute duration of a certain speech segment like a phone or a word. They include the phonetic context, the

prosodic state, speaking rate, personal style of speaking, etc. The variability of above factors leads to large variation of the absolute durations. This made absolute duration not a good feature to use. The study presented in [27] reached a discouraging but realistic conclusion that many duration features are not as consistent as expected and only some of them would be of potential importance for speech recognition purpose. Several approaches can be employed to alleviate this problem. They can be divided into two major categories.

The first category of approaches is to build context-dependent duration models. Contextual factors may include syllable or phone context in [1] [29], speaking rate category context in [30] [31], prosodic state context in [27] [30] - [32].

The second category of methods is to develop robust duration feature. For speaking rate to be a major concern of duration, several features which are expected to be less sensitive to variation in speaking rate were proposed. Normalized state duration was proposed in [10] [22]. Penalty of relative duration of neighboring syllables and syllable structural penalty were proposed in [33]. Speaking rate normalized duration was proposed [26] [27].

Context-dependent duration modeling and modeling of normalized duration by context were found to be effective in improving ASR performance. The difficulty for using them is on the estimation or identification of context. Many contexts are phrase-level or utterance-level information, which cannot be estimated as recognition proceeds. A possible solution is two-pass decoding.

2.3.2. Training of duration model

Like HMM training, training of model θ' involves adjusting model parameters to maximize the likelihood. If all the parameters of model θ' are trained simultaneously, this is referred to as one-pass [2] training. One-pass training has its limitations. Multi-pass training can be used as an alternative approach.

[2] Note that one-pass doesn't mean a single iteration..

2.3.2.1 One-pass training

Among many previous studies, the training of HSMM [5] - [7] employed the one-pass training technique. These training algorithms for HSMM are analogous to the Baum-Welch re-estimation procedures. They guarantee a local optimum solution.

However, one-pass training is not applicable when sophisticated duration features such as relative duration are modeled. Another limitation of this approach is due to the expensive computation. The cost of including duration distribution was rather high. The cost of the increased computation tended to make the one-pass training techniques infeasible in many applications.

2.3.2.2 Multi-pass training

Multi-pass training can be used to simplify the training process of duration models and to model sophisticated duration features. It was employed in many works [19]-[22] [26]-[36].

Generally, the multi-pass training is done by the following steps.

1) The acoustic models (HMMs) are trained.
2) State or word segmentations of training data are obtained by forced alignment with the trained HMMs. Subsequently, duration information are obtained from the segmentations.
3) Duration features concerned are chosen as training samples
4) Finally, the parameters of duration models are estimated over training samples using maximum likelihood techniques.

Multi-pass training approach has its limitation in nature. Parameters of acoustic model are estimated with the absence of parameters of duration model. However, experiences show that its recognition performance is essentially as good as that obtained using the theoretically correct duration model [10].

2.3.3. Incorporation of duration model in decoding

The HMM decoding problem involves searching for the optimal path in an search space formed by HMM states. The probabilistic score of a path is contributed by state output probability and state transition probability. The optimal path is usually determined by Viterbi algorithm.

When duration model is used, an additional score computed based on duration models can be added to the path score. If the recognized speech units in this path exhibit unreasonable duration patterns, the duration score may be used to penalize the path. This is the general idea of using duration modeling to assist HMM-based recognition. Depending on how the duration score is added to the path, there are one-pass decoding method and multi-pass decoding method.

2.3.3.1 One-pass decoding

In one-pass decoding, the solution of the optimal path is obtained in a single pass of search with all the knowledge sources presented. The solution may be optimal or sub-optimal.

Optimal solution

As mentioned earlier, knowledge sources such as explicit duration model are different with HMM. To incorporate these knowledge sources in recognition processes, the dynamic programming algorithm can still be used. However, each path extension needs to cover much longer time span. With different levels of duration models to be incorporated, the decoding algorithm should be changed accordingly for optimal search. If state duration was modeled and the speech dynamics been considered is a time span of an HMM state. In [21], the digit duration was modeled and the speech dynamics been considered is a time span of a whole-word HMM. If D is the maximum duration of speech dynamics considered, a D-fold increase in computation would be required for the one-pass decoding.

If explicit duration modeling is going to be used in an LVCSR task or speech segments for duration modeling are long, substantial computation cost is required by

one-pass decoding. It might be necessary or even compelling use one-pass suboptimal solutions or multi-pass decoding strategies.

Suboptimal solution

In [22] [36], it is suggested to apply duration penalties at each inter-model transitions in Viterbi algorithm. In [20] a different modification of Viterbi algorithm was proposed for English connected digit recognition. It was suggested to apply duration penalties gradually. Rong Dong [26] followed Burstein's approach in a mandarin connected-digit task. These methods achieved significant improvement in recognition performance.

However, this kind of modification theoretically violates the principle of dynamic programming. Duration score does not involved in all of the path extension decisions. In other words, the path is locally not optimal since the decision is made on the subset of knowledge sources. According to Bellman's principle of optimality [11], an optimal policy has the property that, whatever the initial state and decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. Locally non-optimum path will not be global optimum. Therefore, the kind of modification only leads to a sub-optimal solution.

2.3.3.2 Multi-pass decoding

One major kind of multi-pass decoding is the post-processing approach. Generally, N-best list or word lattice are generated from the first pass of search to get a reduced search space. Duration information at various levels can be readily obtained on the basis of N-best list or word lattice. Moreover, speaking rate or other speech context can be readily obtained. Context information can be used for choosing appropriate context-dependent duration model or normalizing duration features. In the second pass of search, obtained duration information contribute to search optimal path in the reduced search space.

Post-processing approaches were adopted in many recognition tasks with explicit duration modeling. [27] [28][37] [38] [30][31].

Multi-pass decoding has the potential to introduce inadmissible pruning. Pruning decisions made in the first pass are based on simplified models (KSs). If a pruning decision is erroneous, the pruning error can not be recovered by successive passes.

Weighted use of duration model

Weighting is commonly used to balance the contribution of multiple information sources such as multiple models, multiple features and multiple modals when they are combined used for speech recognition. The magnitude of weighting factor depends on the relative importance and relative OOM (order of magnitude) from one information source to others. Weighted use of tone information is employed in [39]. Weighted use of Static and Dynamic MFCC features is proposed in [8]. Weighted use of utterance information and speaker information for speaker authentication is proposed in [40].Weighted use of visual information and acoustic information is proposed in [41]. Use of weighting factor shows significant improvement in recognition performance.

Most of the work on explicit duration modeling also employs a weighting factor for the effective contribution of duration models. It was found to be effective to achieve further performance improvement.

Chapter 3

Cantonese connected-digit recognition

In this chapter, HMM based connected-digit recognition will be discussed. We adopt the approach of whole-word modeling, i.e. each Cantonese digit is modeled by a dedicated HMM. A baseline system will be described. It will serve as the basis for our subsequent work on applying duration information. The recognition performance of this baseline system will be analyzed. In particular, the potential benefit of using duration information will be addressed.

3.1. Cantonese connected digit recognition

Connected-digit recognition has many practical applications such as voice dialing, automatic data entry, credit card number entry, pin personal identification, entry of access codes for transactions, etc. These applications generally require very high accuracy. Despite its limited vocabulary size, it is difficult to attain the satisfactory accuracy level because the combination of digits is unrestricted. Previous efforts on high performance connected-digit recognizer were reported in [37] [39] [46] - [49].

3.1.1. Phonetics of Cantonese and Cantonese digit

Cantonese is one of major dialects of Chinese. It is the mother tongue of over 60 million populations in Southern China and Hong Kong. Cantonese is also commonly used in overseas Chinese communities.

As a spoken language, Cantonese is quite different from Western languages. Cantonese is monosyllabic. As shown in Table 3-1, each Cantonese digit is pronounced as a monosyllable sound. Cantonese is tonal. Each syllable is associated

with a specific lexical tone among six Cantonese tones. As seen in Table 3-2, LSHK[3] labels the lexicon with an Arabic digit suffix. Cantonese digits cover all the six tones.

Table 3-1: Phonetic transcriptions of 10 Cantonese digits

Digit	IPA	LSHK
0	lɪŋ	ling4
1	jɐt	jat1
2	ji	ji6
3	sam	saam1
4	sei	sei3
5	ŋ	ng5
6	luk	luk6
7	$ts^hɐt$	cat1
8	pat	baat3
9	kɐu	gau2

As shown in Figure 3-1, each Cantonese syllable is seen as the concatenation of two types of phonological units: Initial and Final. Initials of Cantonese digits have a diversity of semi-vowel (4 cases), fricative (2 cases), affricative (1 case) as well as plosive (2 cases) and no initial (1 case). Finals of Cantonese digits exhibit a large diversity too. They include vowel (1 case), diphthong (2 cases), vowel with nasal coda (2 cases), vowel with stop coda (3 cases) and syllabic nasal (1 case).

<table>
<tr><td colspan="3">BASE SYLLABLE</td></tr>
<tr><td>Initial</td><td colspan="2">Final</td></tr>
<tr><td>[onset]</td><td>Nucleus</td><td>[coda]</td></tr>
</table>

Figure 3-1: Structure of a Cantonese syllable ([] means optional)

The syllable compositions of Cantonese digits are generally very simple. This makes Cantonese connected-digit task relatively difficult comparing with other languages in which the phonetic compositions of spoken digits are more complicated. In English connected-digit task, the major insertions and deletions are caused by the

[3] Transcription scheme developed by Linguistic Society of Hong Kong

vowel-only digit "oh". Other digits are sandwiched by two consonants, like digit "six" or "seven", preceded by two consonants, like digit "two", or succeeded by a consonant, like digit "eight". In machine recognition of continuously spoken digits, the consonant onset or coda help to alleviate inter-digit coarticulation, hence decreasing acoustic confusion [49]. In Cantonese, digits do not contain as many consonant onsets and codas. In particular, digits "2" and "5" can be regarded as a single vowel segment and a single nasal segment respectively. The two mono-phone digits are phonetically very similar to the coda or tail part of other digits, e.g. "0", "3" and "4". Consequently, they will be likely to be co-articulated with other digits and cause severe confusion in recognition. It was reported that "one of the major sources of errors was due to frequent insertions of digit "5", pronounced as a mono-syllabic nasal[ng5], which may be confused with and treated as part of nasal coda in the digits "0" and "3" [50].

Similar observations were reported on Mandarin connected-digit recognition tasks [26] [33] [49] and Korean connected-digit recognition tasks [21] [29]. In Mandarin, all digits are mono-syllabic. Two of them consist of only single vowels, i.e., digit "1" and digit "5". The digit "2" also has a heavily rhotacized vowel. These three digits, due to their short duration and vowel-only structure, are strongly co-articulated with adjacent digits. There are insertions, e.g., digit "7" is often recognized as double digits "7 1", and deletions, e.g., repeated digits "5 5" is often recognized as a single "5". It was found that most of the insertion errors and deletion errors are related to those three digits [49].

Korean is mono-syllabic, too. Two mono-phonemic digits "2" and "5" are said to be involved in most insertion and deletion errors [21].

Explicit duration modeling techniques were applied to Korean and mandarin connected-digit recognition tasks to help reducing the error rate. Significant improvement was achieved.

3.2. The baseline system

The baseline system is an HMM-based ASR system established for Cantonese connected digit recognition. It consists of the feature extraction part, the HMM training part and the HMM decoding part. The first two parts are done by HTK[4].

3.2.1. Speech corpus

All our experiments on Cantonese connected-digit recognition are based on the CUDIGIT speech database, which was developed at the Digital Signal Processing Laboratory, the Chinese University of Hong Kong [51]-[52]. CUDIGIT is part of a whole series of Cantonese spoken language corpora developed for speech processing.

CUDIGIT is a collection of Cantonese digit strings. The data collected are all read speech. Speakers were prompted with the digit strings one by one, with Chinese characters and Arabic digits displayed in parallel. The speech data are of low noise good quality. The recordings were carried out in a confined room providing a closed silent recording environment. The recording was done using high-quality microphone. The signal was passed through a pre-amplification mixer to the DAT recorder for real time A/D conversion at 48 kHz with 16-bit resolution. The digital data was then down-sampled to 16 kHz. All data and annotations in this corpus were manually verified in a two-stage process.

CUDIGIT contains exhaustive permutation of digit strings from one to four syllables. In addition, there are also randomly generated strings of 7, 8 and 16 digit long. There were 25 male speakers and 25 female speakers being recorded. Each speaker spoke about 570 utterances. The average utterance length is 3.65 digits. The reading materials can be classified into seven sections where each section is broken down into partitions for sharing amount a large number of speakers. The table below shows the partition of the sections and the corresponding amount of digit strings per speaker.

---

[4] HTK is a toolkit for building HMMs. It is commonly used for building HMM-based recognition system.

Table 3-2: Distribution of digit string length for each speaker in CUDIGIT

Sections	Contents	Partition	Amount
I	Calibration	1	10
II	Single Digit	1	10
III	Double Digit	1	100
IV	Triple Digit	5	200
V	4-Digit	50	200
VI	Random 7-Digit	Per speaker	20
VII	Random 8-Digit	Per speaker	20
VIII	Random 16-Digit	Per speaker	10
	Number of string per speaker	570	

Recognition experiments were carried out on male and female data separately. For the experiments on male speech, the data include 11,387 training utterances from 20 speakers and 2,847 test utterances from another 5 speakers. Female speech data include 11,393 training utterances from 20 speakers and 2,848 test utterances from another 5 speakers.

3.2.2. Feature extraction

We choose to use MFCC features for recognition. The feature parameters are obtained by the following procedures:

- Frame blocking: to divide the speech signal into consecutive frames for short-time spectral analysis. A length of 20 msec is used for each frame. Adjacent frames overlap by 10 msec.
- Pre-emphasis: to boost high frequency, the pre-emphasis factor is set to 0.97 by following the empirical value employed in HTK.
- Hamming windowing: to minimize the effect of discontinuities at the edges of a frame.
- DFT: to transform acoustic waveform to frequency domain.

- Mel-scale filtering: to approximate the non-linear frequency resolution of human ear. In our experiments, the number of filter banks is set to 32.
- Log compression: to prepare for separation of spectral envelope and details in cepstra domain.
- DCT: to de-correlate the Mel-scale filter-bank output. By DCT, Mel-scale filter bank output is transformed to cepstra domain. Number of cepstra coefficients is set to 12. Liftering is applied to rescale the cepstra coefficients to have similar magnitude with the liftering factor 22. The resultant cepstra coefficients form a vector, which is the MFCC feature vector.
- Cepstral mean subtraction: to compensate for long-term spectral effects such as those caused by different microphones and audio channels.
- Energy term: to augment the MFCC feature vector by energy information.
- Dynamic features: to augment the basic MFCC vector by adding time derivatives. It is aimed to complement the problematic assumption of the HMM that each acoustic feature vector is independent. First order and second order derivatives are used.

For each short-time frame, the resultant acoustic feature vector has 39 components, which include 12 Mel-Frequency Cepstral Coefficient (MFCC), log-energy term, and their first and second order derivatives.

3.2.3. HMM models

Each Cantonese digit was modeled by a whole-word HMM which consists of 6 states. These states are of left-to-right topology without state skipping transition. Each state is defined with an observation probability distribution, which is represented by an 8-mixture components Gaussian. Diagonal covariance matrices are assumed. There are also a six-state "silence" model and a one-state "sp" model to model to non-speech signal. The "silence" model is of left-to-right topology with optional state skipping. It is aimed to model the background silence. The single state of "sp" model is tied with the third state of "silence model". It is aimed to model the transition effect of speech

such as short pause and noise bursts [9]. "sp" model is shown to be useful in reducing insertion errors [36]. Without 'sp' model, those transition segments are more likely to be recognized as digit.

The HMM parameters were estimated by the Baum-Welch algorithm with HTK tools, using the features extracted from the training speech and a transcription file of training speech.

3.2.4. HMM decoding

A Viterbi decoder was used for connected-digit recognition. No grammar constraint was imposed. Details of that decoder will be given in Chapter 4.

3.3. Baseline performance and error analysis

3.3.1. Recognition performance

Table 3-3 gives the recognition performance of the baseline system.

Table 3-3: Performance of the baseline recognition system

baseline	Dig acc (%)	Sen acc(%)	Deletion	substitution	insertion
Male	96.77	89.50	87	69	181
Female	98.12	93.43	30	44	121

The above results were obtained using a decoding algorithm implemented by the author. The same tests were also done using the decoder in the HTK and the results are shown as in Table 3-4. Seen from the two tables, our baseline system performs almost the same as the HTK.

Table 3-4: Recognition performance by HTK

Baseline	Dig acc (%)	Sen acc(%)	Deletion	Substitution	insertion
Male	96.73	89.36	87	69	184
Female	98.09	93.29	30	43	126

Distribution of different types of errors

As shown in Table 3-5, insertion and deletion errors account for about 80% of the recognition errors for both male and female experiments.

Table 3-5: Percentage of deletion, substitution and insertion in the baseline system

	Del%	Sub%	Ins%
Male	25.6	20.3	54.1
Female	15.4	22.6	62.1

3.3.2. Performance for different speaking rates

Two speaking rate[5] thresholds were chosen so that the training data was evenly divided to three speaking rate categories. Using the same threshold, the test data can be also divided to three categories. The recognition performance on each category are given in Table 3-6 and Table 3-7 for male and female respectively.

[5] Speaking rate will be defined in Chapter 4.

Table 3-6: Baseline recognition performance for different speaking rates (male)

	Digit acc%	Sen acc%	Del	Sub	INS
Fast	96.00	85.99	67	54	29
Medial	97.57	91.34	16	15	81
Slow	96.37	91.06	3	0	71

Table 3-7: Baseline Recognition performance for different speaking rates (female)

	Digit acc%	Sen acc%	Del	Sub	INS
Fast	98.08	92.88	25	33	32
Medial	98.24	93.51	5	9	29
Slow	98.11	94.05	0	2	60

Clearly, the recognition performance depends on speaking rate. In terms of accuracy, the best performance is obtained for medial category. It is observed more obviously on male data. In terms of distribution of different errors, fast category tends to have more deletion and substitutions. This is possibly due to the extremely heavy co-articulation. Medial and slow category tends to has more insertions. The possible reason is that long speech segments are not preferred by the implicit duration models in conventional HMMs.

Distribution of recognition errors among utterances

To understand more about the relation between string error rate and digit error rate, the distribution of errors among different utterances is analyzed. The statistics are given as in Table 3-8. As seen in the table, most of wrongly recognized utterances contain only one digit error. In this case, digit error rate multiplies the average number of digits per utterance approximates sentence error rate.

Table 3-8: Distribution of recognition errors among utterances

	Utterances with 1 error	Utterance with 2 errors	Utterance with 3 errors	Utterance with 7 errors
Male	266	29	3	1
Female	180	6	1	N/A

3.3.3. Confusion matrix

Table 3-9: Confusion matrix of baseline recognition result (male)

	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	del	%c/%e
'0'	1028	4	0	0	0	7	0	0	0	0	2	98.9/0.1
'1'	2	1029	0	0	1	0	1	2	2	1	**12**	99.1/0.1
'2'	0	0	985	0	0	1	0	0	0	0	**26**	99.6/0.0
'3'	0	0	0	1022	0	0	0	1	4	0	0	99.5/0.0
'4'	0	1	0	0	1066	0	0	1	0	0	0	99.8/0.0
'5'	3	0	1	0	0	972	1	0	0	0	**39**	99.5/0.0
'6'	0	1	0	0	0	0	1025	0	0	3	4	99.6/0.0
'7'	0	1	0	2	3	0	0	1065	0	0	2	99.4/0.1
'8'	0	1	0	0	0	0	0	0	1050	1	1	99.8/0.0
'9'	0	1	0	0	0	0	5	0	**15**	991	0	97.9/0.2
ins	**34**	**16**	**23**	1	2	**82**	10	4	8	1		

As seen in above confusion matrix, some recognition error frequently occurs. For insertion and deletion, "5" "1" "2" "6" "0" are involved in most cases. For substitution, "9" and "8" are easily confused pair. In the following, we attempt to

explain the error patterns observed on the baseline system. Meanwhile, we will discuss about how duration information can be made helpful to deal with these recognition errors.

Recognition error in relation to the digit "2"

The Cantonese digit "2" is involved mainly in error patterns: 1) A single "2" is split to cause an insertion; 2) A digit "4" is split to "4 2" to cause an insertion. 3) Repetitions of "2" are merged to cause a deletion; and 4) "4" and "2" are merged to cause a deletion. The frequency counts of these errors are given as in Table 3-10.

Table 3-10: Patterns of recognition errors related to digit "2"

Error pattern	"2 2"→ "2"	"2" → "2 2"	"4" → "4 2"	"4 2" → "4"
Frequency	17	3	17	6

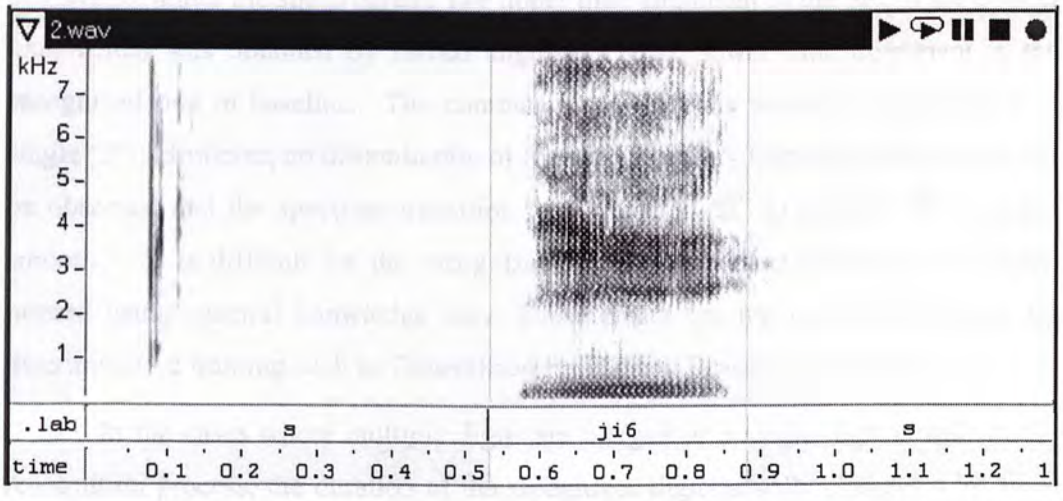


Figure 3-2: Spectrogram plot of Cantonese digit "2"

A digit "2" can be approximately regarded as a single vowel segment. As shown in Figure 3-2, the formant trajectories of "2"are flat over time. When multiple "2" are repeated at a relatively fast speaking rate, they will co-articulate closely with each other, and consequently, there will be great ambiguity on the exact number of digits that are actually spoken. An instance of such a situation is show in Figure 3-3.

