Computational Linguistics and Chinese Language Processing
Vol. 11, No. 1, March 2006, pp. 1-16

Using Duration Information in Cantonese Connected-Digit Recognition

Yu Zhu* and Tan Lee*

Abstract

This paper presents an investigation on the use of explicit statistical duration models for Cantonese connected-digit recognition. Cantonese is a major Chinese dialect. The phonetic compositions of Cantonese digits are generally very simple. Some of them contain only a single vowel or nasal segment. This makes it difficult to attain high accuracy in the automatic recognition of Cantonese digit strings. Recognition errors are mainly due to the insertion or deletion of short digits. It is widely admitted that the hidden Markov model does not impose effective control on the duration of the speech segments being modeled. Our approach uses a set of statistical duration models that are built explicitly from automatically segmented training data. They parametrically describe the distributions of various absolute and relative duration features. The duration models are used to assess recognition hypotheses and produce probabilistic duration scores. The duration scores are added with an empirically determined weight to the acoustic score. In this way, a hypothesis that is competitive in acoustic likelihood, but unfavorable in temporal organization, will be pruned. The conventional Viterbi search algorithms for connected-word recognition are modified to incorporate both state-level and word-level duration features. Experimental results show that absolute state duration gives the most noticeable improvement in digit recognition accuracy. With the use of duration information, insertion errors are much reduced, while deletion errors increase slightly. It is also found that explicit duration models are more effective for slow speech than for fast speech.

Keywords: Explicit Duration Modeling, Duration Features, Connected-Digit Recognition, Cantonese, Hidden Markov Models

* Department of Electronic Engineering, The Chinese University of Hong Kong
Tel: 852-26098267 Fax: 852-26035558
E-mail: tanlee@ee.cunk.edu.hk
The author for correspondence is Tan Lee.

[Received June 15, 2005; Revised January 23, 2006; Accepted January 25, 2006]

1. Introduction

In the past two decades, automatic speech recognition (ASR) has advanced to a high performance level. The state-of-the-art technology predominantly uses hidden Markov models (HMM), which provide a nicely formulated framework for the modeling of speech signals. This framework is amenable to a set of mathematically rigorous algorithms for the estimation of model parameters and pattern classification. For ASR, an HMM consists of a number of states that are arranged into a left-to-right topology. The states can be thought of as a sequence of acoustic targets that constitute a speech segment. The output probability density functions (pdf) associated with individual states describe the spectral variability in the realization of these targets. The temporal structure is reflected mainly in the evolution of the states, which is governed by state transition probabilities.

It is widely acknowledged that an HMM does not impose effective control on the duration of the speech segment being modeled. HMM-based ASR systems frequently make errors. A significant portion of these recognition errors exhibit unreasonable time durations or duration proportions. For the task of connected-digit recognition in various languages in particular, a lot of errors are due to the insertion of short digits [Dong and Zhu 2002; Kwon and Un 1996]. The problem is extremely severe with noise-corrupted speech [Yang 2004].

Connected-digit recognition has many useful applications that often require very high recognition accuracies. Despite its limited vocabulary size, it is not straightforward to attain the desired performance level because the combination of digits is unrestricted. Knowledge sources like lexical constraints and word-level language models are not applicable in this case. Therefore, it becomes particularly important to fully exploit the information embedded in the acoustic signals. Other than the spectral features, prosodic features, like pitch and duration, can be considered.

In this paper, we focus on the use of duration information for Cantonese connected-digit recognition. Our approach uses a set of statistical duration models that are built explicitly from automatically segmented training data. The duration models are used to assess the recognition hypotheses, based on the measured duration at the either state or the model levels. As a result, a probabilistic duration score is generated and added with an empirically determined weight to the conventional acoustic score. In this way, a hypothesis that is competitive in acoustic likelihood, but unfavorable in temporal organization, is pruned.

There have been many studies on explicit duration modeling for ASR. Recognition performance can be improved to various extents. The most commonly used duration features include whole-model duration [Lee *et al*. 1989], absolute state duration [Russell and Moore 1985; Levinson 1986] and normalized (relative) state duration [Rabiner 1989; Power 1996]. The design of duration models has been application-dependent. In most cases, parametric

distributions have been used so that each duration model can be represented by a few parameters.

HMM based speech recognition is formulated as a process of searching for the optimal path among many possibilities. The optimality is measured in terms of the path's accumulated probability or likelihood. With the duration models, the conventional probabilistic path score can be modified to include the duration scores. Unlike the acoustic likelihood, duration scores are not computed on a short-time frame basis. There may be cases in which, when a path extension decision is made, some of the competing paths involve duration scores and others do not. Thus, the search is only sub-optimal. Examples of such sub-optimal methods can be found in [Power 1996].

In this work, we adopt the one-pass approach and aim for an optimal search. The conventional Viterbi search algorithm for connected-word recognition is modified to facilitate the incorporation of explicit duration models at both the state and the model levels. The effectiveness of different duration features is evaluated through recognition experiments.

In the next section, a brief introduction to the Cantonese dialect is given and the task of Cantonese connected-digit recognition is described. Baseline recognition performance is also presented. Statistical modeling of various types of duration features is described in Section 3. The ways of integrating duration models into the speech recognition processes are explained in Section 4. Experimental results are presented and discussed in Section 5. Conclusions are given in Section 6.

2. Cantonese Connected-Digit Recognition

2.1 About Cantonese

Cantonese is one of the major dialects of Chinese. It is the mother tongue of over 60 million people in Southern China and Hong Kong. Like Mandarin, Cantonese is a monosyllabic and tonal language. A Cantonese utterance is considered a string of monosyllabic sounds. Each Chinese character is pronounced as a single syllable that carries a specific tone. A character may have multiple pronunciations, and a syllable typically corresponds to a number of different characters. As shown in Table 1, each Cantonese digit is pronounced as a monosyllable sound.

Table 1. Phonetic transcriptions of the 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^{h}ɐt	cat1
8	pat	baat3
9	kɐu	gau2

2.2 Baseline System

Our baseline system for Cantonese connected-digit recognition was trained with the CUDIGIT database, which is part of a whole series of Cantonese spoken language corpora developed at the Chinese University of Hong Kong [Lee *et al.* 1998a]. CUDIGIT is a collection of Cantonese digit strings. The data collected were all read speech. Speakers were prompted with onr digit string at a time, with Chinese characters and Arabic digits displayed in parallel on a computer screen. The recordings were carried out in a closed quiet room using a high-quality microphone. The speech signal was sampled at 16 kHz. The database contains an exhaustive permutation of digit strings from one to four syllables long. There are also randomly generated strings that are of 7, 8, and 16 digits long. A total of 25 male and 25 female speakers were recorded. Each speaker spoke about 570 digit strings.

For the acoustic models of the baseline system, the training data included 11,387 utterances from 20 male speakers. In addition, 2,847 utterances from the other 5 male speakers in CUDIGIT were reserved as development data, which were used as the estimation of the weighting factor for the duration models (see Section 5.1).

The utterances for performance evaluation were from a different database, which was recently collected for speaker recognition research. It contains Cantonese digit strings recorded under the same acoustic conditions as CUDIGIT. About 900 utterances from 5 male speakers were used in this study. In terms of the total number of digit occurrences, the amount of the evaluation data is similar to the development data.

Feature extraction was done with a 20-msec Hamming window and 10-msec window overlapping. 32 nonlinearly spaced (Mel-scale) filter banks were used to cover the bandwidth of 8 kHz and the first 12 cepstral coefficients were computed. Each feature vector had 39 components, including the 12 Mel-Frequency Cepstral Coefficient (MFCC), log-energy, and their first and second order derivatives. Cepstral liftering was applied to the cepstral coefficients.

Each Cantonese digit was modeled by a whole-word HMM. The HMM had 6 left-to-right connected states. There was no state-skipping transition. Each state was associated with a mixture of 8 Gaussian distributions. Diagonal covariance matrices were assumed. There were also a six-state "silence" model and a one-state "sp" model for the non-speech signal. The baseline recognition performance is given in Table 2.