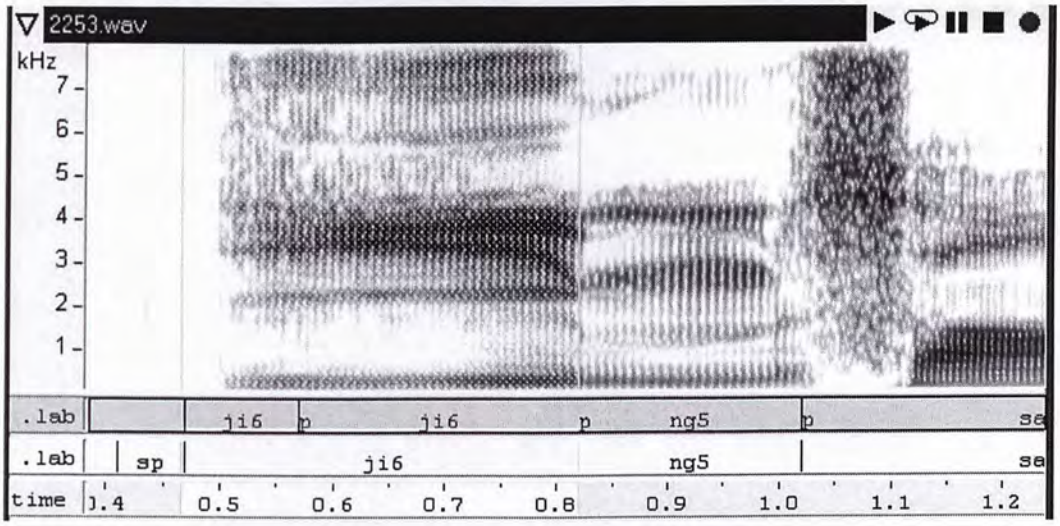


Figure 3-3: Spectrogram plot that explains the deletion error "22" →"2"

Figure 3-3 gives the spectrogram of an utterance. There are two time alignments under the spectrogram. The upper time alignment in the plot is the correct one, which was obtained by forced alignment. The lower time alignment is the recognized one in baseline. The combination of "22" is wrongly recognized as a single "2". However, no discontinuity of formant trajectory from the spectrogram can be observed and the spectrum transition from the digit "2" to another "2" is rather smooth. It is difficult for the recognizer to judge the exact number of "2" been uttered using spectral knowledge only. These errors can not be corrected even by discriminative training such as Generalized Probability Descent method [21].

In the cases where multiple digits are merged or a single digit is split in the recognition process, the duration of the recognized digits usually changes a lot from what it used to be. The duration would be much shorter or longer than those in the normal cases. In this situation, duration knowledge would be helpful to correct the error.

"2" is phonetically similar to the tail part of digit "4". Figure 3-4 shows a typical spectrogram of "4".

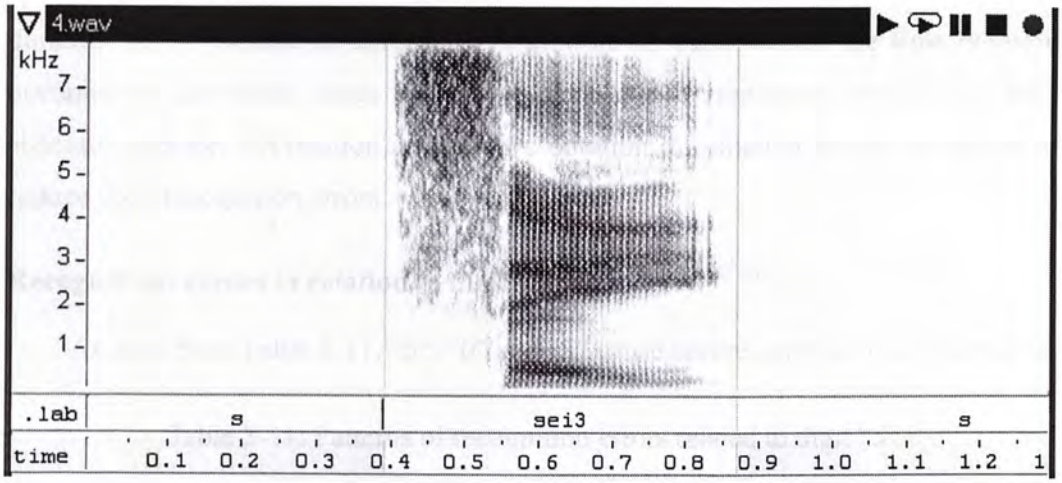


Figure 3-4: Spectrogram plot of Cantonese digit "4"

Therefore, "2" may be confused with and treated as part of "4". The possible errors will be that a digit "4" is recognized as combination of "42" or a "42" combination is recognized as "4" only. Figure 3-5 shows an instance of such an insertion. "2" is exactly inserted at the end of "4".

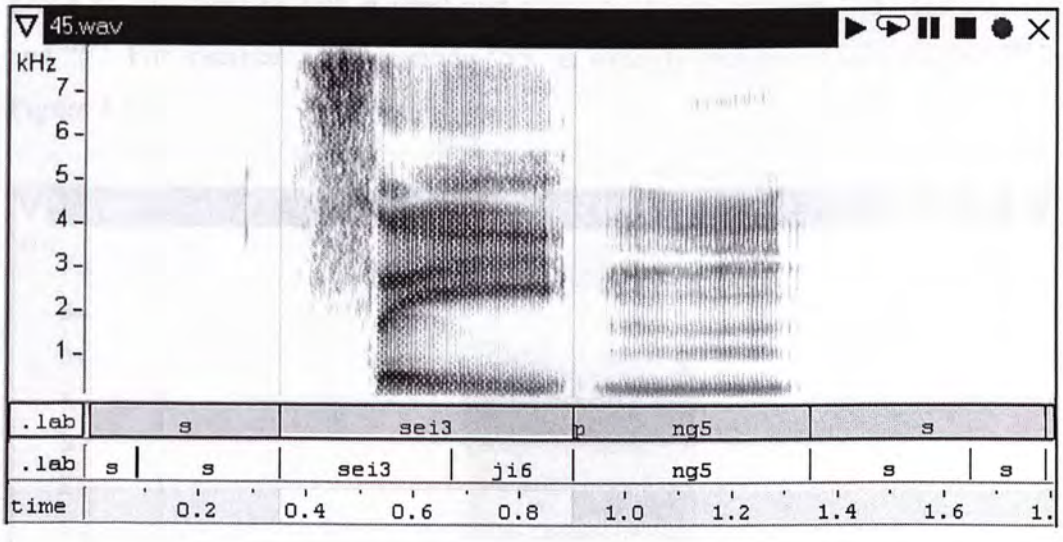


Figure 3-5: Spectrogram plot that explains the insertion "4" →"42"

When a "42"combination is misrecognized as "4" or vice versa, absolute duration of recognized digit(s) would be unreasonable. In addition to absolute duration, there is another kind of duration, relative duration of sub-segments of a digit. When "42" combination is misrecognized as "4" or vice versa, the relative

duration of "4" would be unreasonable as well. In other words, the time duration occupied by individual states of "4" would be not in appropriate proportion. Both absolute duration information and relative duration information would be helpful to reduce such recognition errors.

Recognitions errors in relation to digit "5"

As seen from Table 3-11, "5", "0" and "3" cause severe confusion in recognition.

Table 3-11: Patterns of recognition errors related to digit "5"

Error pattern	"3" →"35"	"0"→"0 5"	"3 5"→"3"	"05"→"0"	"5"→"55"	"55"→"5"
Frequency	30	21	5	11	17	9

"5" (/ng/) is a nasal-only syllable. Sometimes it is pronounced as another nasal-only syllable /m/. A typical spectrogram of "5" is shown in Figure 3-6. Formant trajectories of "5" are flat over time. Therefore, similar to the situation of "2", "5" is prone to co-articulate with it itself and cause confusion between "55" combination and "5". For instance, a combination "55" is wrongly recognized as a single "5" in Figure 3-7.

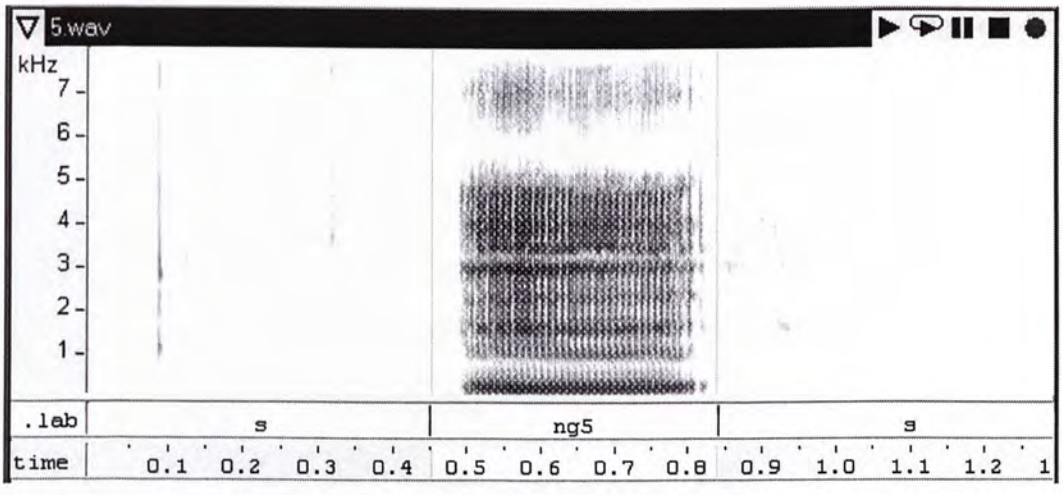


Figure 3-6: Spectrogram plot of Cantonese digit "5"

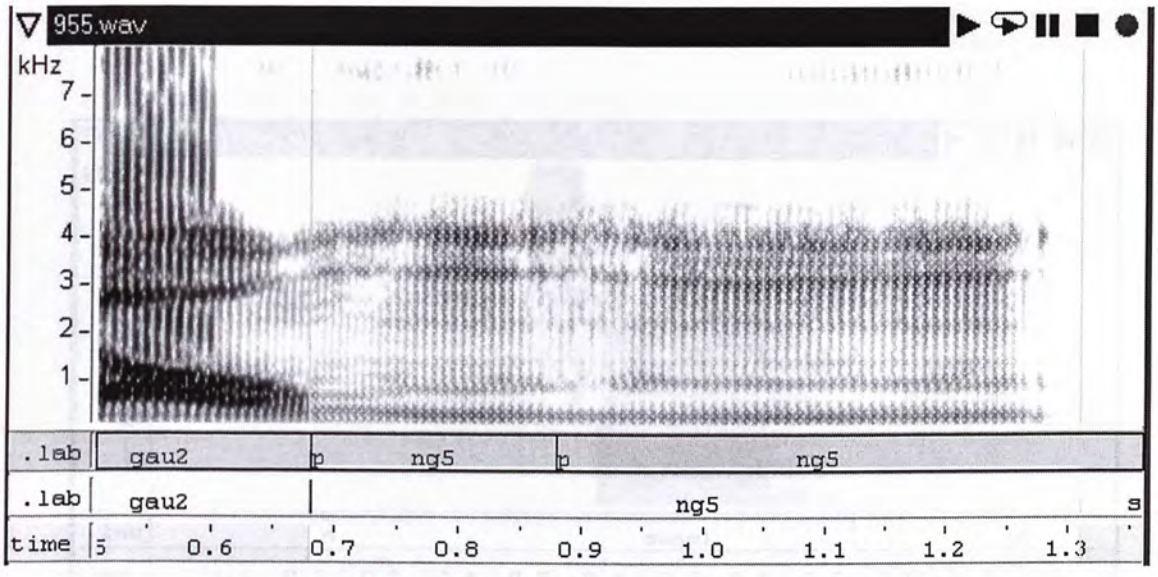


Figure 3-7: Spectrogram plot that explains the deletion "55" →"5"

When "55" combination is misrecognized as "5" or vice versa, the absolute duration of the recognized digits would changes a lot from what it used to be and become unreasonable. Prior knowledge about time duration of "5" would be useful.

Confusion between "35" combination and "3" and confusion between "05" combination and "0" are due to digit "3" and "0" being end with nasal sound, which is also the only sound of digit "5". "5" would be confused with and treated as part of "0" ("3"). The typical spectrogram of '0' and '3' are plotted in Figure 3-8 and Figure 3-9.

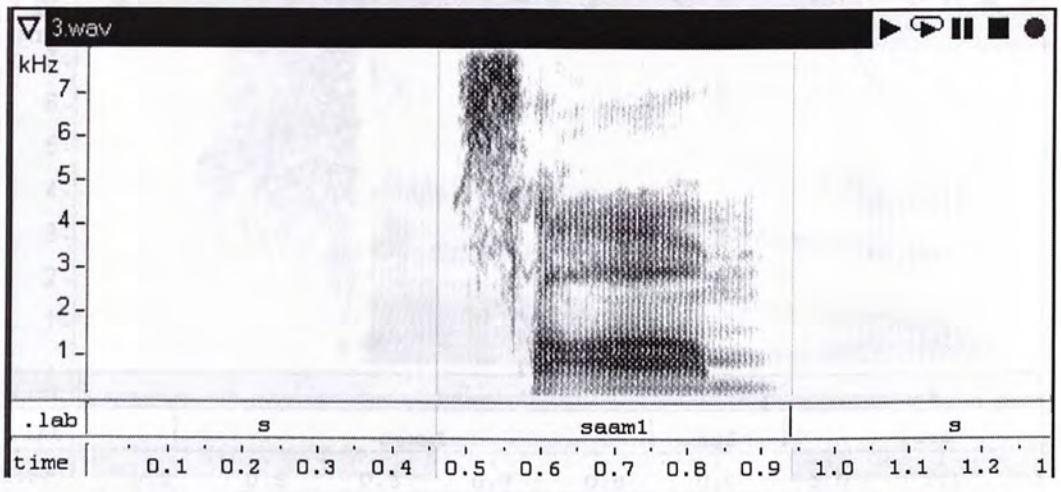


Figure 3-8: Spectrogram plot of Cantonese digit "3"

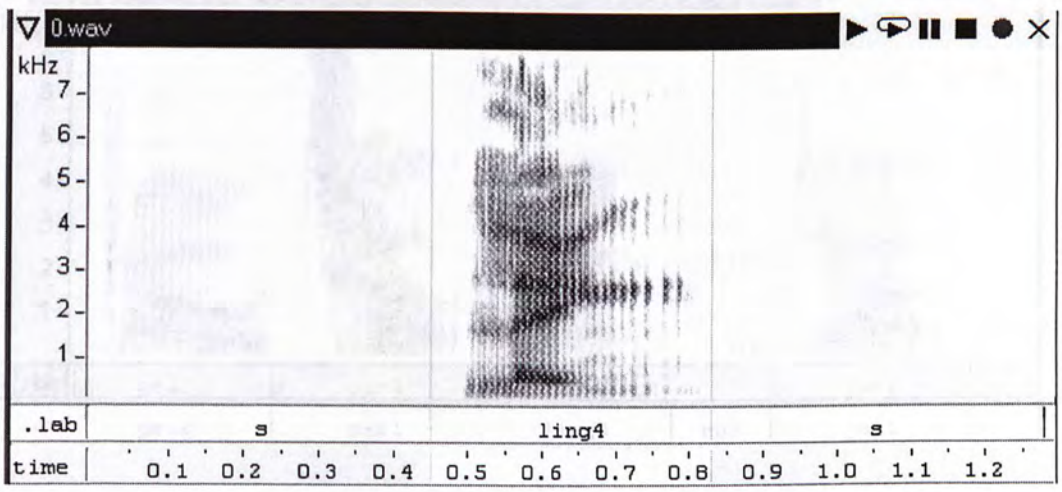


Figure 3-9: Spectrogram plot of Cantonese digit "0"

Therefore, it is very hard to differentiate the "05" combination from a single digit "0" or the "35" combination from "3" using spectral knowledge only. Figure 3-10 shows an instance that "3" is misrecognized as the combination of "35". Figure 3-11 shows an instance that "0" is misrecognized as the combination of "05".

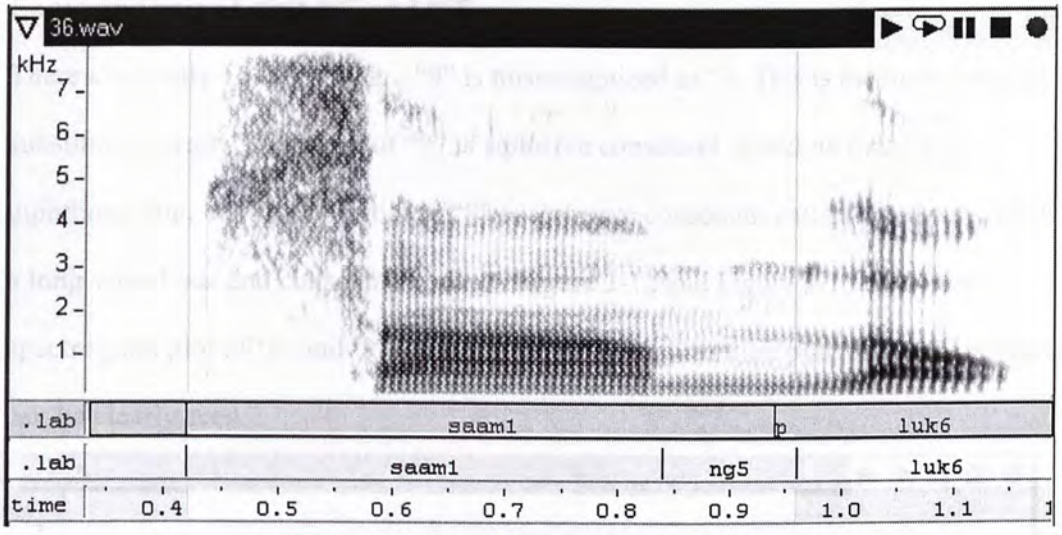


Figure 3-10: Spectrogram plot that explains the insertion "3" →"35"

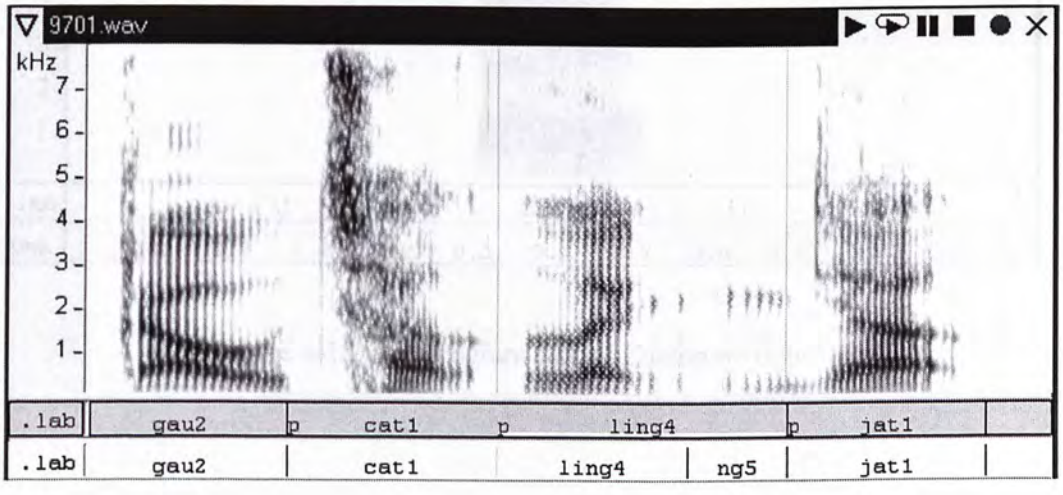


Figure 3-11: Spectrogram plot that explains insertion "0" →"05"

Similarly, if the "05" ("35") combination is misrecognized as "0" ("3") or vice versa, the absolute duration and the relative duration of "0" and "3" would be unreasonable. The absolute duration of "5" would be unreasonable as well. Consequently, the absolute and relative duration information of relevant digits would be helpful.

Confusing pair of digit "9" and "8"

There are totally 15 cases where "9" is misrecognized as "8. This is the most frequent substitution error. The Initial of "9" is a plosive consonant /g/and its Final is a diphthong /au/. Whilst the Initial of "8" is a plosive consonant and its Final consists of a long vowel /aa/ and consonant coda /t/. Figure 3-12 and Figure 3-13 show the spectrogram plot of '8' and '9' respectively. In the figures their phonological structure can be clearly seen.