Table 2. Baseline performance for Cantonese connected-digit recognition

Digit accuracy	Deletions	Substitutions	Insertions
95.09%	82	116	418

2.3 Discussion

As shown in Table 2, insertions and deletions accounted for over 80% of the recognition errors. It must also be noted that 68.2% of the insertion and deletion errors were due to the digits "2" and "5" [Zhu 2005]. The phonetic compositions of Cantonese digits are generally very simple. This makes it difficult to attain high accuracy in the automatic recognition of Cantonese digit strings. For example, the digit "2" can be regarded as a single vowel segment. When this digit is repetitively spoken in a continuous utterance, the boundaries between them tend to be blurred because the signal's spectrum remains virtually unchanged. This will cause deletion and insertion errors in speech recognition. Moreover, "2" is phonetically very similar to the coda part of the digit "4". It is easily confused with this coda, and recognition errors will occur.

Figure 1 shows the spectrogram of an example utterance. It contains the digit string "22" during the period of 0.5 – 0.81 sec. There is no observable spectral discontinuity that signifies the boundary between the two digits. Similarly, in the example shown in Figure 2, the coda of digit "4" is likely to be recognized as an inserted "2".

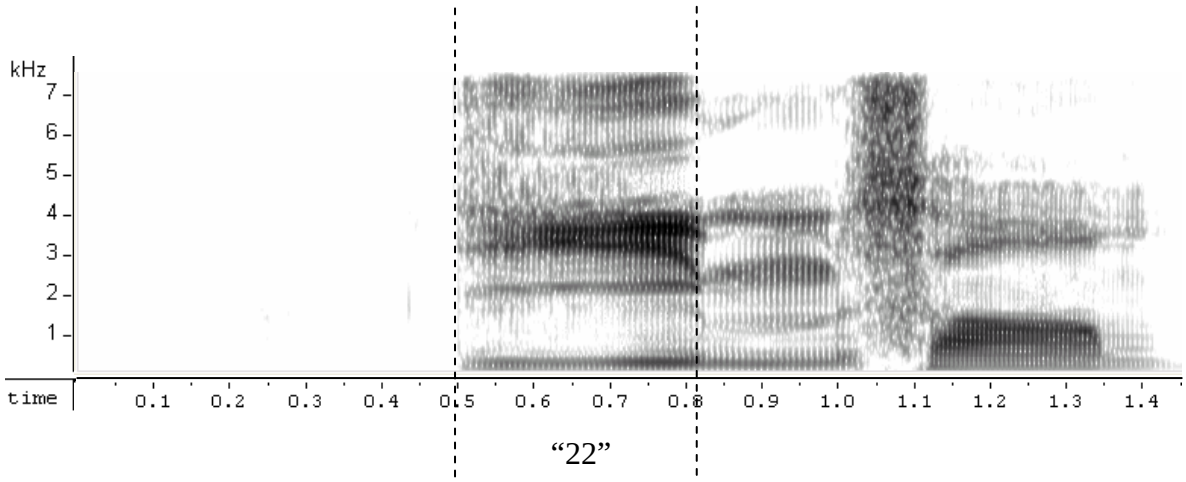

Figure 1. Spectrogram of an utterance that contains the digit string “22”

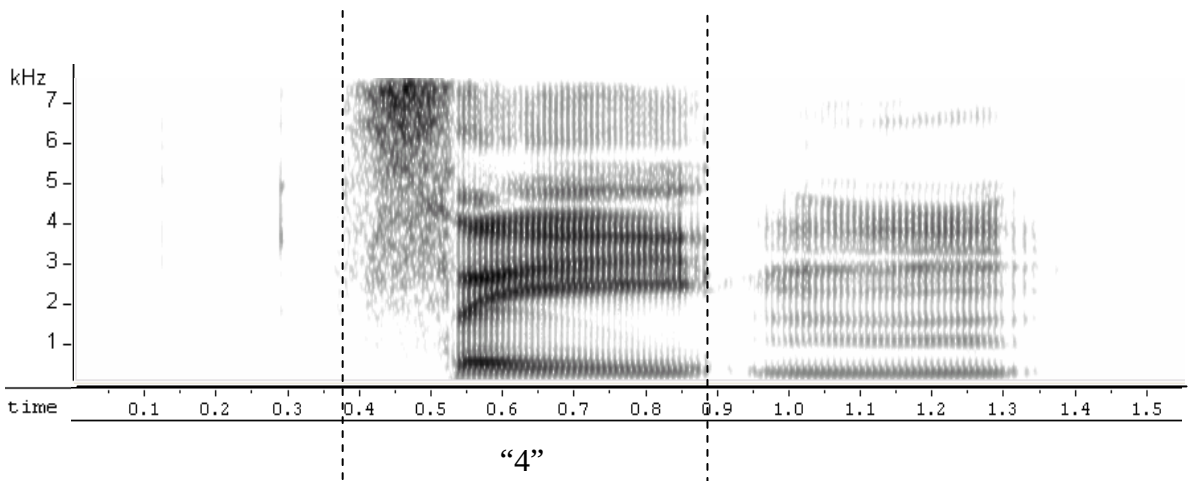

Figure 2. Spectrogram of an utterance that contains the digit “4”

Another problematic digit is “5”, which can be approximated as a single nasal segment. Like “2”, if the digit “5” is uttered repetitively in a continuous utterance, the spectral cues are not sufficient for detecting the digit boundaries. It is easily confused with the nasal codas of the digits “0” and “3.”

Although the duration of a digit is affected by many different factors, it by no means has an unlimited range of variation, especially in those applications where the speaking style and the speaking rate are relatively stable. In the cases in which repetitive “2” or “5” segments are merged or a single segment is split, the durations of the recognized digit segments usually deviate much from their nominal values. Similar argument can be made when the string “42”

is recognized as a single digit "4" or vice versa. Prior knowledge about digit durations would be helpful to correct such errors. In addition to the absolute duration, relative duration features, e.g. the ratio between the duration of certain state(s) and that of the whole digit, are also useful. These features reflect the regularity that governs possible internal adjustments among the sub-components of a digit segment. In the next section, the statistical modeling of both absolute and relative duration features is discussed.

3. Duration Modeling for Cantonese Digits

3.1 Duration Features

Duration can be measured and modeled at segments of various lengths. The measurements of duration information are referred to as duration features. In an HMM-based system, HMMs are used to model and segment speech signals. In our baseline system, each Cantonese digit was modeled by a whole-word HMM. Given a digit string, the durations of individual digits were given directly by the model-level segmentation. State durations were derived from the state-level time alignment.

Both state duration and model duration have been found to be useful for speech recognition, but their effectiveness varies across applications. It was reported that the use of the relative state duration (with respect to the model duration) leads to better recognition performance than the use of the absolute state and model durations [Power 1996].

In this study, both the absolute state duration and the absolute digit duration were investigated. As for the relative duration features, the relative state duration (with respect to the digit duration) and the so-called tail part ratio were used. The tail part ratio measures the relative duration of the tail part of a digit. The tail part is defined to cover the last two states of an HMM. The tail part ratio can be considered a variation of normalized state duration. From the baseline recognition results, it is observed that the tail part corresponds roughly to the last phonetic unit of the digit. As mentioned in Section 2.3, the two mono-phone digits, i.e., "2" and "5", are easily confused with the tail part of other digits. When the tail part is deleted or prolonged, the tail part ratio becomes unreasonable.

3.2 Statistical Modeling

In [Russell and Moore 1985], Poisson distribution was used to model state duration. While the model is simple to estimate (only one free parameter), it is not generally applicable because it demands that the variance be equal to the mean. It was found that Gaussian and Gamma distributions are more appropriate [Levinson 1986]. In [Gadde 2000], a mixture of Gaussian distributions was used to model multivariate duration features. In [Burshtein 1995], it was shown that Gamma distribution fits the empirical data better than the Gaussian distribution for

both state and model durations. In [Dong and Zhu 2002], it was also found that duration models using Gamma distributions are superior to other parametric distributions in terms of speech recognition accuracy.

Figure 3 shows the empirical distribution of the absolute duration of digit "0" as well as the corresponding Gamma fit. The empirical distribution was obtained through supervised segmentation (also known as forced alignment) of the training data in CUDIGIT. It can be seen that the Gamma distribution fits the empirical measurements quite well. This is also true for all other digits [Zhu 2005].