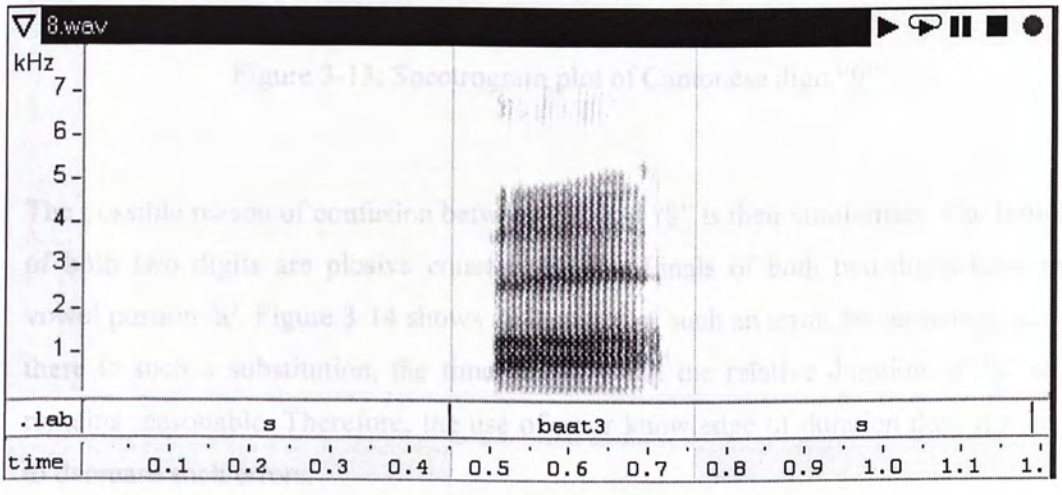


Figure 3-12: Spectrogram plot of Cantonese digit "8"

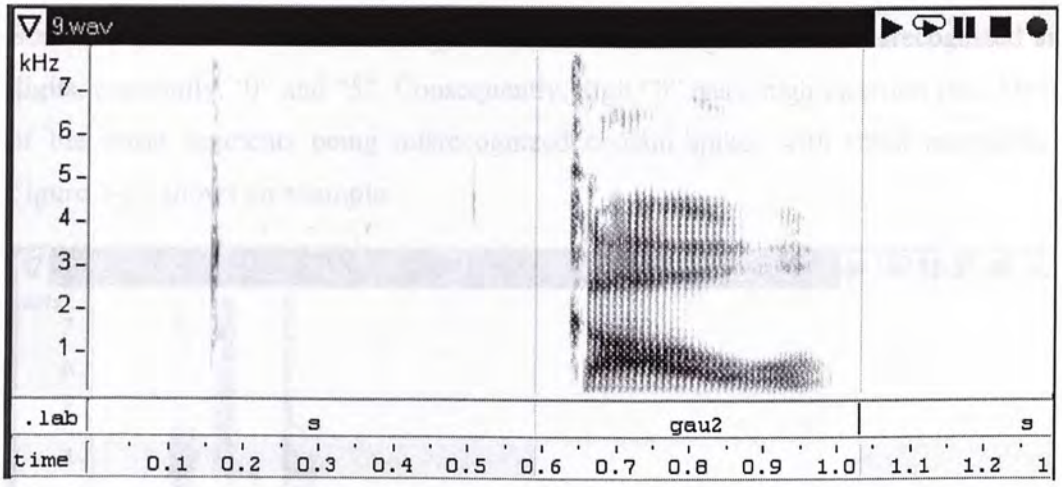


Figure 3-13: Spectrogram plot of Cantonese digit "9"

The possible reason of confusion between "9" and "8" is their similarities. The Initials of both two digits are plosive consonants. The Finals of both two digits have the vowel portion /a/. Figure 3-14 shows an instance of such an error. Nevertheless, when there is such a substitution, the time duration and the relative duration of "8" still remains reasonable. Therefore, the use of prior knowledge of duration does not help to decrease such errors.

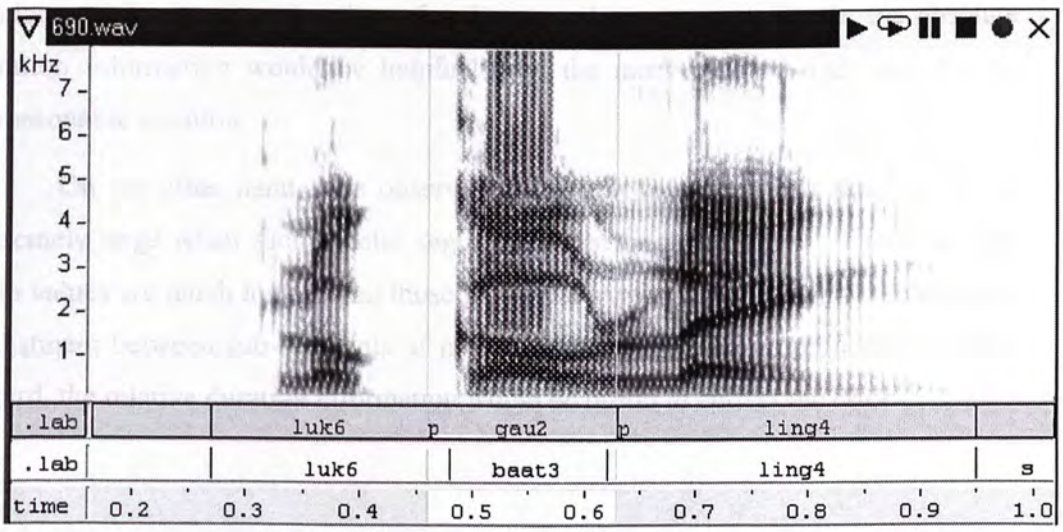


Figure 3-14: Spectrogram plot that shows the substitution "9" →"8"

Noise-induced recognition errors

Some cases of error are caused by noise. The noise segments are misrecognized as digits, especially, '0" and "5". Consequently, digit "0" has a high insertion rate. Most of the noise segments being misrecognized contain spikes with small magnitude. Figure 3-15 shows an example.

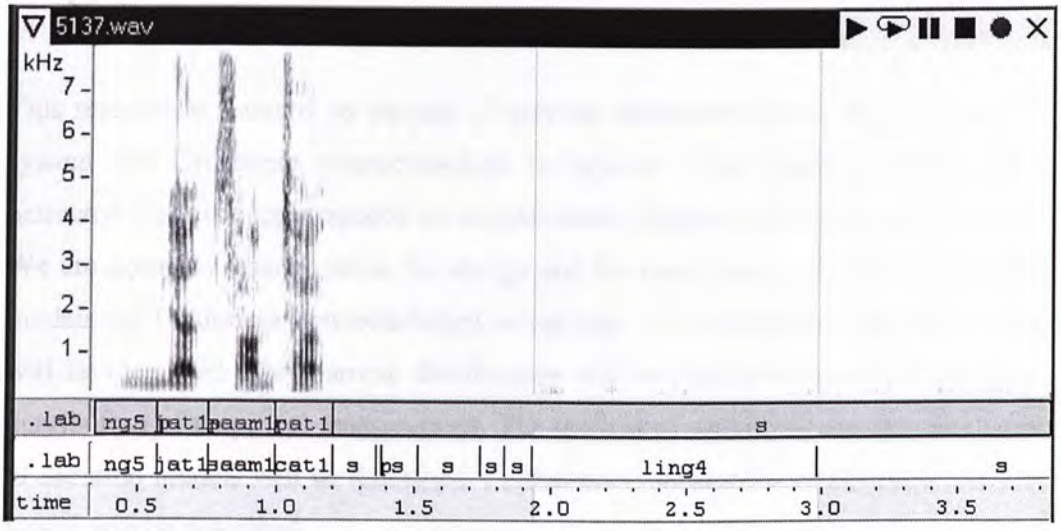


Figure 3-15: Spectrogram plot that illustrates the insertion due to a noise segment being misrecognized as "0"

In Figure 3-15, the noise segment from 1.9s to 3.0s is misrecognized as digit "0". This is an unreasonably long duration for digit "0". The time duration of a noise segment is random, while the time duration of a digit should be regular. Hence, the absolute duration information would be helpful when the misrecognized digit lasts for an unreasonable duration.

On the other hand, it is observed that the value of tail part ratio[6] of "0" is extremely large when such a noise segment is misrecognized as "0". These tail part ratio values are much longer than those in normal cases. It indicates that the duration adjustment between sub-segments of misrecognized "0" is not in proportion. In this regard, the relative duration information would be useful to correct the error.

[6] Tail part ratio is a kind of relative duration. It is defined as the relative duration of last two states in a digit.

Chapter 4

Duration modeling for Cantonese digits

This research is focused on the use of duration information in an HMM-based ASR system for Cantonese connected-digit recognition. The duration information is extracted from the input speech as supplementary information for the ASR process. We are going to discuss about the design and the construction of statistical duration models for Cantonese connected-digit recognition. The appropriate duration features will be identified. The Gamma distributions will be examined in comparison to the distribution of empirical measurement. The method of estimating the free parameters of duration models will be described. Furthermore, speaking-rate-dependent duration model will be described.

4.1. Duration features

As mentioned in the background review, the absolute and relative duration of speech segments have been commonly used as the features for duration modeling. In the following, we will investigate on the use of both of them. Our work is focused on duration modeling at state and model level. In our application, model-level duration essentially gives the duration of a Cantonese digit. State-level duration contains information about the length of sub-segments of a digit.

4.1.1. Absolute duration feature

In an HMM, the implicitly assumed distribution for state duration is inappropriate for real speech signals. This can be observed in our baseline recognition system. Figure 4-1 shows the distribution of state duration in the HMM for Cantonese digit "0". The empirical distribution is obtained by supervised segmentation of training data with the

HMM, and the implicit distribution is computed from the transition matrix of the HMM by the following equation:

$$p_i(d) = a_{ii}^{d-1} \times (1 - a_{ii}) \\ = \text{probability of } d \text{ consecutive observations in state } i \tag{4-1}$$

From the figure, it can be seen that the actual state duration derived from empirical data does not approximate a Geometric distribution. This is observed consistently for other digits as well (see Appendix). Therefore, absolute state duration (AS) is a useful feature to describe speech signal and should be modeled accurately as a supplementary cue to assist HMM based recognition.

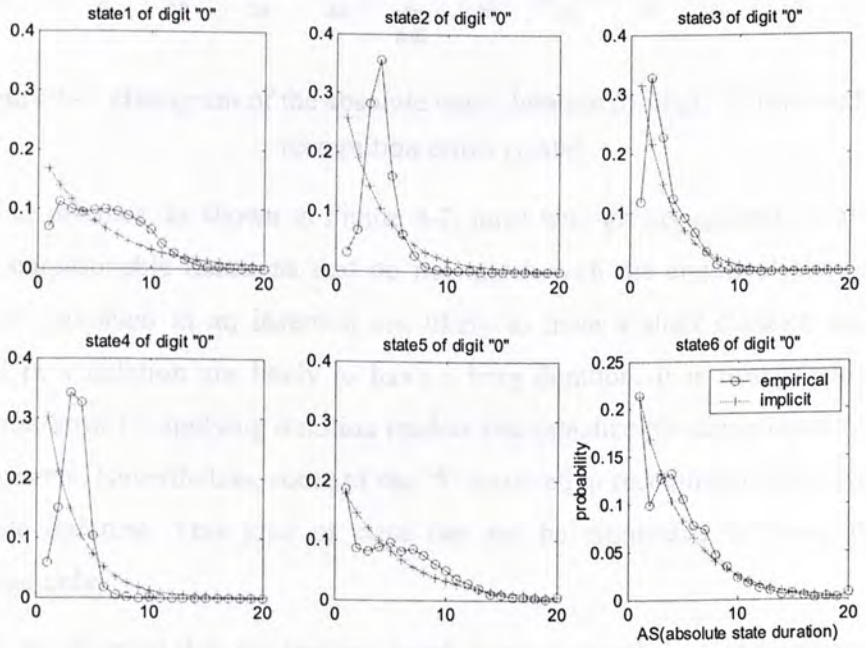


Figure 4-1: Distribution of the absolute state duration (AS) for Cantonese digit "0"

In addition to the absolute state duration, we consider the absolute word duration (AW). As is mentioned earlier, when an insertion or deletion occurs, absolute duration of recognized digit(s) tends to be unreasonable. In this case, accurate modeling of absolute word duration would be helpful.

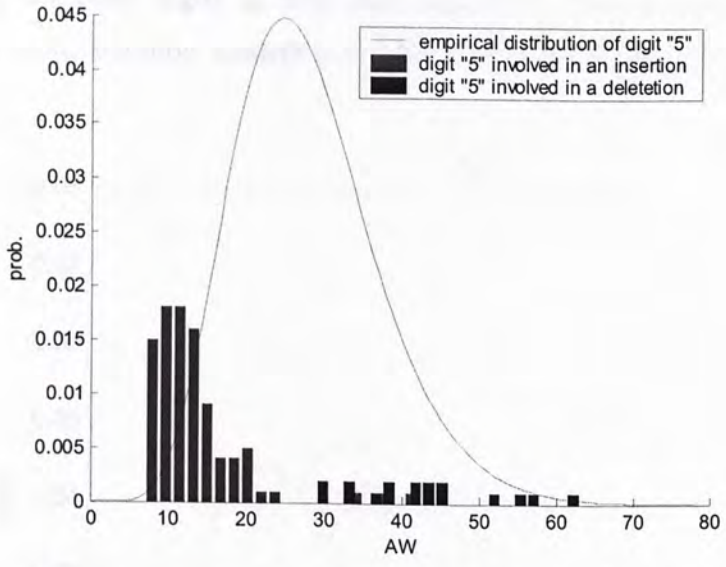


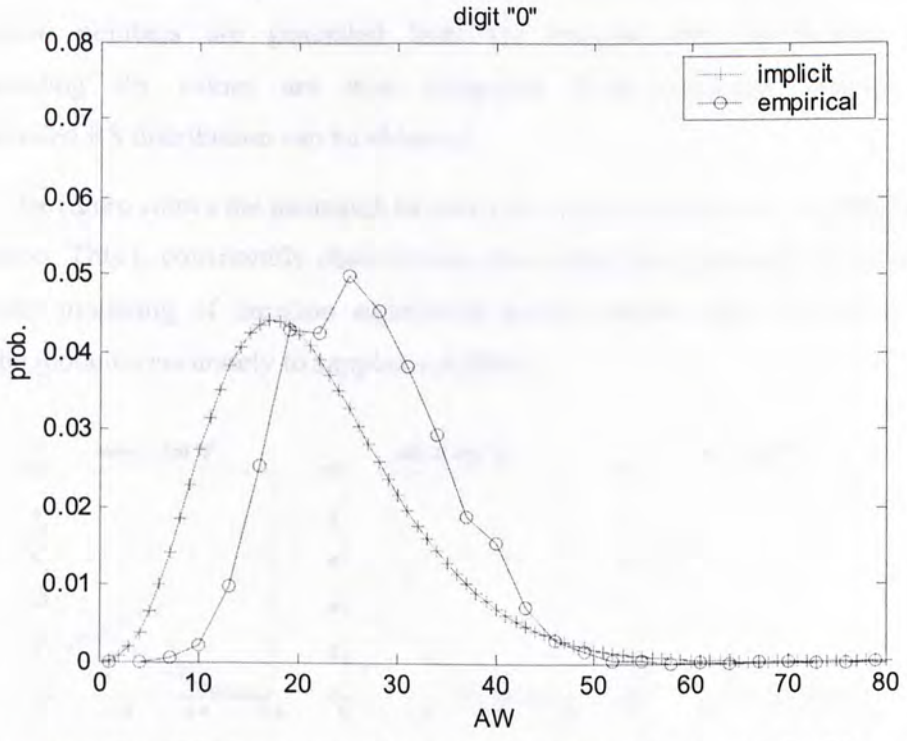
Figure 4-2: Histogram of the absolute word duration for digit '5' involved in recognition errors (male)

For instance, as shown in Figure 4-2, most wrongly recognized '5' are found to have unreasonable durations that do not match with the empirical observations. Digits "5" involved in an insertion are likely to have a short duration and those involved in a deletion are likely to have a long duration. It is possible to correct recognition error by applying duration models that penalize the unreasonable short or long segments. Nevertheless, some of the '5' involved in recognition errors posses an acceptable duration. This kind of error can not be eliminated by using duration knowledge only.

It is observed that the implicit word duration distribution is inappropriate in our recognition system. For instance, the implicit digit duration distribution mismatches the empirical one for digit "0" as shown in Figure 4-3. The implicit distribution is obtained by the following steps:

1) duration distribution of each states in the same HMM is calculated
2) Model duration is obtained as the summation of state duration in that HMM. Assuming that state durations are independent, the duration distribution of a word is the convolution of state duration distribution.

The figure reveals that HMM favors short digit durations. This is consistently observed for other digits as well (see Appendix). This indicates the inadequacy absolute word duration modeling in HMM. It may explain why there are more



insertion errors than deletion errors in the baseline recognition results. In this regard, AW should be utilized as a supplementary cue.

Figure 4-3: Distribution of the absolute word duration for digit "0"

Many factors of variability cumbered absolute duration to be a useful information source. Obviously, the speaking rate is a major concern of time duration. Here we address the speaking rate variability only and let other variability embodied in the natural variation. Construction of speaking-rate-dependent models will be explained in 4.4.

4.1.2. Relative duration feature

The relative duration of HMM states (RS) are investigated. It is observed that the implicit RS distribution is inappropriate in our recognition system. Figure 4-4 compares the implicit distribution and empirical distribution of all the states in digit

"0". It is possible to get the implicit distribution of RS analytically or numerically in the same way as we calculated the implicit AW distribution. However, the implementation is difficult for $RS_i = d_i / \sum_{i=1}^{n} d_i$. The implicit duration is obtained by simulation. The simulation procedure is as follows. For each HMM states, a large set of random numbers are generated from the implicit state distribution. The corresponding RS values are then computed. With sufficient samples, an approximated RS distribution can be obtained.

The figure shows the mismatch between the implicit distribution and empirical distribution. This is consistently observed for other digits (see Appendix). It indicates inadequate modeling of duration adjustment among HMM states. Therefore, RS should be modeled accurately to supplement HMM.

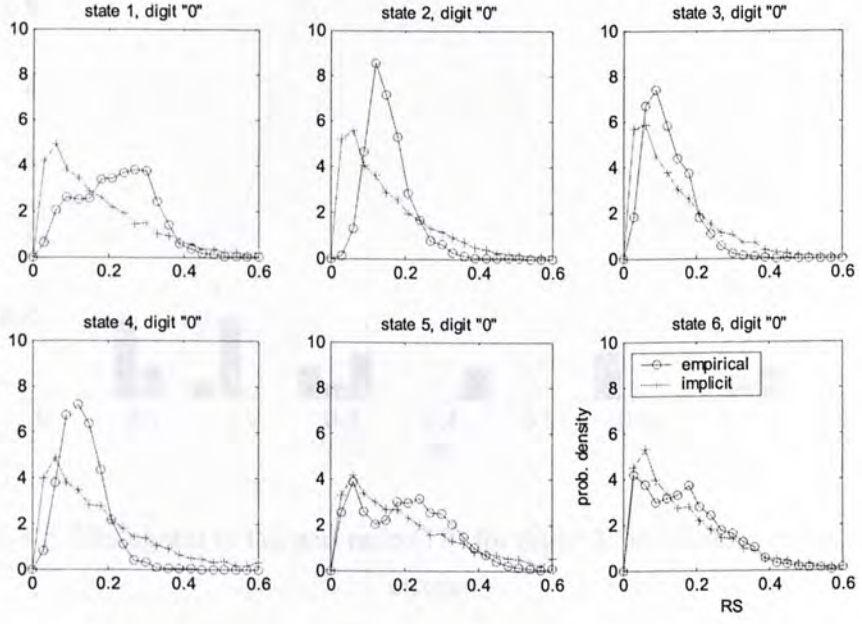


Figure 4-4: Distribution of the relative duration of HMM states for digit "0"

We also propose to use the tail part ratio (TP) for Cantonese connected-digit recognition, which measures the relative duration of the tail part of a digit. The tail part is defined to cover the last two states of the 6-state HMM. The tail part ratio can be considered as a variation of normalized state duration. Observing many cases in the baseline recognition result, we find that the tail part defined above roughly

corresponds to the last phonetic unit in a Cantonese digit. As mentioned in Chapter 3, the two mono-phone digits are very similar to the tail part of another three digits. If this tail part happens to be deleted or prolonged, the tail part ratio would be unreasonable. In this regard, accurate modeling of the tail part ratio would be helpful.

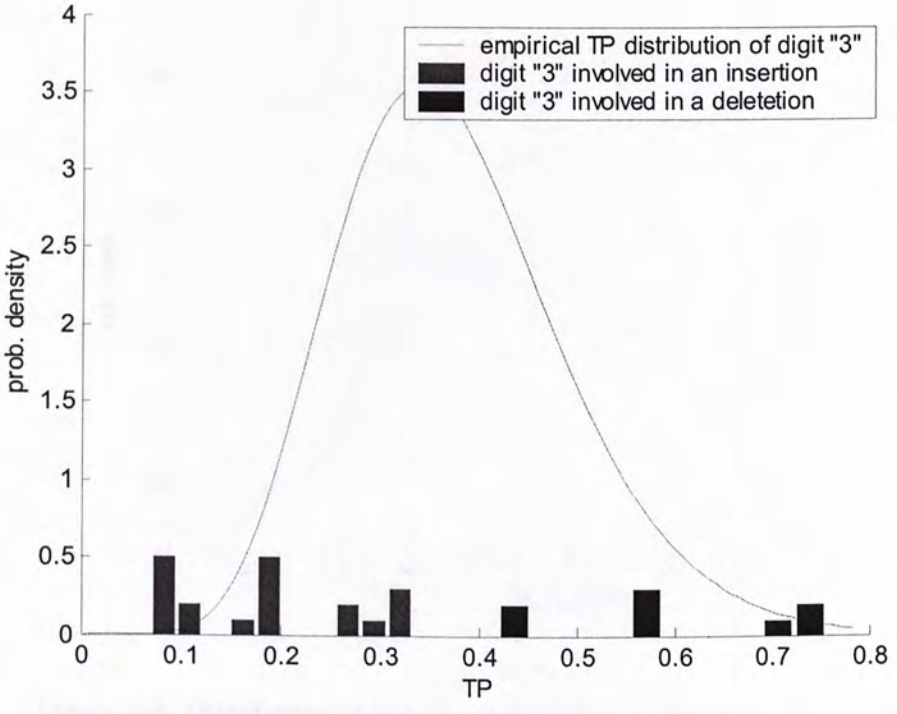


Figure 4-5: Histogram of tail part ratio (TP) for digit "3" involved in recognition errors

As shown in Figure 4-5, we find when '3' is involved in an insertion it is likely to have a small TP value. Similarly, when '3' is involved in a deletion, it is likely to have a large TP value. Such recognition error is possible to be corrected by applying duration modeling so as to penalize the speech segments with unreasonable duration adjustment among its sub-segments. However, there are cases that digits involved in recognition error posses a very reasonable TP values. Again, this kind of error can not be eliminated by using duration model.