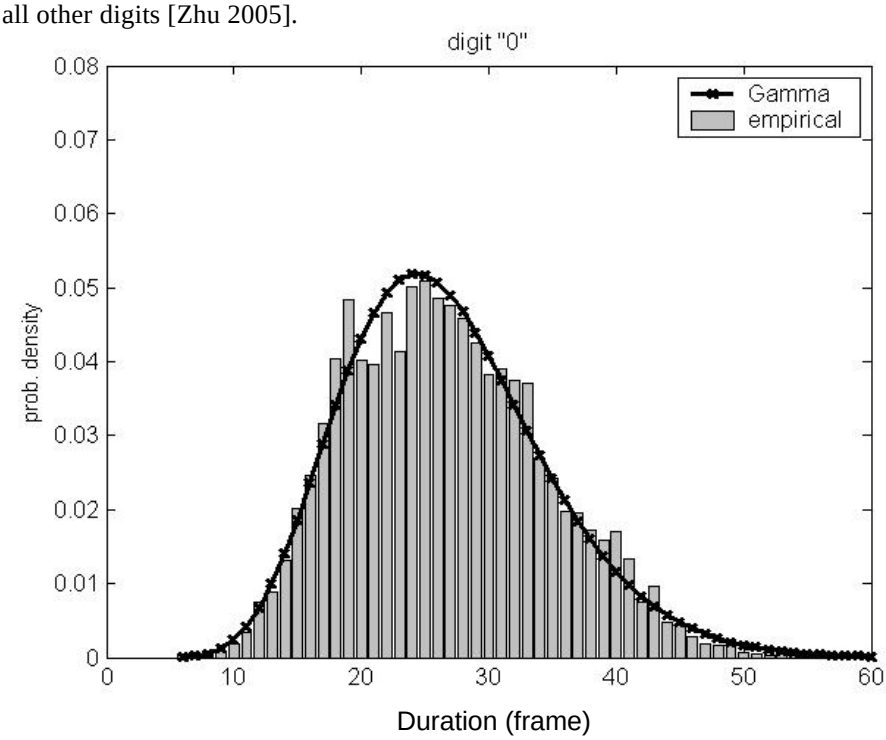

Figure 3. Distribution for the absolute digit duration for the digit "0"

For each HMM state, there is one distribution for absolute state duration and one distribution for relative state duration to be modeled. Thus, the total number of state duration distributions is 120. More than 70% of these empirical distributions can be approximated quite well as Gamma functions [Zhu 2005]. The distributions that do not fit well have complicated shapes, e.g., multi-modal. Similar observations are made concerning the modeling of relative state duration. For simplicity, uni-modal Gamma distribution is used in all state duration models.

As for the tail part ratios, the empirical distributions can all be nicely modeled with uni-modal Gamma functions. One of the examples is given in Figure 4.

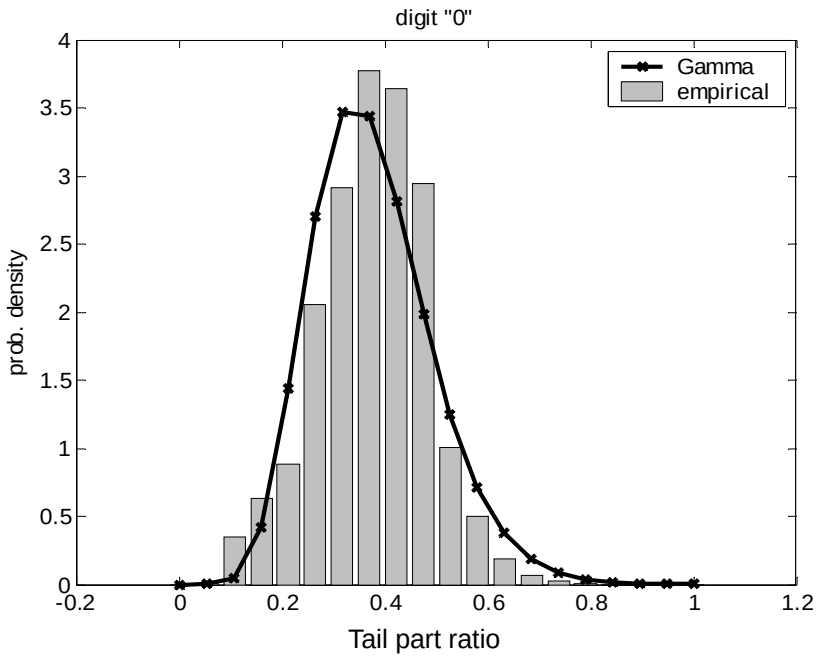

Figure 4. Distribution of the tail part ratio for the digit "0"