Furthermore, implicit TP distribution is found inappropriate in our recognition system. Figure 4-6 plots the empirical distribution and the implicit distribution of TP for digit "3".The implicit distribution of TP is obtained by simulation. It can be seen from this figure that the implicit duration distribution has a much larger variation than the empirical observation. This is consistently seen for other digits (see Appendix). This implies that unreasonable tail part ratios are usually allowed. In this sense, TP should be modeled accurately as a supplementary cue.

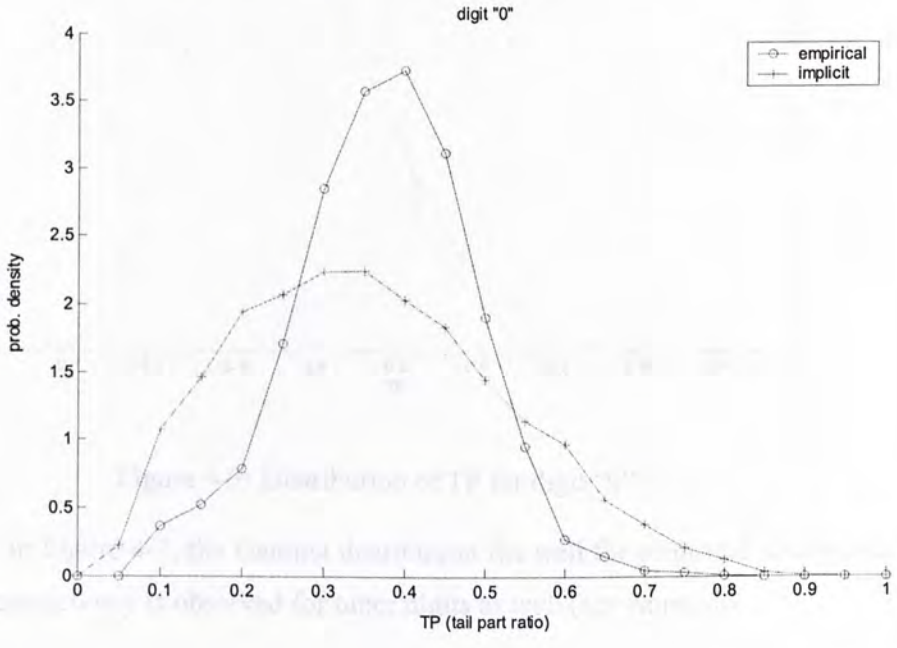


Figure 4-6: Distribution of tail part ratio (TP) for Cantonese digit "0" (male)

4.2. Parametric distribution for duration modeling

The Gamma distribution has been widely accepted as the best candidate for duration modeling. We try to investigate whether the Gamma distribution is appropriate for duration modeling in our system.

For each duration feature that we considered, the Gamma fit is investigated. The following plots show how the distribution derived from empirical data is fit. These plots are for Cantonese digit "0". Plots for other digits are given in Appendix.

1) Plot for TP

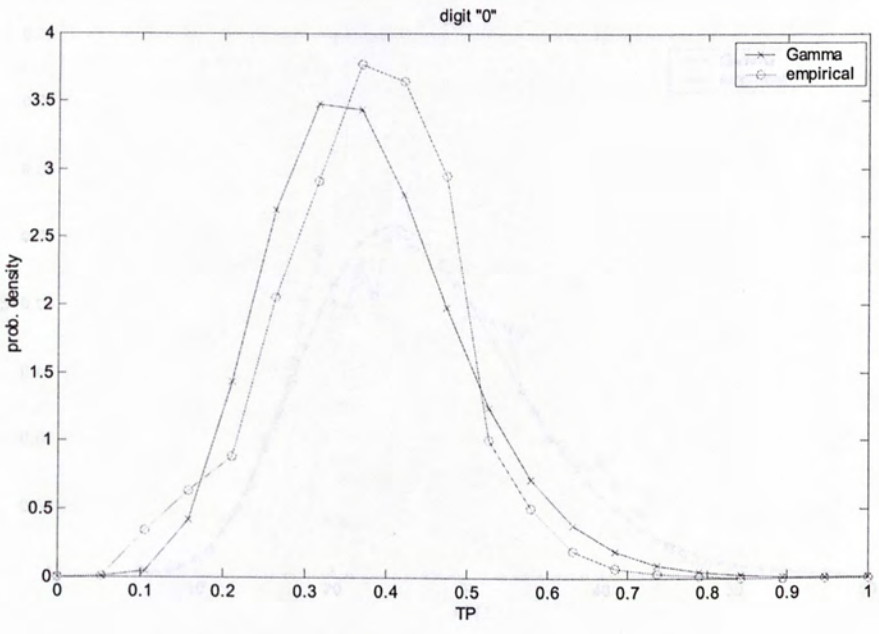


Figure 4-7: Distribution of TP for digit "0"

As is shown in Figure 4-7, the Gamma distribution fits well for empirical distribution of TP. This consistency is observed for other digits as well (see Appendix).

2) Plot for AW

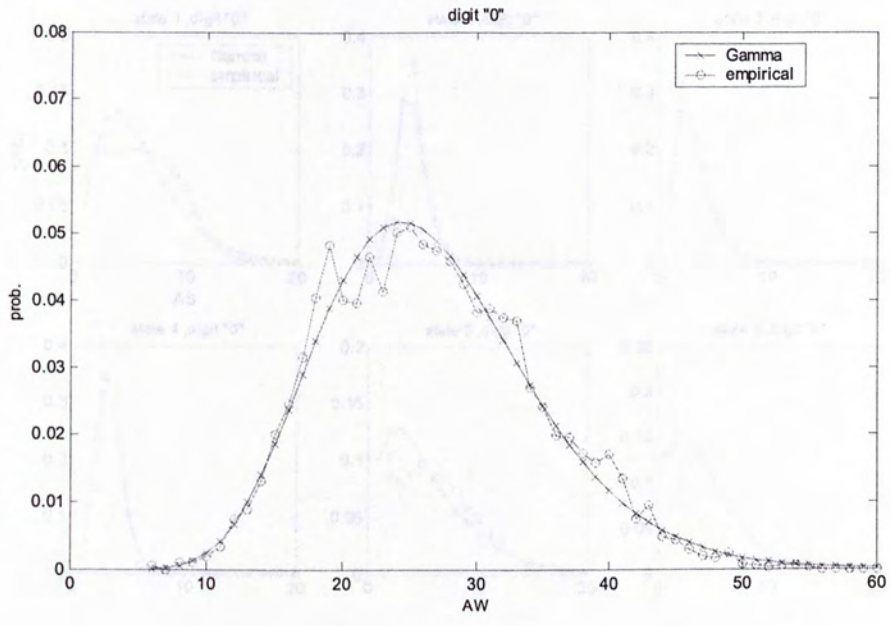


Figure 4-8: Distribution of AW for digit "0"

It is shown in Figure 4-8 that Gamma distribution fits well for empirical distribution of AW. This is observed consistently for other digits (see Appendix).

3) Plot for AS

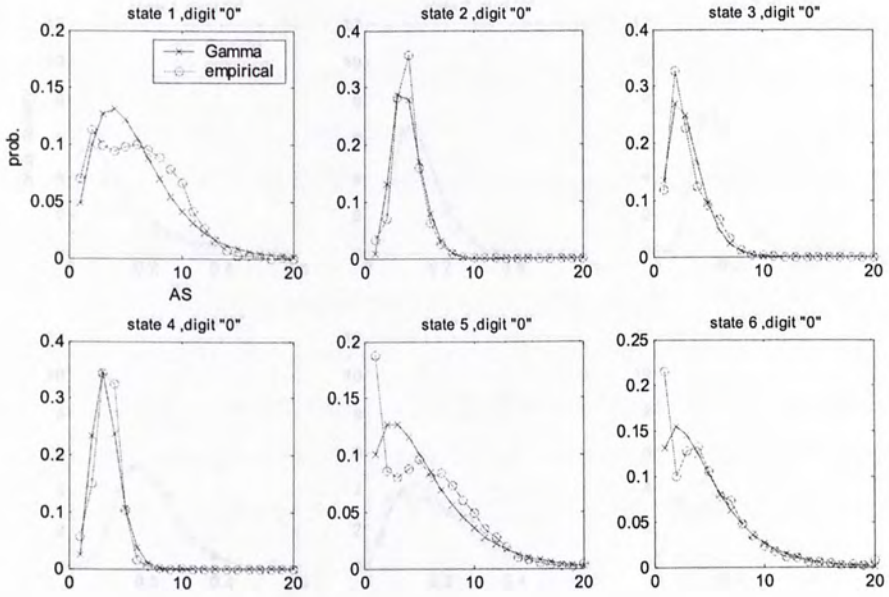


Figure 4-9: Distribution of AS for the HMM states of digit '0'

As seen in Figure 4-9, in most cases the Gamma distribution fits well for empirical distribution of AS. This is observed consistently for other digits (see Appendix). However, the Gamma assumption is inappropriate in a few cases. In Figure 4-9, the Gamma distribution fits badly for the fifth state since the shape of empirical distribution is irregular.

4) Plots for RS

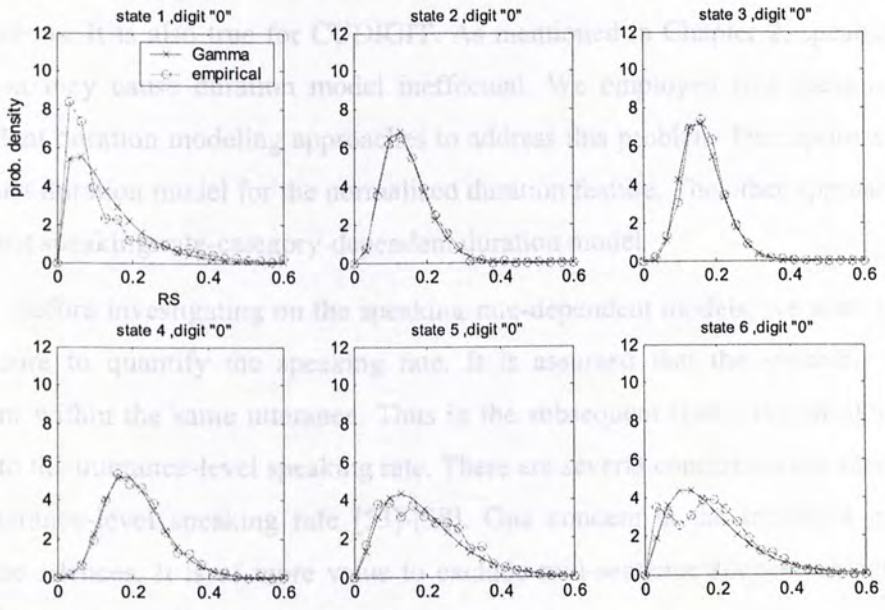


Figure 4-10: Distribution of RS for the HMM states of digit '0'

It can be seen from Figure 4-10 that in most cases Gamma distribution fits well for the empirical distribution of RS. This is observed consistently for other digits (see Appendix). However, problems are seen in a few cases. In Figure 4-10, Gamma fit for the first and sixth state are bad due to the irregular shape of empirical distribution.

4.3. Estimation of the model parameters

In Chapter 2, it was mentioned that there are one-pass training and multi-pass training for estimating the parameters of duration models. For sophisticated duration feature such as TP and RS, one-pass training method would be difficult to be applicable. Thus we will use multi-pass training method to estimate the parameters of the designed duration models.

4.4. Speaking-rate-dependent duration model

Large dynamic range of speaking rate in data is often seen in practical ASR applications. It is also true for CUDIGIT. As mentioned in Chapter 2, speaking rate variation may cause duration model ineffectual. We employed two speaking-rate-dependent duration modeling approaches to address this problem. One approach is to construct duration model for the normalized duration feature. The other approach is to construct speaking-rate-category-dependent duration model.

Before investigating on the speaking-rate-dependent models, we need to have a measure to quantify the speaking rate. It is assumed that the speaking rate is constant within the same utterance. Thus in the subsequent study, the speaking rate refers to the utterance-level speaking rate. There are several concerns when measuring the utterance-level speaking rate [53]-[58]. One concern is the treatment of mid-sentence silences. It is of more value to exclude mid-sentence silence periods since these durations may be dependent on factors other than speech rate [53].

Another concern is the metric of measuring speaking rate. We choose to use the number of words per second as the metric. However, word counting has its disadvantages. It has been pointed out that word rate is unsatisfactory because of unpredictability in the structure and length of a word [54]. Although Cantonese are all monosyllabic, different digits vary in phonetic compositions. Consequently, they have different intrinsic duration. We may approximate the intrinsic duration by the average duration over the training data of that digit. Table 4-1 and Table 4-2 give the average duration from male training data and female training data respectively.

Table 4-1: Average duration of different digits (male)

W	ling4	jat1	ji6	saam1	sei3	ng5	luk6	cat1	baat3	gau2
$\mu_{DUR(w)}$ (frame)	26.74	19.59	27.42	33.14	32.14	27.75	21.46	23.14	22.66	28.97

Table 4-2: Average duration of different digits (female)

W	ling4	jat1	ji6	saam1	sei3	ng5	luk6	cat1	baat3	gau2
$\mu_{DUR(w)}$ (frame)	30.78	21.16	30.12	35.25	33.33	30.07	23.02	24.43	24.55	31.21

The above tables show that the intrinsic durations of digits are quite different from each other. It is desirable to exclude this factor when measuring speaking rate. We choose to follow the UROS (utterance-level rate of speech) metric proposed in [lee98]. It was proposed to address similar problem in a Swedish LVCSR task. Another observation is that the average durations from female data are consistently longer than those from male data. This might be one of the reasons that higher recognition accuracy can be achieved on female data in baseline.

The formula of UROS metric in [27] is as below:

$$WROS(w) = \frac{DUR(w)}{\mu_{DUR}(w)} \tag{4-2}$$

$$UROS(w) = average_{w}(WROS(w)) \tag{4-3}$$

The first equation calculates the instantaneous *WROS* (word-level *ROS*). In the second equation, *UROS* is defined as the average of *WROS* in that utterance. The following plots show histogram of *UROS* values for female and male training data of CUDIGIT respectively.

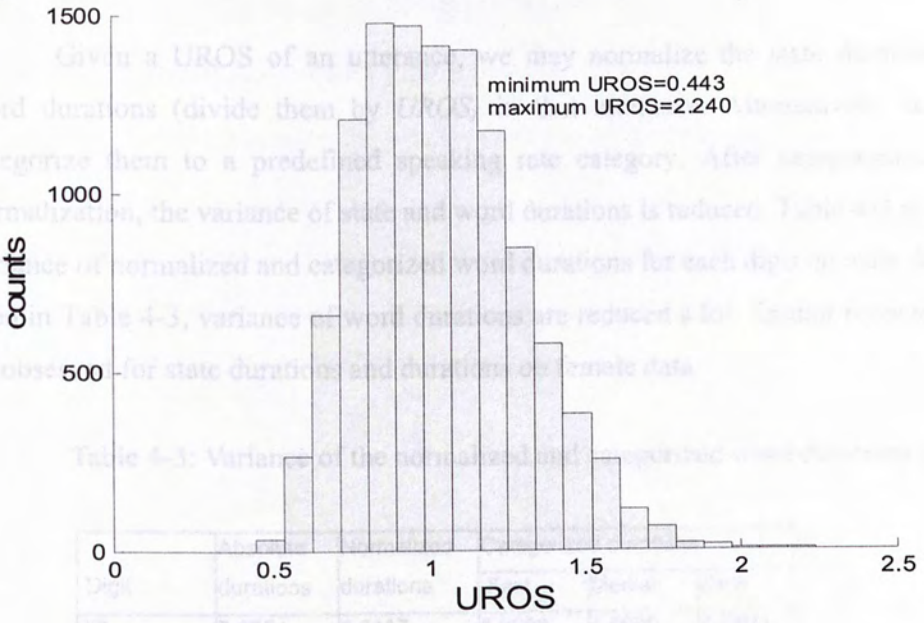


Figure 4-11: Histogram of UROS (male)

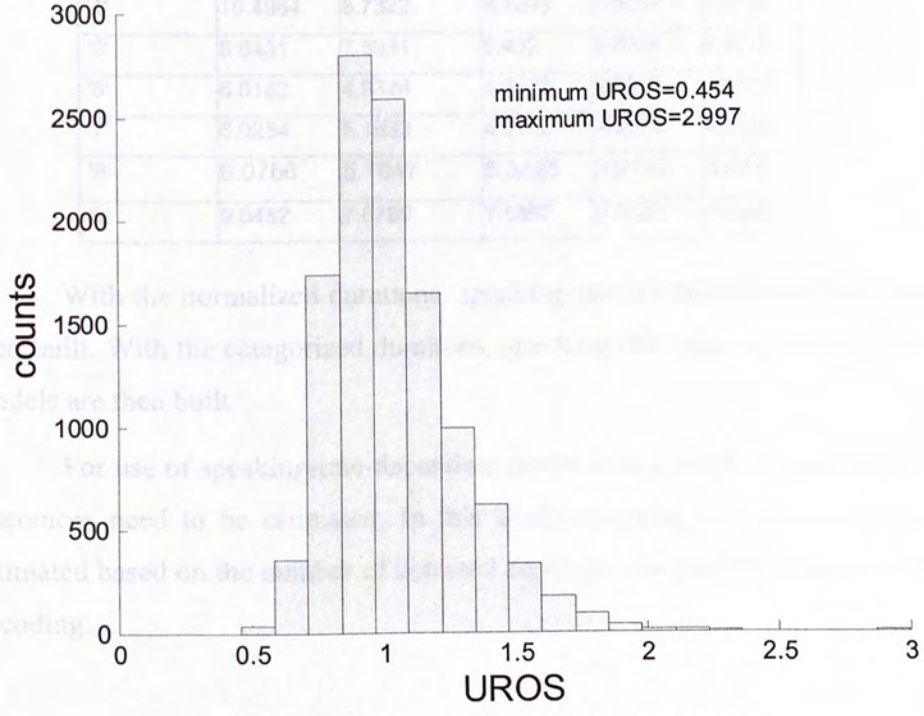


Figure 4-12: Histogram of UROS (female)

The plots of *UROS* of both male and female show a large dynamic range of speaking rate.

Given a UROS of an utterance, we may normalize the state durations and word durations (divide them by *UROS)* in that utterance. Alternatively, we may categorize them to a predefined speaking rate category. After categorization and normalization, the variance of state and word durations is reduced. Table 4-3 gives the variance of normalized and categorized word durations for each digit on male data. As seen in Table 4-3, variance of word durations are reduced a lot. Similar reduction can be observed for state durations and durations on female data.

Table 4-3: Variance of the normalized and categorized word durations (male)

Digit	Absolute durations	Normalized durations	Categorized durations		
			Fast	Medial	Slow
'0'	7.9354	6.3417	5.9686	4.9009	4.7409
'1'	5.9422	4.919	5.0219	3.7974	3.8946
'2'	9.254	7.7547	6.8454	5.7846	6.2336
'3'	9.5616	7.8667	7.5359	6.1028	6.4543
'4'	10.4964	8.7322	8.1263	7.0563	6.9786
'5'	8.8431	7.5931	6.432	6.2024	6.6225
'6'	6.0162	4.9324	4.8772	3.8516	3.5721
'7'	6.0254	5.1633	4.6108	4.3318	4.2532
'8'	6.0766	5.1047	5.3895	3.9747	3.973
'9'	9.0482	7.6789	7.5895	6.3628	6.1421

With the normalized durations, speaking-rate normalized duration models are then built. With the categorized durations, speaking-rate-category dependent duration models are then built.

For use of speaking-rate-dependent model in recognition, speaking rate of test utterances need to be estimated. In this work, speaking rate of a test utterance is estimated based on the number of detected digits per second with a preliminary run of decoding.

In summary, we establish the following explicit duration models for Cantonese connected-digit recognition:

AW model: duration model for absolute model-level duration

AS model: duration model for absolute state duration

RS model: duration model for relative duration of a HMM state

TP model: duration model for relative duration of last two HMM states (tail part)

All the above duration models are digit-dependent regarding that different digits have different phonetic compositions. For these models, they are further improved to speaking-rate-dependent models, namely, speaking-rate category dependent duration model and speaking-rate normalized duration model.

Chapter 5

Using Duration modeling for Cantonese digit recognition

Optimal decoding is superior to other approaches. Certainly, we have to pay attention to the extra computation required and make sure that it is affordable. Considering that Cantonese connect-digit recognition is a small-vocabulary task, we choose to use one-pass decoding. Algorithms are developed for incorporating state-level duration model and word-level duration model to HMM-based recognition respectively.

5.1. Baseline decoder

As mentioned in Chapter 2, the connected-word recognition problem involves optimal path searching in an search space formed by HMM states. The optimal path can be obtained by dynamic programming algorithm. The problem of optimal path search is divided into sub-problems at frame level. In the baseline system, Viterbi algorithm is employed. The step size for path extension is one frame.

For Cantonese connected digit recognition, the HMMs representing different Cantonese digits and silence are connected. The HMM states form the search space. The sub-problem at a particular frame t is to find the optimal partial path extends to each legitimate state. Let (t, v, j) denote the optimal partial path extends to state j of model v and at frame t. The accumulated path score is denoted by $L(t, v, j)$. A sub-problem at frame t can be solved given the solution of sub-problems one step before, i.e., the immediately preceding frame. The algorithm is described as follows:

- Initialization: to solve the sub-problems at frame 1:

$$L(1, v, 1) = 1 \tag{5-1}$$

$$L(1, v, j) = \infty \tag{5-2}$$

- Recursion (path extension): to solve the sub-problems at frame t given the solution of sub-problems at frame t-1.

1) If the path is extended to the first state of a HMM, its predecessor can be the last state of any HMM or the current state itself. Each possible predecessor $(t-1,u,N_u)$ or $(t-1,v,1)$ is evaluated, where N_u is the last state of HMM u. A decision is made to choose the best predecessor according to the following equation.

$$L(t,v,1)=\max_u\left\{L(t-1,u,N_u)\times a_{N_u-1,N_u},L(t-1,v,1)\times a_{11}\right\}\times b_1(o_t) \quad (5\text{-}3)$$

At the same time, a back pointer pointing to the best predecessor is recorded

$$B(t,v,1)=\arg\max_u\left\{L(t-1,u,N_u)\times a_{N_u-1,N_u}\right\} \quad (5\text{-}4)$$

2) If the path is extended to the jth state of a HMM with $j\neq 1$, its predecessor can be any state of the same HMM that can make transition to it. Each possible predecessor $(t-1,v,i)$ will be evaluated. The best predecessor is chosen according to the following equation.

$$L(t,v,j)=\max_{1\le i\le j}\left\{L(t-1,v,i)\times a_{ij}\right\}\times b_j(o_t) \quad (5\text{-}5)$$

Similarly, the corresponding back pointer is recorded.

$$B(t,v,j)=\arg\max_{1\le i\le j}\left\{L(t-1,v,i)\times a_{ij}\right\} \quad (5\text{-}6)$$

- Termination: to terminate at frame T. Token q_T with highest likelihood δ is selected as the end of the optimal path.

$$\delta=\max_u\left\{L(T,u,N_u)\right\} \quad (5\text{-}7)$$

$$q_T=\arg\max_u\left\{L(T,u,N_u)\right\} \quad (5\text{-}8)$$

- Backtracking: to trace back the most likely sequence from q_T according to the back pointers recorded before.

$$q_{t-1}=B(q_t)\quad t=T,\ T-1,\ T-2,\ldots,\ 1 \quad (5\text{-}9)$$

As mentioned in the chapter of background, knowledge sources such as explicit duration model are different with HMM based acoustic models. To incorporate these knowledge sources into the recognition processes, dynamic programming algorithm can still be used. However, the path extension needs to cover much longer time span. To use duration model for state-level features, the path extension should cover a time span of a HMM state. To use duration model for model-level features, the path extension should cover a time span of a model.

5.2. Incorporation of state-level duration model

A dynamic programming algorithm is developed to incorporate duration model for state-level duration feature, i.e. absolute state duration feature, into the HMM-based connect word recognition.

In this algorithm, the step size of path extension is chosen to be a time span of state. Duration probability is incorporated into the path extension. After each extension decision, the path is extended to the beginning of a new state.

The optimal path problem is divided into sub-problems at consecutive frames. The sub-problem at a particular frame t is to find the optimal partial path just extends to each legitimate state. Let (t, v, j) denote the optimal partial path just extends to state j of model v and at frame t, and $L(t, v, j)$ be the corresponding accumulated path score. A back pointer $B(t,v,j)$ is also recorded for back tracing the most likely path . A sub-problem at frame t can be solved given the solution of sub-problems at the previous state. The algorithm is described in details below:

- Initialization: to solve the sub-problems at Frame 1:

$$L(1, v, 1) = 1 \tag{5-10}$$

$$L(1, v, j) = \infty \ (j > 1) \tag{5-11}$$

- Recursion (path extension), to solve the sub-problems at Frame t given the solution of sub-problems at one step size before:

1) If the path is extended to the first state of a HMM, its predecessor can be the last state of any HMM. For each possible predecessor $(t - d, u, N_u)$, the

duration probability $D_{u,N_u}(d)$ is calculated. $D_{u,N_u}(d)$ is the probability that state Nu in model u has a time duration of d. $D_{u,N_u}(d)$ is then incorporated in the path extension decision by

$$L(t,v,1)=\max_{\substack{u\\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,u,N_u)\times a_{N_u-1,N_u}\times\prod_{t-d<\tau<t}b_{N_u}(o_\tau)\times\left[D_{u,N_u}(d)\right]^w\right\}\times b_1(o_t) \quad (5\text{-}12)$$

where w is for balance use. d_{max} and d_{min} are the up bound and low bound of state duration respectively. A back pointer pointing to the best predecessor is recorded:

$$B(t,v,1)=\arg\max_{\substack{u\\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,u,N_u)\times a_{N_u-1,N_u}\times\prod_{t-d<\tau<t}b_{N_u}(o_\tau)\times\left[D_{u,N_u}(d)\right]^w\right\} \quad (5\text{-}13)$$

2) If the path is extended to the jth state of HMM with $j\neq 1$, its predecessor can be any state of the same HMM that can transit to it. For each possible predecessor $(t-d,v,i)$, the duration probability $D_{v,i}(d)$ is calculated and incorporated into the path extension decision as in the following equation.

$$L(t,v,j)=\max_{\substack{1\le i<j\\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,v,i)\times a_{ij}\times\prod_{t-d<\tau<t}b_i(o_\tau)\times\left[D_{u,i}(d)\right]^w\right\}\times b_j(o_t) \quad (5\text{-}14)$$

Similarly, the back pointer is recorded.

$$B(t,v,j)=\arg\max_{\substack{1\le i<j\\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,v,i)\times a_{ij}\times\prod_{t-d<\tau<t}b_i(o_\tau)\times\left[D_{u,i}(d)\right]^w\right\} \quad (5\text{-}15)$$

- Termination: to terminate at frame T and select the token q_s with highest likelihood δ_T as the final solution to the optimal path.

$$\delta_T=\max_{\substack{u\\ d_{\min}\le d\le d_{\max}}}\left\{L(T-d,u,N_u)\times a_{N_u-1,N_u}\times\prod_{t-d<\tau<t}b_{N_u}(o_\tau)\times\left[D_{u,N_u}(d)\right]^w\right\} \quad (5\text{-}16)$$

$$q_S = \arg\max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(T-d,u,N_u) \times a_{N_u-1,N_u} \times \prod_{t-d<\tau<t} b_{N_u}(o_\tau) \times \left[D_{u,N_u}(d) \right]^w \right\} \quad (5\text{-}17)$$

S is the number of path extensions required to constitute the optimal path.

- Backtracking: to trace back the most likely sequence from q_S according to the back pointers recorded before.

$$q_{s-1} = B(q_s) \quad t = \mathrm{S}, \mathrm{S}-1, \ldots, 1 \quad (5\text{-}18)$$

The above formulation is referred to as the 3-dimensional optimal decoder because the token (t, v, j) has three elements. This 3D optimal decoder has a large number of repetitive computations of $\prod_{t-d<\tau<t} b_i(o_\tau)$ due to common partial paths. The main advantage of dynamic programming is to save this kind of repetitive computations by having intermediate results stored. Hence, dynamic programming techniques can be applied in this case. A new variable state duration "*d*" is introduced to the path token. The token *(t,v,j,d)* refers to a path gets to the *d*th frame of the *j*th state in HMM *v* at frame *t*. Re-write the 3D decoder and we have:

- Initialization:

$$L(1,v,1,1) = 1 \quad (5\text{-}19)$$

$$L(1,v,1,d) = \infty \quad (d>1) \quad (5\text{-}20)$$

- Recursion:

$$L(t,v,j,1) = \max_{\substack{1 \le i < j \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-1,v,i,d) \times a_{ij} \times \left[D_{u,i}(d) \right]^w \right\} \times b_j(o_t) \quad (5\text{-}21)$$

$$L(t,v,j,d) = L(t-1,v,j,d-1) \times b_j(o_t) \quad (5\text{-}22)$$

$$L(t,v,1,1) = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-1,u,N_u,d) \times a_{N_u-1,N_u} \times \left[D_{u,N_u}(d) \right]^w \right\} \times b_1(o_t) \quad (5\text{-}23)$$

Back pointer:

$$B(t,v,j,1) = \arg\max_{\substack{1 \le i < j \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-1,v,i,d) \times a_{ij} \times \left[D_{u,i}(d) \right]^w \right\} \quad (5\text{-}24)$$

$$B(t,v,j,d) = (t-1,v,j,d-1) \quad (5\text{-}25)$$

$$B(t,v,1,1) = \arg\max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-1,u,N_u,d) \times a_{N_u-1,N_u} \times \left[D_{u,N_u}(d) \right]^w \right\} \quad (5\text{-}26)$$

- Termination:

$$\delta = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(T,u,N_u,d) \times a_{N_u-1,N_u} \times \left[D_{u,N_u}(d) \right]^w \right\} \quad (5\text{-}27)$$

$$q_{T+1} = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(T,u,N_u,d) \times a_{N_u-1,N_u} \times \left[D_{u,N_u}(d) \right]^w \right\} \quad (5\text{-}28)$$

- Backtracking:

$$q_t = B(q_{t+1}) \quad t = T,\ T-1,\ T-2,\ldots,\ 1 \quad (5\text{-}29)$$

This resultant 4D decoder is essentially equivalent to the decoding framework in [18], which is oriented for isolated word recognition task. The computation of this decoder will be d_{max} times that of baseline decoder. In implementation, we will limit the d_{max} to a reasonable value, say, 15 to limit the searching region of path extension.

5.3. Incorporation word-level duration model

A dynamic programming algorithm is developed to incorporate duration model for word-level duration feature, i.e. AW, RS and TP.

In this algorithm, the step size of path extension is chosen to be a time span of a word when incorporating the duration probability. After each path extension, the path is extended to the beginning of a word. The sub-problem at a particular frame *t* is to find the optimal partial path just extends to each legitimate word. Let *(t,v)* denote the optimal partial path just extends to model *v* and at frame *t*, and *L(t,v)* be the corresponding accumulated path score. A back pointer *B(t,v)* is recorded for back trace the most likely path . A sub-problem at Frame *t* can be solved given the solution of sub-problems one step before, i.e., the preceding word. The algorithm is described as follows:

- Initialization: to solve the sub-problems at Frame 1:

$$L(1,v)=1 \tag{5-30}$$

- Recursion (path extension): to solve the sub-problems at Frame t given the solution of sub-problems one step size before. The predecessor of *(t,v)* can be the beginning of any HMM. For each possible predecessor *(t-d,u)*, the duration probability $D_u(d)$ is calculated and incorporated in the path extension decisions as in the following equation:

$$L(t,v)=\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,u)\times warp(u,t-d,t-1)\times\left[D_u(d)\right]^w\right\} \tag{5-31}$$

, where $warp(u,t-d,t-1)$ is the probability that the feature vector sequence *t-d* through *t*-1 is generated by HMM *u*. It is a dynamic programming at another level. d_{max} and d_{min} are the up bound and low bound of word duration respectively. $D_u(d)$ can be the duration score contributed by one or more word-level duration models, which include the AM models, RS models, TP models. If it is RS model, $D_u(d)$ will be $\prod_s P_u(d_s)$. Where d_s is the state duration value, $P_u(d_s)$ is the probability that duration of state *s* in model *u* has a value of d_s. We assume probability from each state are independent and can be multiplied to give the overall probability.

At the same time, the back pointer point to the best predecessor is recorded.

$$B(t,v)=\arg\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(t-d,u)\times warp(u,t-d,t-1)\times\left[D_u(d)\right]^w\right\} \tag{5-32}$$

- Termination: to terminate at Frame T and select the token q_s with highest likelihood δ_T as the final solution to the optimal path.

$$\delta_T=\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(T-d+1,u)\times warp(u,T-d+1,T)\times\left[D_u(d)\right]^w\right\} \tag{5-33}$$

$$q_S=\arg\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(T-d+1,u)\times warp(u,T-d+1,T)\times\left[D_u(d)\right]^w\right\} \tag{5-34}$$

- Backtracking: to trace back the most likely sequence according to the back pointers recorded before according to

$$q_s=B(q_{s+1}) \quad t=\mathrm{S}-1,\ \mathrm{S}-2,\ldots,\ 1 \tag{5-35}$$

, where S is the number of path extensions to constitute the optimal path.

Similarly, there are a large number of repetitive warp operations. To reduce cost of computation, a new variable, word duration *"d"* is introduced to the path token. Token *(t,v,j,d)* refer to a path gets to the *d*th frame of HMM *v* with state *j* at time *t*. Re-writing above 2D decoder and we have:

- Initialization:

$$L(1,v,1,1)=1 \tag{5-36}$$

$$L(1,v,1,d)=1 \quad d>1 \tag{5-37}$$

- Recursion:

$$L(t,v,1,1)=\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(t-1,u,N_u,d)\times a_{N_u-1,N_u}\times[D_u(d)]^w\right\}\times b_1(o_t) \tag{5-38}$$

$$L(t,v,j,d)=\max_{1\le i<j}\left\{L(t-1,v,i,d-1)\times a_{ij}\right\}\times b_j(o_t) \tag{5-39}$$

$$B(t,v,1,1)=\arg\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(t-1,u,N_u,d)\times a_{N_u-1,N_u}\times[D_u(d)]^w\right\} \tag{5-40}$$

$$B(t,v,j,d)=\arg\max_{1\le i<j}\left\{L(t-1,v,i,d-1)\times a_{ij}\right\} \tag{5-41}$$

- Termination:

$$\delta=\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(T,u,N_u,d)\times a_{N_u-1,N_u}\times[D_u(d)]^w\right\} \tag{5-42}$$

$$q_{T+1}=\arg\max_{\substack{u \\ d_{\min}\le d\le d_{\max}}}\left\{L(T,u,N_u,d)\times a_{N_u-1,N_u}\times[D_u(d)]^w\right\} \tag{5-43}$$

- Backtracking:

$$q_t=B(q_{t+1}) \quad t=T,\ T-1,\ T-2,\ldots,\ 1 \tag{5-44}$$

This resulted 4D decoder is essentially the same as that was proposed in [21] for Korean connected-digit recognition. The computation cost of this decoder is d_{max}

times that of baseline decoder. In practical implementation, we can limit the d_{max} to a reasonable value, say, 80 to restrict the searchable region in path extension. The computation and storage caused is much heavier as compared to that of baseline decoder and decoder for use of state-level duration model.

5.4. Weighted use of duration model

Experimental observations in our experiments show that, the acoustic score from HMM always have a much larger OOM than that of the duration score. Score from HMM and Duration models in some random selected cases are given in Table 5-1. Therefore, the effect of duration models tends to be overshadowed by that of HMM. In this work, a positive weighting factor w is used to balance the situation.

Table 5-1: Score from HMM and duration models for "0" in random selected cases

HMM	AW	RS	TP	AS
-1758.8	-3.1	-4.2	-0.4	-10.8
-1786.5	-3	-3.4	-0.6	-10.7
-1889.6	-3	-4.2	-0.3	-8.1
-2419.6	-3.2	-2.1	-0.5	-9.8
-1660.6	-3.1	-4.1	-0.1	-9
-2619.5	-3.4	-6.5	-2.2	-8
-1735.4	-3.1	-3.5	-0.1	-10

Chapter 6

Experimental results and analysis

In the previous chapters, we have discussed about duration models at various levels. We also explained two optimal decoding algorithms that can incorporate the duration models into HMM-based recognition of Cantonese connected-digits. Recognition experiments have been carried out to evaluate the effectiveness of the duration models for different duration features with optimal decoding algorithms. Furthermore, use of speaking-rate-dependent duration models has been evaluated.

Experiments with speaking-rate-independent duration models will be referred to as the 'g' scheme. Speaking-rate-dependent duration models include speaking-rate normalized duration models and speaking-rate category dependent duration models. Experiments with them are denoted as the 'n' scheme and the 'c' scheme respectively. The list of our experiments is given in Table 6.1. Similar to the experiments with the baseline system, speech data from male and female speakers in the CUDIGIT are tested separately.

Table 6-1: List of experiments

	Word-level duration model					State-level duration model
'g'	AM	RS	TP	AM+RS	AM+TP	AS
'c'	AM+RS				AM+TP	AS
'n'	AM+RS				AM+TP	AS

6.1. Experiments with speaking-rate-independent duration models

Table 6-2: Recognition performance with different duration features (male)

	Digit acc%	Sen acc%	Del	Sub	Ins
Baseline	96.77	89.50	87	69	181
AW	97.45	91.64	126	67	72
RS	97.58	92.41	103	70	78
TP	97.40	91.85	85	60	125
AW+RS	97.68	92.34	128	61	52
AW+TP	97.82	92.87	121	61	44
AS	97.83	92.83	109	64	52

Many trials of experiments have been carried out with different weights within a limited range on duration scores. The results in Table 6.2 are the performance with the best weight.

Table 6-3: Best weight obtained from trials on male speech data

	AW	RS	TP	AW+RS	AW+TP	AS
Weight	6	4	4	(6,2)	(6,4)	3

For the subsequent recognition experiments, we will continue using the weight obtained in this stage for balance the contribution of duration models.

Table 6-4: Recognition performance with different duration features (female)

	Digit acc%	Sen acc%	Del	Sub	Ins
Baseline	98.12	93.43	30	44	121
AW	98.55	94.77	61	48	42
RS	98.48	94.66	51	48	59
TP	98.26	93.96	40	45	96
AW+RS	98.63	95.08	69	45	28
AW+TP	98.59	94.91	64	47	37
AS	98.60	95.01	54	40	52

6.1.1. Discussion

Table 6-2 and Table 6-4 show the recognition performance with different duration features on male and female data respectively. It can be seen that in all cases the recognition performance is improved as compared with the baseline system. For male speech, the recognition accuracy is improved by up to 1.06%. For female data, the recognition accuracy is improved by up to 0.51%.

Word-level duration features

From Table 6-2, it can be seen that using word-level duration model noticeably improves the recognition accuracy. Among the three word-level features, RS (relative state duration) offers the best improvement. If the AW model and the RS model are used in combination, a slight degree of further improvement can be achieved. This implies that they are complementary to each other to some extent. For example, the baseline recognition output of an utterance is as follows:

Table 6-5: Baseline recognition output of "data/CD16M/data/5667.mfc"

Duration (frame)	TP	Recognized digit
37		s
24	0.083333	"6"
14	0.428571	"6"
22	0.454545	"7"
51		s

As can be seen in Table 6-5, a combination of "56" of this utterance is misrecognized as "6". Use of AW failed to reduce the error, since the time duration of recognized "6" still reasonable. However, use of TP and RS is able to reduce this error since sub-segments of recognized "6" are in unreasonable portion. The combination of AW and TP attained 97.82% accuracy, which is comparable to that attained by combination of AW and RS model. It shows that TP is equally effective as RS in supplementing the AW information.

In Table 6-4, the use of AW or RS model consistently achieves noticeable improvement. Nevertheless, the TP model shows only marginal improvement. As discussed in Chapter 4, the motivation of TP was to deal with the possible insertion or deletion that come with the digits "0","3" and "4". It was observed, for male data in CUDIGIT, that if the coda of these happen to be deleted or prolonged, the TP of these digits would become unreasonable. However, this observation does not hold on female data. Therefore, TP is not as good a duration feature as RS for relative duration modeling. Combined use of AW and either of relative duration model gives further improvement

State-level duration features

As shown in Table 6.2 and Table 6.3, the use of state-level duration model shows better recognition performance than use of any of the word-level duration model. On the other hand, it shows similar performance to the combined use of word-level features. It indicates that the state-level information is as useful as word-level

information. In this regard, using of state-level duration model is a better approach because it does not take as much computation as that of using word-level duration model.

Confusion matrix of digits

With use of duration modeling in recognition, noticeable change has been taken place in confusion matrix from that of baseline. For expedience of illustration, the confusion matrix of using AS on male data is given in Table 6.6. It is observed that:

1) Number of insertions is greatly reduced for most of digits, in particular, the digit "2" and "5". Duration model shows to be effective to reduce recognition errors that caused by insertion of digits with short duration.
2) Number of deletion for most digits is increased for most of digits, especially the digit "5". The reason will be analyzed later.
3) Insertions of "0" are reduced a lot. As we mentioned earlier, "0" has a high insertion rate because some noise segments are misrecognized to "0". It indicates that duration modeling would be helpful in noise robust ASR.
4) No significant change is observed in the number of substitution. Duration model shows marginal effect on substitution error.

Table 6-6: Confusion matrix of recognition results with duration model (male: AS)

	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	del	%c/%e
'0'	1028	6	0	0	0	2	0	0	0	0	5	98.9/0.1
'1'	1	1030	0	0	1	0	0	0	2	0	**16**	99.1/0.1
'2'	0	3	987	0	0	1	0	0	0	0	**24**	99.6/0.0
'3'	0	0	0	1018	0	0	3	0	5	0	1	99.5/0.0
'4'	0	1	0	0	1066	0	0	1	0	0	0	99.8/0.0
'5'	0	0	1	0	0	965	1	0	0	0	**49**	99.5/0.0
'6'	0	1	0	0	0	0	1022	0	0	3	7	99.6/0.0
'7'	0	2	0	2	3	0	0	1062	0	0	4	99.4/0.1
'8'	0	2	0	0	0	0	0	0	1048	0	4	99.8/0.0
'9'	0	2	0	0	0	0	4	1	**15**	989	0	97.9/0.2
ins	**4**	**7**	**10**	1	0	**25**	2	0	1	2		

6.1.2. Analysis of the error patterns

In the experiments, it is observed that some recognition errors are reduced with the use of duration information, but at the same time, some new errors are introduced. Statistics are given in Table 6-7 and Table 6-8 for male and female experiments respectively.

Table 6-7: Reduced recognition errors and newly introduced ones (male)

Duration model	No of reduced	No of newly introduced
AW+TP	159	49
AS	164	53

Table 6-8: Reduced recognition errors and newly introduced ones (female)

Duration model	No of reduced	No of newly introduced
AW+TP	93	45
AS	110	47

In comparison to the number of reduced errors, the number of newly introduced errors is not a negligible number. There are two possible reasons that can be used to explain the trade-off between reduced errors and newly introduced errors. The first reason is that some correctly recognized digits may exhibit unreasonable duration as well. Duration model may falsely penalize these digits and introduce new recognition errors. The other reason is that our duration models have a tendency to prefer longer duration. This helps to penalize the cases with unreasonably short duration and prevent to penalize the unreasonable long durations. Consequently, there will be a tradeoff between reduced insertion errors and new deletion errors. This tradeoff will be elaborated in details in the next section.

6.1.3. Reduction of deletion, substitution and insertion

The trade-off between reduced errors and newly introduced errors is further analyzed in terms of deletion, substitution and insertion. Table 6-9 and Table 6-10 give the statistics on the experiments with duration models on male data.

Table 6-9: Reduced recognition errors and newly introduced ones (male: AW+TP)

	Reduced	Newly introduced
Del	2	37
Sub	17	10
INS	140	2

Table 6-10: Reduced recognition errors and newly introduced ones (male: AS)

	Reduced	Newly introduced
Del	10	33
Sub	17	12
INS	137	8

From these tables, it is observed that:

1) Substitution errors do not change as much as insertion and deletion errors. It may due to that digits been substituted for still exhibit reasonable durations.
2) Most of the newly introduced errors are deletion errors.
3) Most of the reduced errors are insertion errors.

The last two observations lead to a conclusion that longer duration is preferred in the decoding with the proposed duration models. The preference to longer duration or shorter duration depends on the strength and polarity of duration score. Negative supra-segmental scores tend to prefer less digits and longer duration, which will lead to more deletion errors. In our experiment, the duration scores are all negative. As a result, we see this tendency in the experiment results. Furthermore, when the positive weight is applied on the duration score is heavier, there will be more deletion errors and insertion errors.

Similar phenomenon can be observed from the experiment results on female data. As shown in Table 6-11 and Table 6-12, most of the reduced errors are insertion errors and most of the newly introduced errors are deletion errors.

Table 6-11: Reduced recognition errors and newly introduced ones (female: AM+TP)

	Reduced	Newly introduced
Del	0	33
Sub	9	12
INS	84	0

Table 6-12: Reduced recognition errors and newly introduced ones (female: AS)

	Reduced	Newly introduced
Del	7	25
Sub	15	11
INS	78	11

Insertion penalty is usually used to adjust the balance of insertion errors and increase the deletion errors. Many trials have been carried out with different insertion penalty on male speech data. With a value of -18 the highest accuracy can be obtained. Table 6-13 gives the recognition performance of using duration penalty with the empirical weight -18 on male and female speech data respectively.

Table 6-13: Recognition performance with insertion penalty

	Digit acc (%)	Sent acc (%)	Del	Sub	Ins
male	97.46	91.64	122	66	66
female	98.35	94.14	58	47	67

6.1.4. Recognition performance at different speaking rates

Using the same threshold of defining speaking rate for training data, we divide the test data into three categories. Table 6-13 shows the baseline recognition performance for each speaking rate category. Table 6-14 and Table 6-15 give the recognition performance with duration models. All these results are on male speech data.

Table 6-14: Baseline recognition performance at different speaking rates (male)

	Digit acc%	Sen acc%	Del	Sub	INS
Fast	96.00	85.99	67	54	29
Medial	97.57	91.34	16	15	81
Slow	96.37	91.06	3	0	71

Table 6-15: Recognition performance at different speaking rates (male: AW+TP)

	Digit acc%	Sen acc%	Del	Sub	INS
Fast	95.78	85.14	100	47	11
Medial	99.02	96.38	17	14	14
Slow	98.87	97.21	4	0	19

By comparing Table 6-14 and Table 6-15, it is seen that the recognition performance for the fast category drops from 96.00% to 95.78% due to a large number of newly introduced deletion errors. Recognition performances on the medial category and slow category are improved because of significant reduction on insertion errors. This trade-off is expected. Fast speech has shorter digit durations, which is not preferred by our duration models.

Similar phenomenon is observed in the experimental results on the use of AS model for male data. As shown in Table 6-15, the medial category and slow category achieve better performance than Fast category.

Table 6-16: Recognition performance for different speaking rate category (male: AS)

	Digit acc%	Sen acc%	Del	Sub	INS
Fast	95.89	95.24	95	49	10
Medial	99.04	96.47	13	15	16
Slow	98.67	96.79	1	0	26

Figure 6-1 shows the recognition performance on different speaking rate category for experiments on female data. It is also observed that the medial and slow category achieves better performance than the fast category.

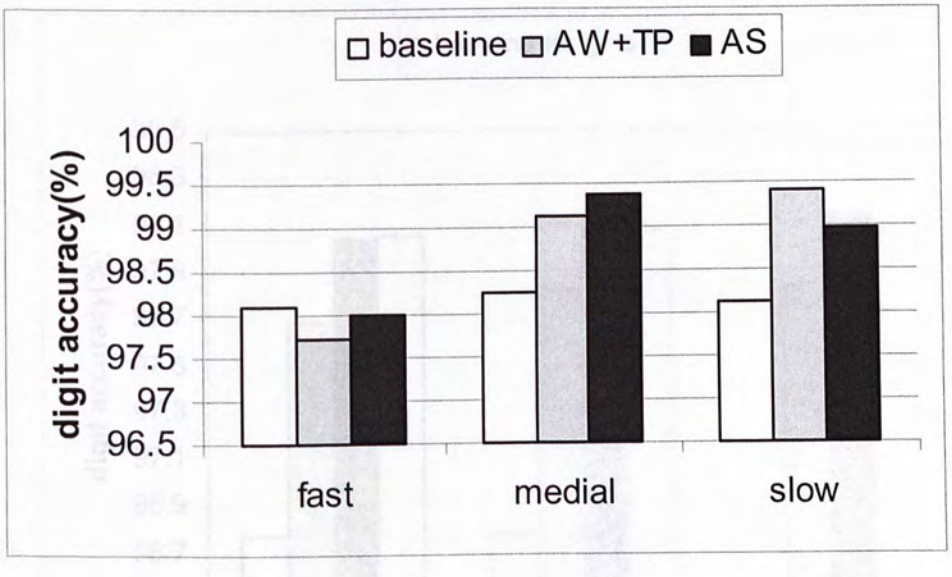


Figure 6-1: Recognition performance at different speaking rates (female)

6.2. Experiments with speaking-rate-dependent duration models

Experiments with speaking-rate-dependent duration models include two runs of decoding. The speaking rate is estimated based on the detected number of digits per second in a preliminary run of decoding. Corresponding duration model is then integrated into the second run of decoding.

6.2.1. Using true speaking rate

To probe the up bound of the performance improvement with speak-rate-dependent models, we use the true speaking rate in an oracle experiment instead of using estimated speaking rate first.

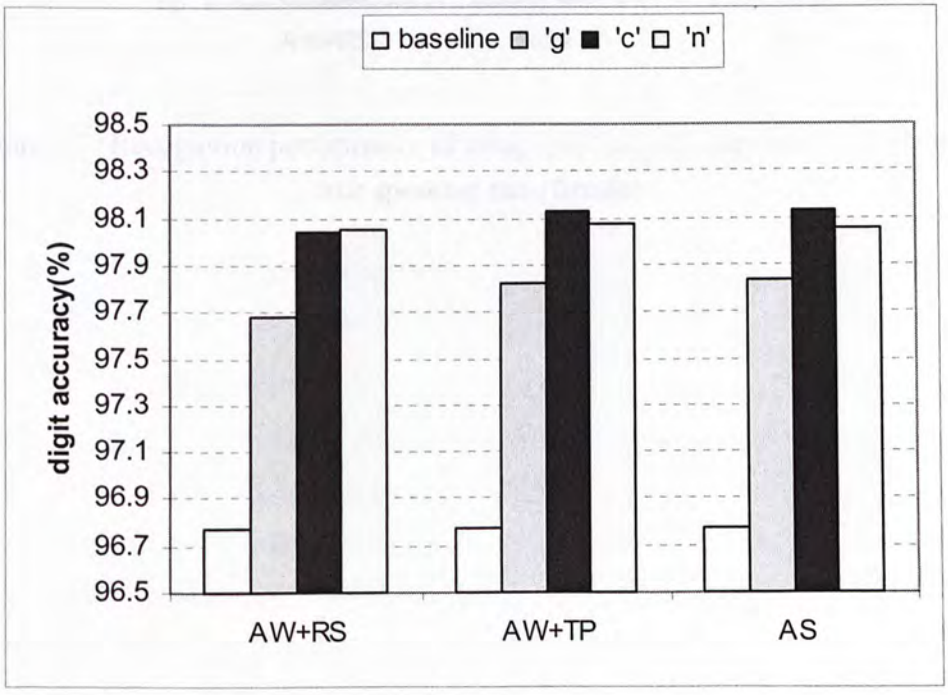


Figure 6-2: Recognition performance of using speaking-rate-dependent models with true speaking rate (male)

Figure 6-2 and Figure 6-3 give recognition performance of using speaking-rate-dependent models with true speaking rate. The recognition performance of

baseline and using speaking-rate-independent duration models is shown for compare. As shown in the figures, noticeable improvement is achieved by using speaking-rate-dependent models in comparison to using speaking-rate-independent duration models.

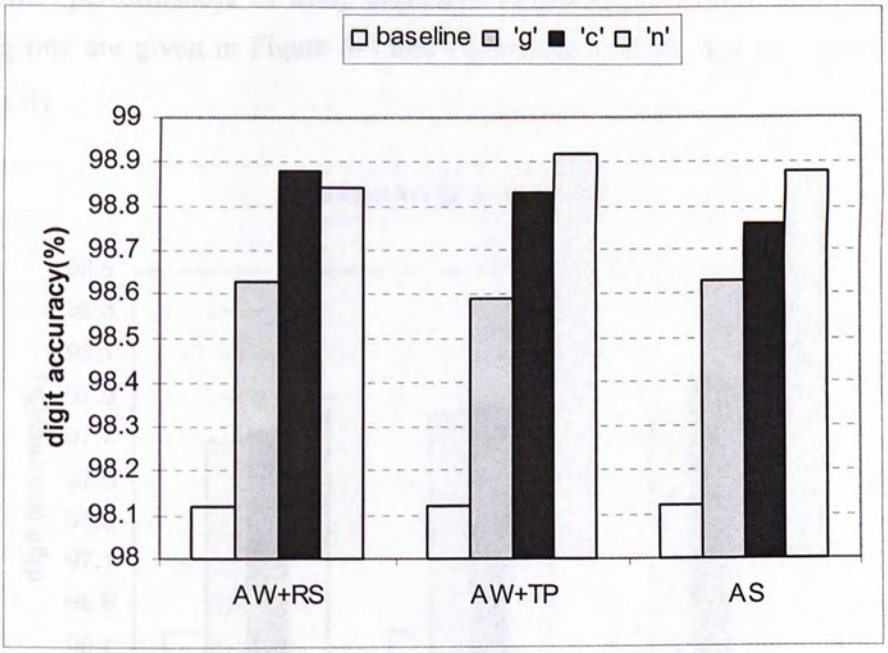


Figure 6-3: Recognition performance of using speaking-rate-dependent models with true speaking rate (female)

6.2.2. Using estimated speaking rate

Recognition performances of using speaking-rate-dependent models with estimated speaking rate are given in Figure 6-4 and Figure 6-5 on male data and female data respectively.

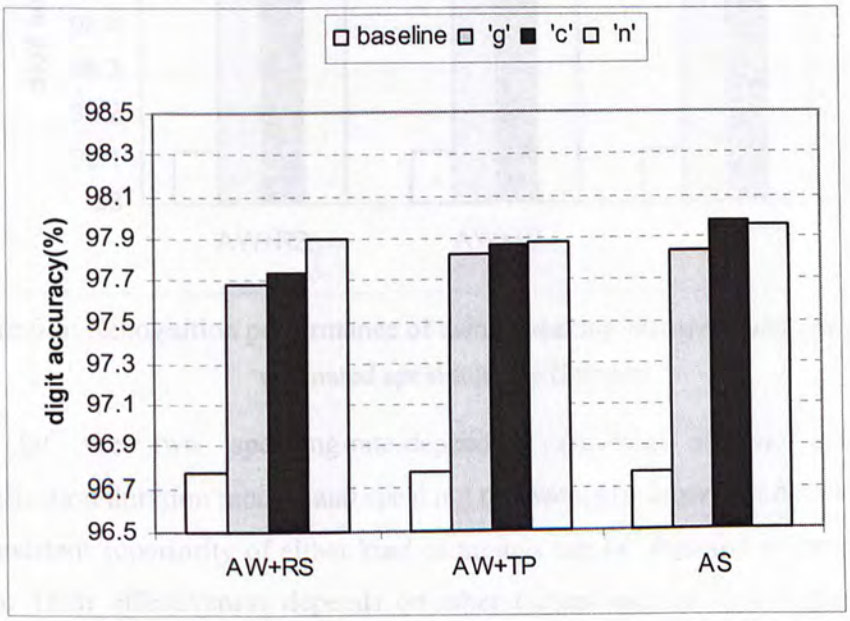


Figure 6-4: Recognition performance of using speaking-rate-dependent models with estimated speaking rate (male)

The speaking-rate-dependent duration models do offer improvement over speaking-rate-independent duration models. It indicates that estimated speaking rate information is useful to some extent in guiding the second run of decoding. However, the improvement is limited as compared to the upper bound. More accurate speaking rate information is necessary for effective use of the speaking-rate-dependent models. Nevertheless, accurate speaking rate is difficult to be obtained currently. This is the limitation of using speaking-rate-dependent duration models.

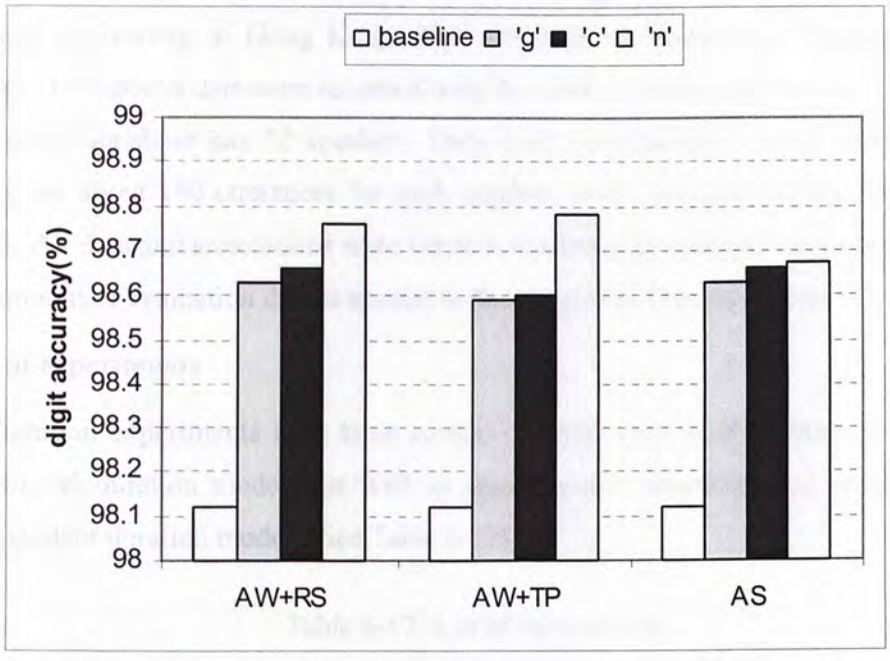


Figure 6-5: Recognition performance of using speaking-rate-dependent models with estimated speaking rate (female)

Of the two speaking-rate-dependent duration models, speaking-rate normalization duration models and speaking rate category dependent duration models, no consistent superiority of either kind of models can be observed in our experiment results. Their effectiveness depends on other factors such as speech data, duration feature and so on.

6.3. Evaluation on another speech database

In Section 6.1, the weights used for duration model score were trained by many trials on test data set. The values are therefore data specific. One may argue that the recognition results could not fully reveal the effectiveness of our method. In this section, further experimental work will be presented to address this issue.

6.3.1. Experimental setup

Evaluation data

In the following experiments, the evaluation data are not from CUDIGIT. Instead,

they are a subset of speech data for speaker recognition recently collected at the Chinese University of Hong Kong. This subset is a collection of Cantonese digit strings. The speech data were recorded with the same acoustic condition as CUDIGIT. The entire database has 52 speakers. Only 5 of them are used in our experiments. There are about 180 utterances for each speaker. Each utterance contains exactly 14 digits. All data and annotations were verified manually. In terms of numbers of digits, the amount of evaluation data is similar to that of used in Section 6.1 and 6.2.

List of experiments

Recognition experiments have been carried out with state-level duration model and word-level duration models, as well as speaking-rate dependent and speaking-rate independent duration model. (See Table 6-17)

Table 6-17: List of experiments

	Word-level duration model		State-level duration model
'g'	AM+RS	AM+TP	AS
'c'	AM+RS	AM+TP	AS
'n'	AM+RS	AM+TP	AS

Weights for duration models

The best weights developed from CUDIGIT in the previous sections as listed in Table 6-18 are used.

Table 6-18: The best weights obtained for male data of CUDIGIT

	Insertion Penalty	AW+RS	AW+TP	AS
Weight	-18	(6,2)	(6,4)	3

6.3.2. Experiment results and analysis

6.3.2.1 Experiments with speaking-rate-independent duration models

Table 6-19: Recognition performance with different duration features

	Digit acc%	Del	Sub	Ins
Baseline	95.09	82	116	418
AW+RS	97.22	142	90	116
AW+TP	97.24	133	90	124
AS	97.45	105	88	127
Insertion penalty	96.37	124	117	215

In the baseline recognition system, insertion and deletion errors account for more than 80% of the recognition errors. Seen from confusion matrix in Table 6-20, 68.2% of insertion errors and deletion errors are due to "2" and "5".

Table 6-20: Confusion matrix of baseline recognition results

	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	del
'0'	1333	1	0	0	0	0	0	0	0	0	1
'1'	12	1236	3	1	3	8	0	4	0	0	**7**
'2'	2	0	1274	0	0	0	0	0	1	0	**16**
'3'	3	0	0	1196	0	4	0	4	1	2	0
'4'	2	0	2	0	1288	0	0	1	0	0	0
'5'	5	0	1	1	0	1133	1	0	0	3	**51**
'6'	3	0	0	0	0	0	1268	0	4	3	6
'7'	0	0	0	5	4	4	0	1205	0	1	1
'8'	1	0	0	2	0	1	2	1	1301	7	0
'9'	2	3	0	1	0	0	5	0	**2**	1124	0
ins	**75**	4	**55**	1	2	**224**	4	14	5	34	

Table 6-19 gives the recognition results on the new evaluation data with different duration features. The result with insertion penalty is also shown for comparison. Obviously, the use of explicit duration model results in noticeable improvement over the baseline. State-level models show better performance than use of word-level models. The combination of AW and RS attains similar improvement to the combination of AW and TP.

Similar to the observation in Section 6.1 and 6.2, the insertion errors are greatly reduced while the deletion errors are increased at the same time. Additionally, the substitution errors are reduced in all cases. Use of insertion penalty can also improve the recognition accuracy. However, it is less effective than the proposed duration models.

Figure 6-6 shows the recognition performance for each speaking rate category on the new evaluation data.

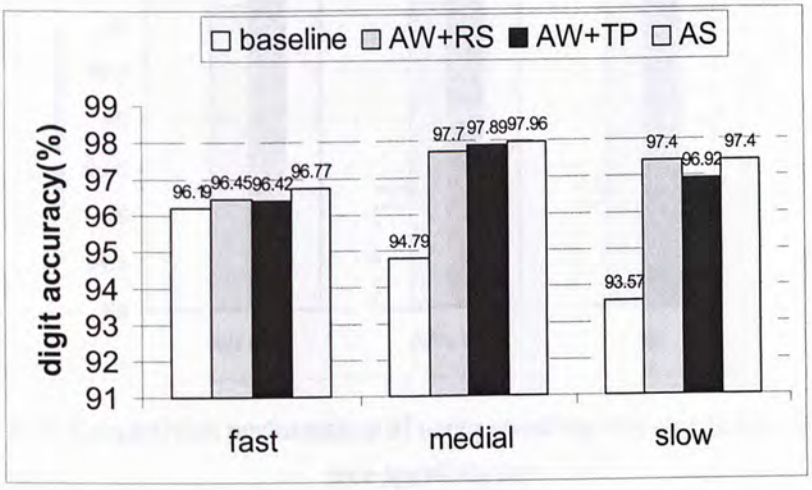


Figure 6-6: Recognition performances for different speaking rates

Recognition performances are improved for all categories. The slow category achieves the most significant improvement while the fast category achieves marginal improvement due to a large number of new deletion errors. The improvement of the medial category is also significant.

6.3.2.2 Experiments with speaking-rate-dependent duration models

Experiments with speaking-rate-dependent duration models include two runs of decoding. The speaking rate is estimated in a preliminary run of decoding, and the respective duration model is then integrated into the second run of decoding.

Using true speaking rate

To know the up bound of the performance improvement with speak-rate-dependent models, we use the true speaking rate in an oracle experiment, and the results are given in

Figure 6-7. Noticeable improvement can be observed on the use of speaking-rate-dependent models as compared to using speaking-rate-independent duration models.

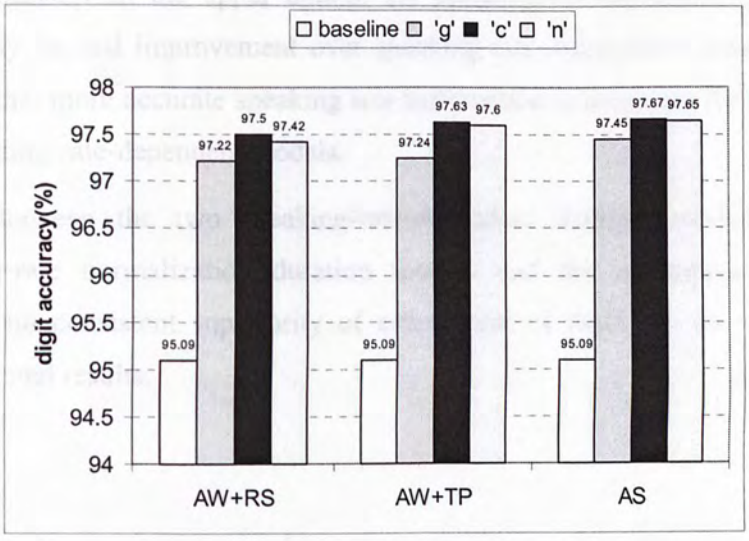


Figure 6-7: Recognition performance of using speaking-rate-dependent models with true speaking rate

Using estimated speaking rate

Recognition performances of using speaking-rate-dependent models with estimated speaking rate are given in Figure 6-8 .

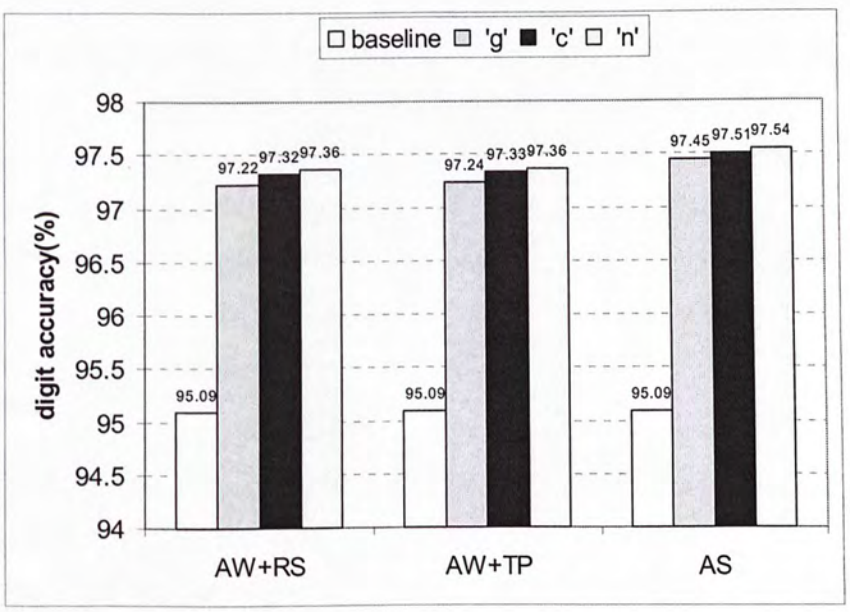


Figure 6-8: Recognition performance of using speaking-rate-dependent models with estimated speaking rate

As compared with the upper bound, the speaking-rate-dependent duration models offer only limited improvement over speaking-rate-independent duration models. It implies that more accurate speaking rate information is necessary for effective use of the speaking-rate-dependent models.

Between the two speaking-rate-dependent duration models, namely, the speaking-rate normalization duration models and the category-specific duration models, no consistent superiority of either kind of them can be observed in our experimental results.

Chapter 7

Conclusions and future work

7.1. Conclusion and understanding of current work

The basic assumptions of HMM are inappropriate to characterize the temporal structure of speech signals. It does not give effective control on both the time duration and relative duration of speech segments that it models. For HMM-based Cantonese digit recognition, it is observed that a significant portion of recognition errors exhibit unreasonable duration properties.

We have investigated on the use of both absolute duration models and relative duration models to confine the duration of recognized digits. In particular, the relative duration of the tail part of a Cantonese digit has been proposed. Algorithms have developed to incorporate state-level duration score and word-level duration score to HMM-based recognition process. For each decision on path extension, the duration models are used to contribute an additional probabilistic score to the conventional path score. Use of word-level duration model is computationally much expensive than state-level duration models. In the decoding algorithm, a weighting factor is used to balance the contribution of acoustic model and duration model.

Experiments have been carried out on CUDIGIT corpus. A set of empirical weights for different duration information are obtained by many trials on male test data set. With these weights, use of different duration information shows performance improvement of various degrees. For male speech, the recognition accuracy is improved by up to 1.06%. For female data, the recognition accuracy is improved by up to 0.51%.

Among the three word-level features, the relative state duration (RS) and the absolute word (AW) consistently offer noticeable improvement. The relative duration of tail part (TP) provides noticeable improvement only on male speech. TP is not as good as RS for relative duration modeling. Relative duration features are shown to be a good supplement to the absolute duration features. If the AW model and the RS model or TP model are used in combination, further improvement can be achieved.

The use of state-level duration feature AS always shows better recognition performance than any of the word-level duration feature. On the other hand, the combined use of word level duration features provides comparable recognition performance to the use of AS.

With the use of duration model, the performance improvement is mainly reflected in the reduction of insertion errors. Slow speech benefits more than fast speech. In the case of fast speech, new errors are introduced. Use of explicit duration model attains higher accuracy than the use of word insertion penalty.

Regarding that speaking rate variation may weaken the effectiveness of the duration model. Speaking-rate-dependent duration models are suggested. With the estimated speaking rate by a preliminary run of decoding, use of speaking-rate-dependent model shows further improvement in comparison to the speaking-rate-independent models. However, the improvement is not significant. Further improvement on the characterization of speaking rate information is needed for effective use of the speaking-rate-dependent model.

In view of that those empirical weights are data specific, further evaluation experiments have been carried out on a different set of speech data. The experiments results are consistent with earlier observations. The improvement of recognition accuracy is up to 2.36%, which is achieved by the use of AS. The most significant improvement is observed on slow speech. Using speaking-rate-dependent duration model consistently shows further improvement.

7.2. Future work

Explicit duration modeling would be useful for ASR under noise conditions. As we mentioned in experiment result, insertion errors related to noise are reduced a lot with duration modeling. Noise segments do not posses regular duration properties like speech units. If the noise segments are misrecognized as speech units, the recognized speech units tend to have unreasonable duration properties. In this regard, explicit duration modeling would be helpful.

A Appendix

1) Plots for AS

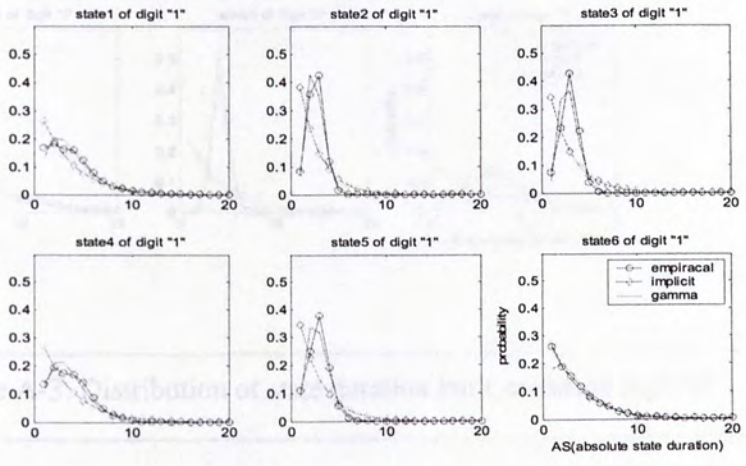


Figure A-1: Distribution of state duration For Cantonese digit "1"

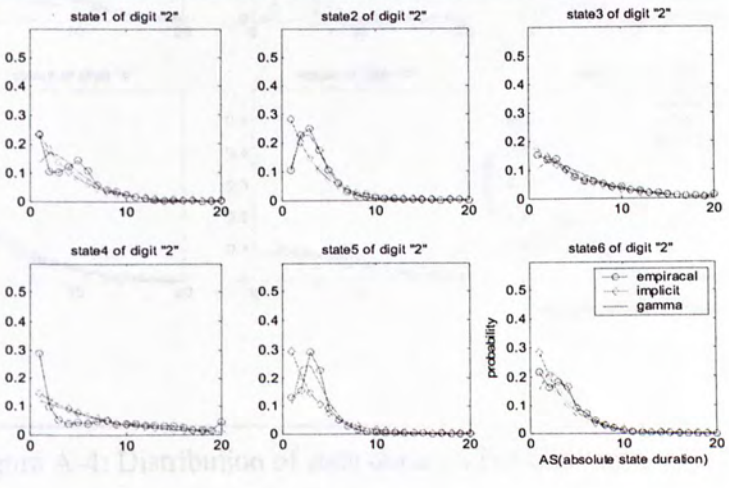


Figure A-2: Distribution of state duration For Cantonese digit "2"

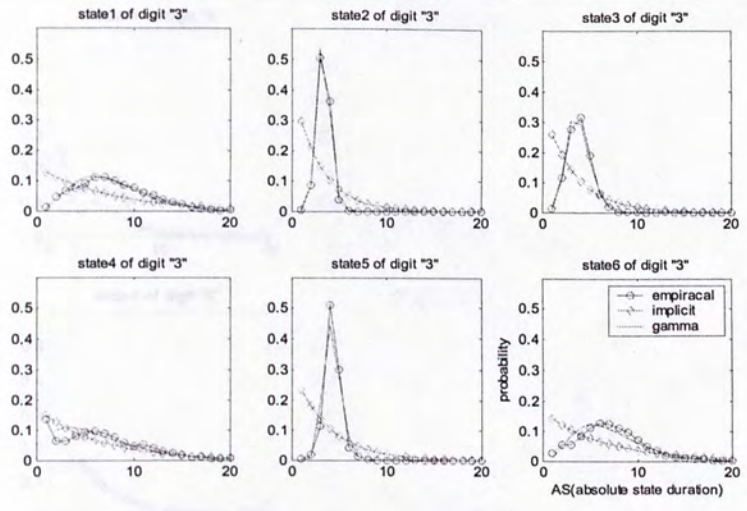


Figure A-3: Distribution of state duration For Cantonese digit "3"

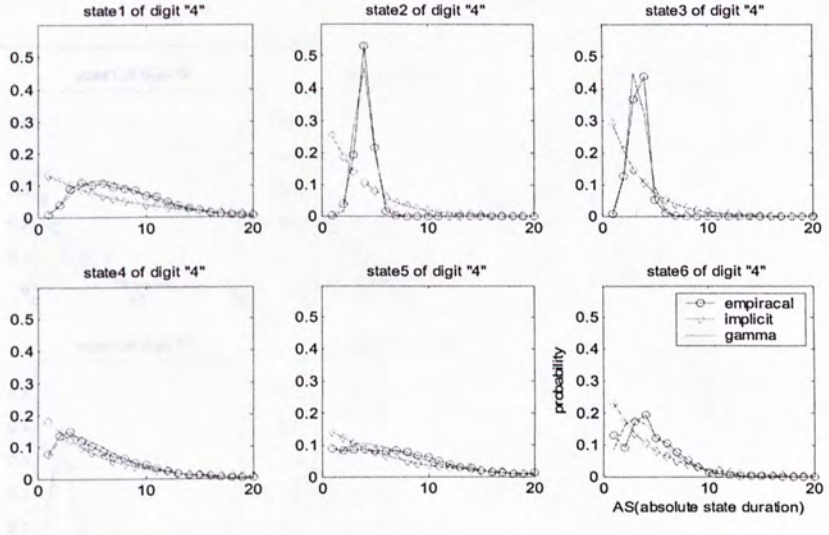


Figure A-4: Distribution of state duration For Cantonese digit "4"

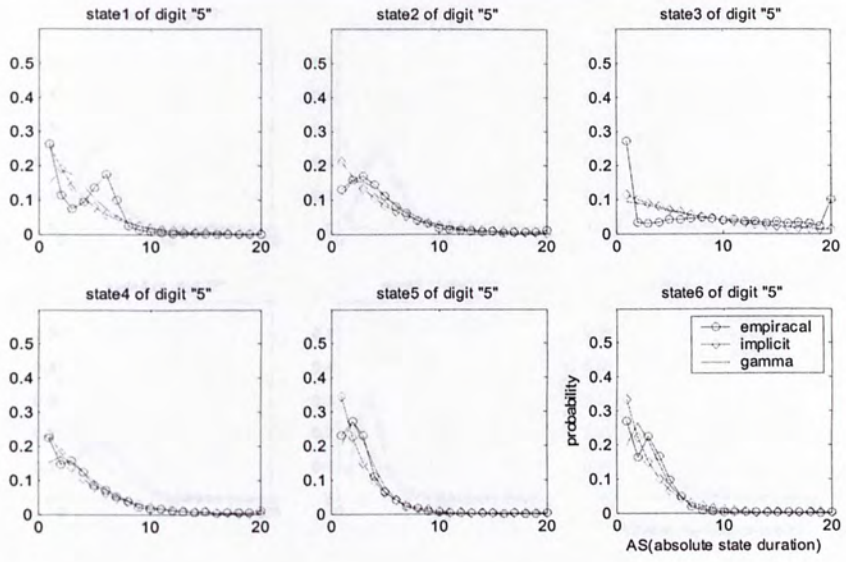


Figure A-5: Distribution of state duration For Cantonese digit "5"

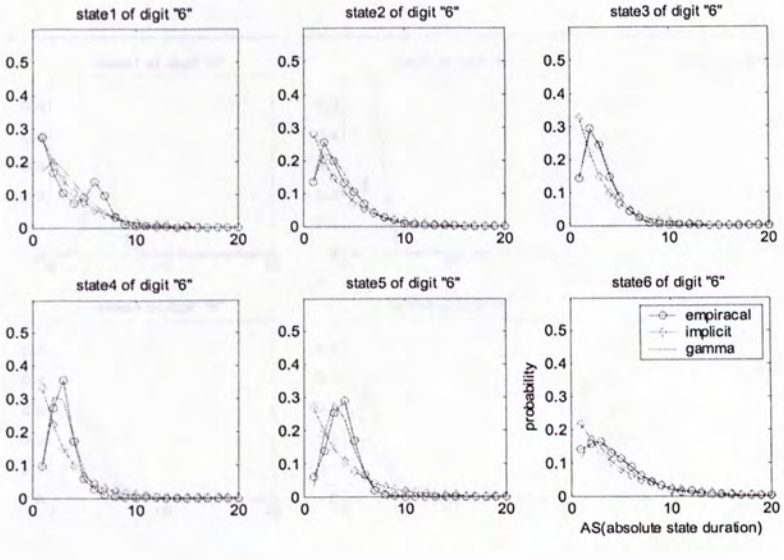


Figure A-6: Distribution of state duration For Cantonese digit "6"

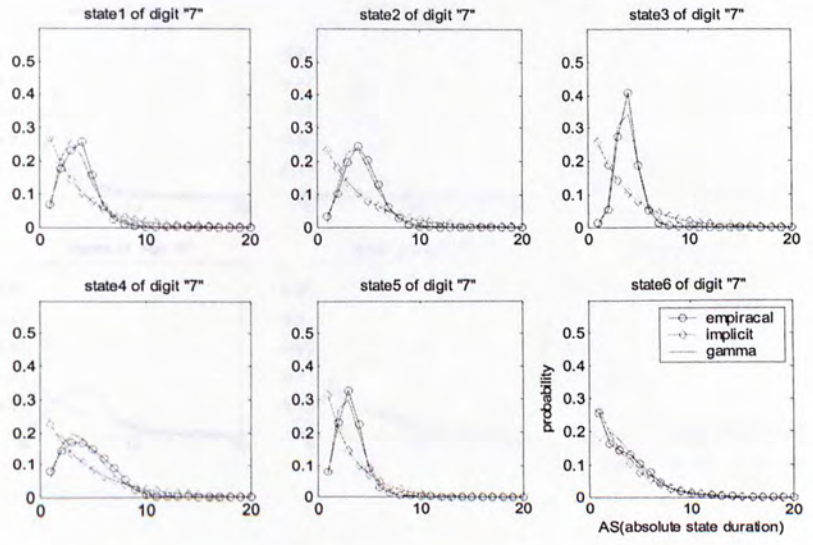


Figure A-7: Distribution of state duration For Cantonese digit "7"

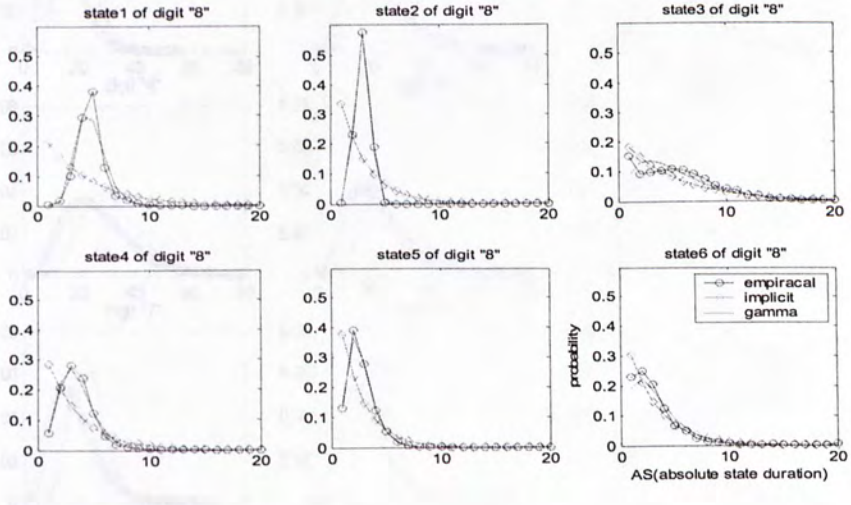


Figure A-8: Distribution of state duration For Cantonese digit "8"

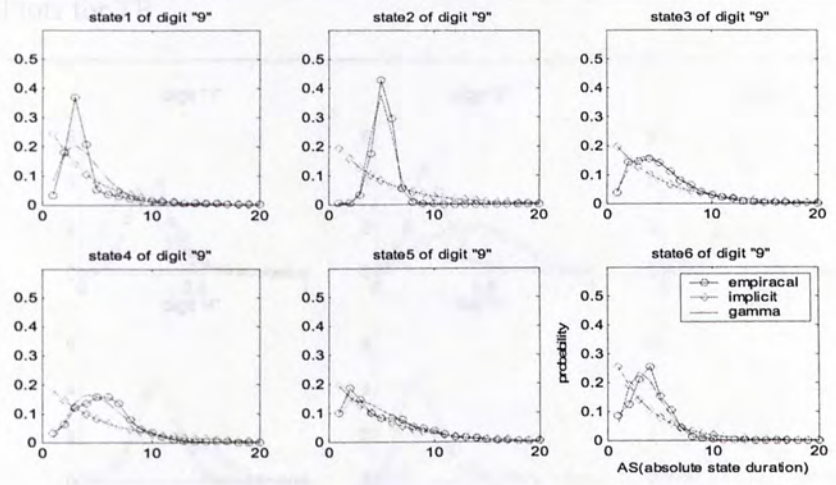


Figure A- 9:Distribution of state duration For Cantonese digit "9"

2) Plots for AW

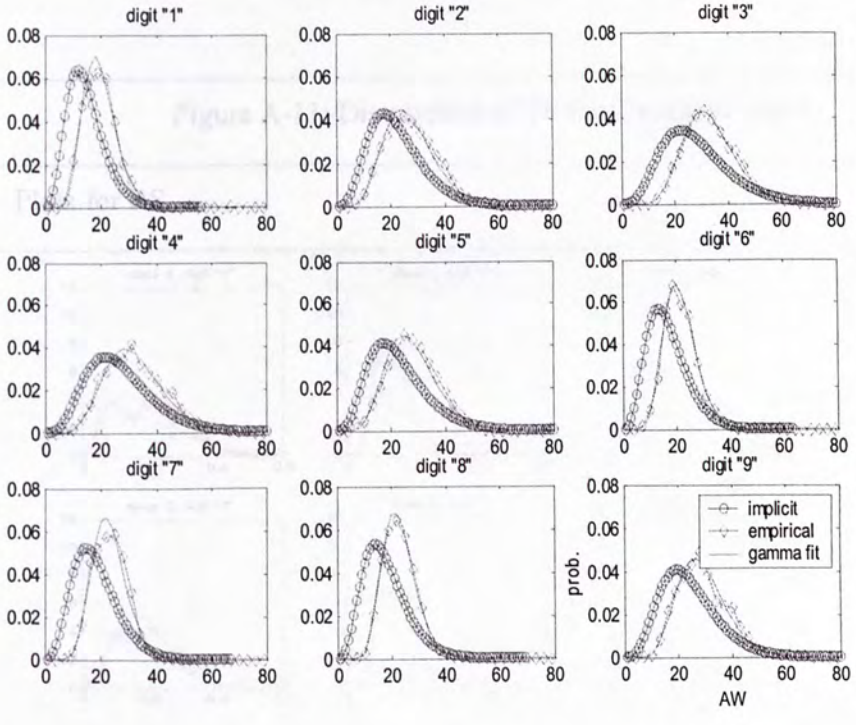


Figure A-10: Distribution of AW For Cantonese digits

3) Plots for TP

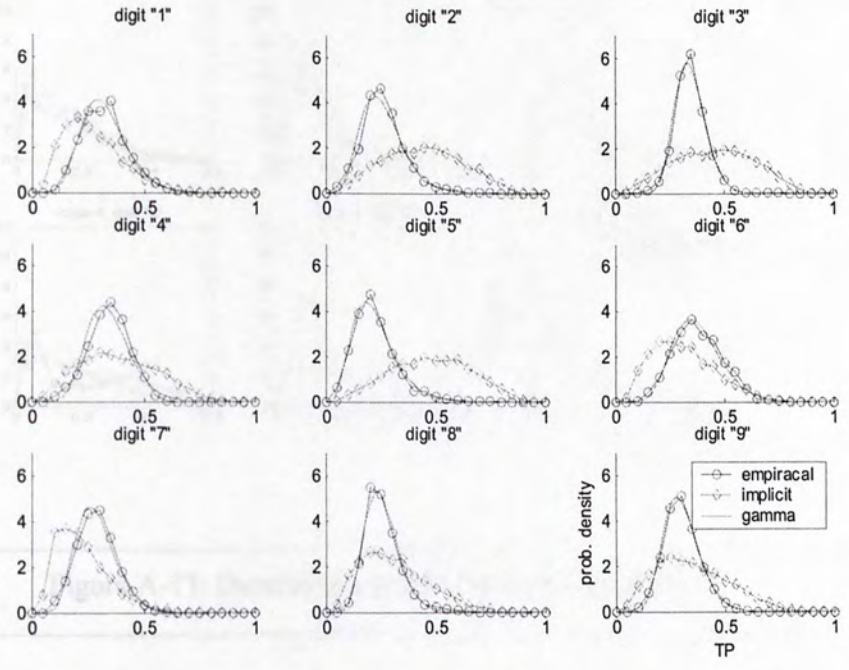


Figure A-11: Distribution of TP For Cantonese digits

4) Plots for RS

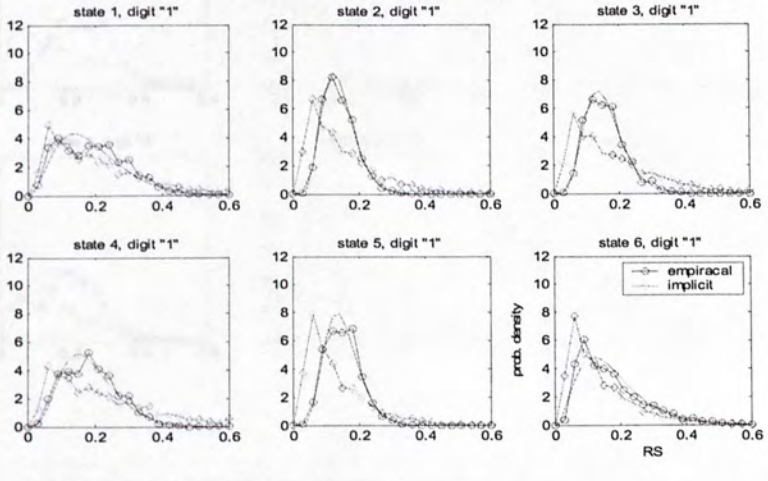


Figure A-12:Distribution of RS For Cantonese digit "1"

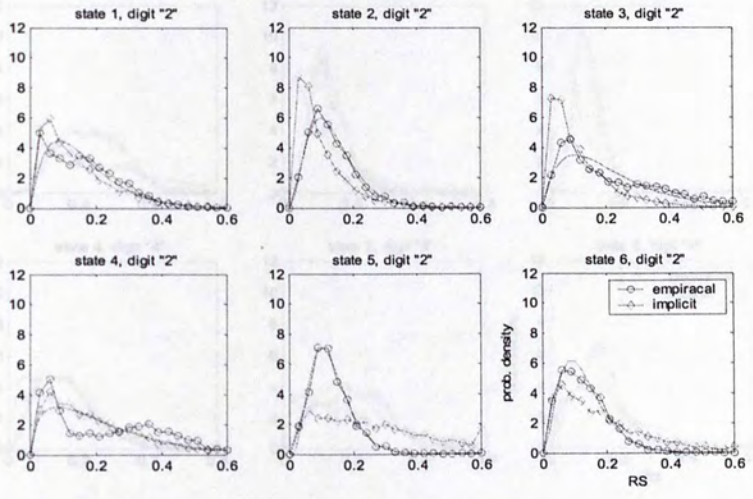


Figure A-13: Distribution of RS For Cantonese digit "2"

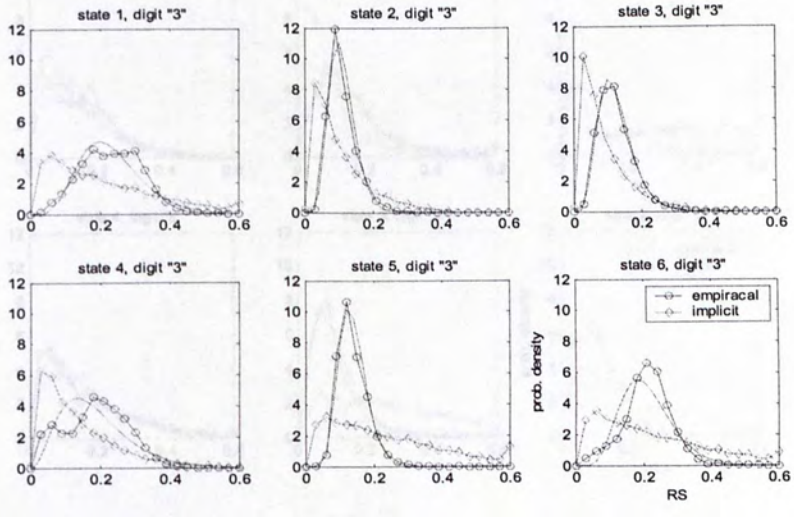


Figure A-14: Distribution of RS for Cantonese digit "3"

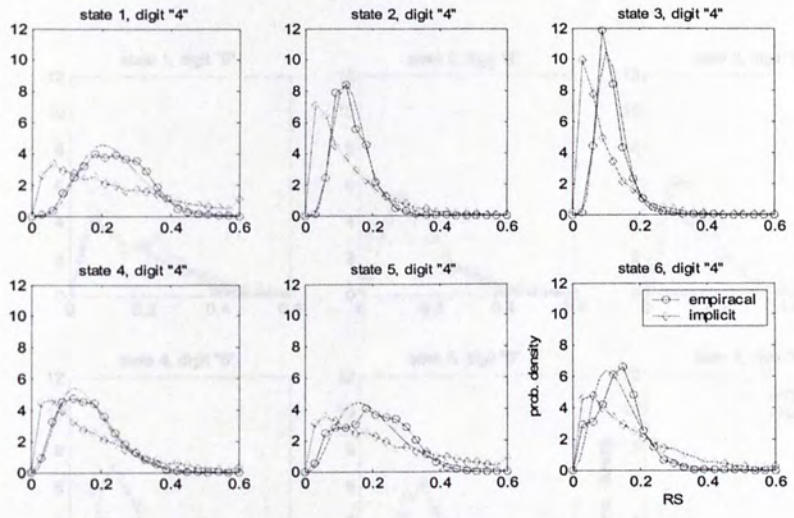


Figure A-15: Distribution of RS for Cantonese digit "4"

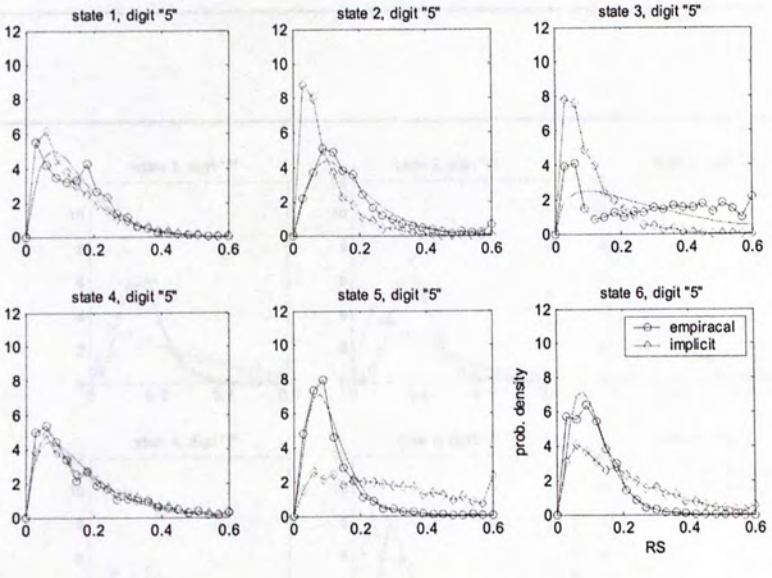


Figure A-16: Distribution of RS For Cantonese digit "5"

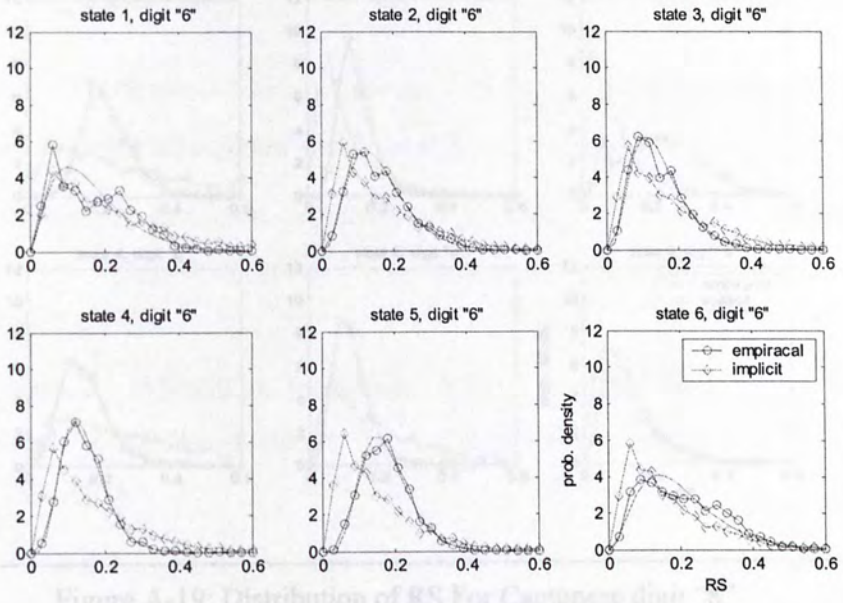


Figure A-17: Distribution of RS For Cantonese digit "6"

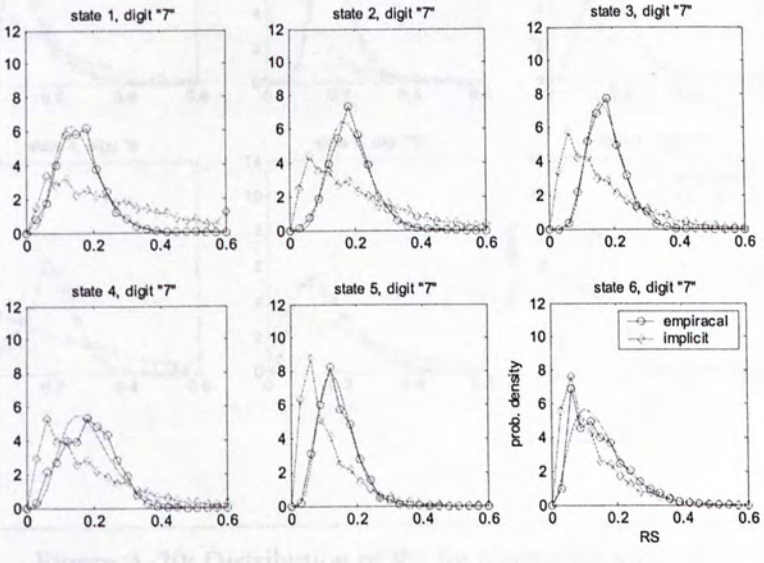


Figure A-18: Distribution of RS For Cantonese digit "7"

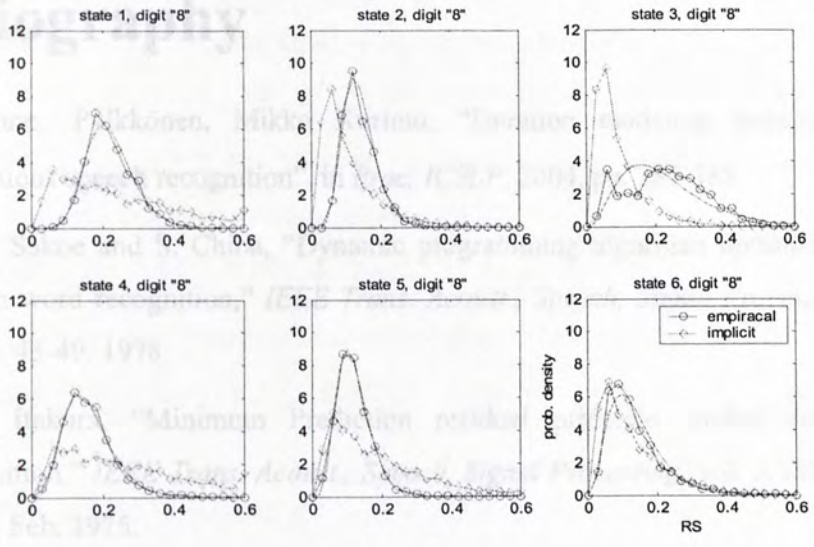


Figure A-19: Distribution of RS For Cantonese digit "8"

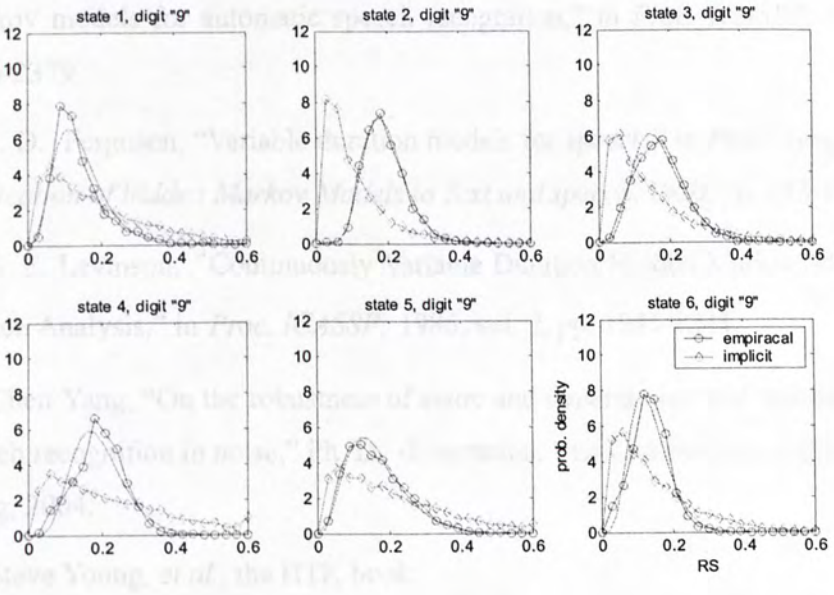


Figure A-20: Distribution of RS for Cantonese digit "9"

Bibliography

[1] Janne, Pylkkönen, Mikko Kurimo, "Duration modeling techniques for continuous speech recognition", in *Proc. ICSLP*, 2004, pp. 385-388.

[2] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 43-49, 1978.

[3] F. Itakura, "Minimum Prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 57-72, Feb. 1975.

[4] C. Myers, L.R. Rabiner, and A.E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 623-635, Dec. 1980.

[5] M. Russell and R. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP*, 1985, pp. 2376-2379.

[6] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. On the application of hidden Markov Models to Text and speech*, 1980, pp. 143-179

[7] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Speech Analysis," in *Proc. ICASSP*, 1986, vol. 2, pp. 1241-1244.

[8] Chen Yang, "On the robustness of static and dynamic spectral information for speech recognition in noise," Ph. D. dissertation, The Chinese University of Hong Kong, 2004.

[9] Steve Young, *et al.*, the HTK book.

[10] L.R.Rabiner, B.H. Juang, *Fundamentals of speech recognition*, New Jersey: Prentice Hall, 1993.

[11] Viterbi, A. ,"Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. Information Theory*, vol. 13, pp. 260 – 269, Apr. 1967.

[12] R.E. Bellman, *Dynamic programming*. Princeton University Press, 1957.

[13] H. Sakoe, "Two-Level DP matching-a dynamic programming-based pattern matching algorithm for connected word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27 (6), pp. 588-595, Dec. 1979.

[14] C.S. Myers and L.R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29 (2), pp. 284-297, Dec. 1981.

[15] T.K. Vintsyuk, "Element-Wise Recognition of Continuous Speech consisting of Words From a Specified Vocabulary," *Kibernetika*, No. 2, pp. 133-143, Mar.-Apr. 1971.

[16] J.S. Bridle, M.D. Brown, and R.M. Chamberlain, "An algorithm for connected word recognition," in *Proc. ICASSP*, 1982, pp. 899-902.

[17] H. Ney, "The use of a one-stage dynamic programming algorithm for connected Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32 (2), pp. 263-271, April. 1984.

[18] Hung-yan Gu, Chiu-yu Tseng and Lin-shan Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with bounded State Duration," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1743-1751, Aug. 1991.

[19] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, Feb. 1989.

[20] D. Burshtein. "Robust parametric modeling of durations in Hidden Markov Models," in *Proc. ICASSP*, 1995, pp 548-551

[21] O. W. Kwon, C. K. Un, "Performance of connected digit recognizers with context-dependent word duration modeling," in *Proc. APCCAS*, 1996, pp. 243-246.

[22] K. Power, "Duration modeling for improved connected digit recognition," in *Proc. ICSLP*, 1996, pp. 885-888.

[23] T.K. Vintsyuk, "Speech Discrimination by Dynamic programming," *Kibernetika*, 4 (2), pp. 81-88, Jan.-Feb. 1968.

[24] Russell M., Cook A., "Experimental evaluation of duration modeling techniques for automatic speech recognition", in Proc. ICASSP, 1987, pp. 2376 – 2379.

[25] P. Ramesh and J. Wilpon, "Modeling state durations in Hidden Markov Models for automatic speech recognition," in *Proc. ICASSP*, 1992, pp. 381-384.

[26] Rong Dong, Jie Zhu, "On use of duration modeling for continuous digits speech recognition," in *Proc.ICSLP,* 2002, pp. 385-388.

[27] Tan Lee, Rolf Carlson and Björn Granström, "Context-dependent duration modeling for continuous speech recognition," in *Proc. ICSLP,* 1998, pp.2955-2958.

[28] V. R. R. Gadde, "Modeling word duratons,", in *Proc. ICSLP*, 2000, pp. 601-604.

[29] Myoung-wan Koo, Sung-Joon Park, Dan-Young Son, "Context dependent phoneme duration modeling with tree-based state tying," in *Proc. ICSLP*, 2004, pp. 721-724.

[30] M. Jones and P. C. woodland, "Using relative duration in large vocabulary speech recognition," in *Proc. EUROSPEECH*, 1993, pp. 311-314.

[31] A. Anastasakos, R. Schwartz, Shu Han, "Duration modeling in large vocabulary speech recognition," in *Proc. ICASSP*, 1995, pp. 628-631.

[32] Wern-Jun Wang, Chun-Jen Lee, "Duration modeling for Mandarin speech recognition using prosodic information," in *Proc. Speech Prosody*, 2004, pp.591-594.

[33] Gang Peng, Bo Zhang and William S.-Y. Wang, "Duration modeling in Mandarin connected digit recognition," in *Proc. ISCSLP*, 2000, pp. 227-30.

[34] Gang Peng, William S.-Y. Wang, "An innovative prosody modeling method for Chinese speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2-3, pp. 129-140, Apr. 2004.

[35] Gang Peng, "Reliability index guided prosody modeling in speech recognition," Ph. D. dissertation, The City University of Hong Kong, 2004.

[36] Yu Zhu, Tan lee, "Explicit Duration Modeling for Cantonese Connected-digit Recognition," in *Proc. ICSLP*, 2004, pp. 685-688.

[37] L.R. Rabiner, J. G. Wilpon and F. K. Soong, "High performance connected digit recognition using hidden markove models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1214-1225, Aug. 1989.

[38] C.-H. Lee and L. Rabiner, "A frame-synchronous network search algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1649-1658, Nov. 1989.

[39] Yao Qian, Tan Lee and Frank K. Soong, "Use of tone information in continuous Cantonese speech recognition," in *Proc. of Speech Prosody*, 2004, pp.587- 590.

[40] L.Rodriguez-linares, C.Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task,", in *Proc. ICSLP*, 1998, pp. 1084-1087.

[41] Heckmann M., Berthommier F., Kroschel K., "Optimal weighting of posteriors for audio-visual speech recognition," in *Proc. ICASSP*, 2001, pp. 161-164.

[42] C. Chesta, P. Laface and F. Ravera, "Connected digit recognition using short and long duration models," in *Proc. ICASSP*, 1999, pp. 557-560.

[43] Jen-Tzung Chine,chih-hsien Huang, "Bayesian learning of speech duration models," *IEEE Trans. Speech Audio Processin*, vol. 11, pp. 558-567, Nov. 2003.

[44] J. P. Nedel and R.M. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. ICASSP*, 2001, pp. 313-316

[45] J. P. Nedel and R.M. Stern, "Duration normalization and Hypothesis Combination for improved spontaneous speech recognition," in *Proc. EUROSPEECH*, 2003, pp. 1509-1512.

[46] Yonggang Deng, Taiyi Huang and Bo xu, "Towards high performance continuous mandarin digit string recognition," in *Proc. ICSLP*, 2000, pp. 642-645.

[47] Yun Tang, Wenju Liu and Bo Xu, "Trigram duration modeling in speech recognition," in *Proc. ISCSLP*, 2004, pp. 225-228.

[48] Chao Huang and Daowen Chen, "High performance Chinese continuous digits Recognition system," in *Proc. ICCC*, 1996, pp. 21-25.

[49] J.-K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, pp. 206-216, Jan. 1994.

[50] Tan Lee, W.K. Lo, P.C. Ching and Helen Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol.36, No.3-4, pp. 327-342, Mar. 2002.

[51] W. K. Lo, "Cudigit readme," DSP Lab, Dept. of Electronic Engineering, The Chinese University of Hong Kong, 1999

[52] Tan Lee, W.K. Lo, P.C. Ching, "Development of Cantonese spoken language corpora for speech applications", in *Proc. ISCSLP,* 1998, pp. 102-107.

[53] W.N. Campbell, "Extacting speech-rate values from a real-speech datatbase," in *Proc. ICASSP*, 1988, pp. 683-686.

[54] N. Miraghafori, E. Fosler, N.Morgan, "Towards robustness to fast speech in ASR," in *Proc. ICASSP*, 1996, pp. 335-338.

[55] M.A. Siegler, R.M. Stern, "On the effects of speech rate in large vocabulary speech recognition system," in *Proc. ICASSP*, 1995, pp. 612-615.

[56] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes," in *Proc. EUROSPEECH*, 1995, pp. 491-494.

[57] N. Mirghafori, E. Fosler, and N. Morgan, "Speech Recognition using on-line estimation of Speaking Rate," in *Proc. Eurospeech,* 1997, pp. 2079-2082.

[58] T. Pfau, G. Ruske, "Estimating the speaking rate by vowel detection", in *Proc. ICASSP,* 1998, pp. 945-948.

[59] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, *Spoken language processing: A Guide to Theory, Algorithm and system development.* Carnegie Mellon University, 2001.