3.3 Training of Duration Models

In this study, Gamma distribution was used for the statistic modeling of duration features. Ther training of a duration model refers to the process of estimating the parameters of the Gamma distribution from segmented training utterances. Given a large amount of training utterances, manual segmentation at the word (digit) level is not realistic, let alone at the state level. Supervised automatic segmentation can be done with a set of acoustic models (HMMs) that are trained beforehand. This is referred to as the multi-pass training approach.

To obtain a truly optimal solution, the parameters of duration models must be estimated jointly with the HMM parameters, because they depend on each other [Russell and Moore 1985; Levinson 1986]. This one-pass approach is computationally expensive. Moreover, it is not applicable when sophisticated duration features, like relative state duration, are being modeled. Experimental results also showed that multi-pass training can be just as effective as one-pass training in terms of recognition performance [Rabiner 1989]. In this study, the duration models were trained through the multi-pass approach.

In summary, for our study regarding Cantonese connected-digit recognition, explicit duration models were established for the absolute digit duration, the absolute state duration, the relative state duration, and the tail part ratio. Each duration model was represented by a Gamma distribution, which was trained with CUDIGIT training data through the multi-pass

approach. In the subsequent discussion, the abbreviations in Table 3 are used to refer to the different duration features.

Table 3. Different duration features

AD	Absolute digit duration
AS	Absolute state duration
RS	Relative state duration
TR	Tail part ratio

4. Integrating Duration Models into Speech Recognition

As described earlier, the problem of connected-word recognition concerns the search for an optimal word string among many possibilities. The search space is formed by the HMM states, and a word string is in this way essentially a path connected by the states. The basic idea behind incorporating duration models into the search process is to make the duration probabilities contributive to the path probability. The challenge is to ensure that each path extension decision is optimal, considering that the duration probability is computed in a different time scale from the acoustic probability.

In the conventional Viterbi algorithm [Ney 1984], the problem of searching for an optimal complete path can be decomposed into many sub-problems at the frame level. The sub-problem at a particular frame *t* is to find the optimal partial path extended to each legitimate state. Let (t, v, j) denote the optimal partial path extended to state *j* of model *v* and at frame *t*. The accumulated path score is denoted by $L(t, v, j)$. The sub-problem at frame *t* can be solved given the solutions to the sub-problems at $t-1$, i.e., the immediately preceding frame. The path extension algorithm is explained as follows:

1) If the path is extended to the first state of an HMM, the predecessor can be the last state of any HMM or the current state itself. The path extension is done by,

$$L(t,v,1) = \max_{u} \left\{ L(t-1,u,N) \times a_{N,N+1}, L(t-1,v,1) \times a_{11} \right\} \times b_1(o_t), \tag{1}$$

where *N* is the number of states in the model and $a_{N,N+1}$ is the probability of exit from state *N*. Here we assume that all HMMs have the same number of states.

2) For a path extended to state *j* of a model, where $j \neq 1$, we have

$$L(t,v,j) = \max_{\substack{i=j \text{ or} \\ i=j-1}} \left\{ L(t-1,v,i) \times a_{ij} \right\} \times b_j(o_t). \tag{2}$$

That is, the predecessor can be either state *j* itself or state *j-1* of the same HMM, because we have assumed there is no state skipping.

The path extension is performed with a step size of one frame. To incorporate the duration model scores, the path extension needs to cover a longer time span. For state-level duration features, it should cover the time span of an HMM state. For word-level features, it should cover the span of a model.

4.1 Incorporation of State-Level Duration Model

For state-level duration features, the duration scores can only be computed if there is a state transition. In this case, the notion of path extension is defined differently. The step size of the path extension is a state instead of a frame. The path extension stretches from the beginning frame of one state to the beginning frame of another state. The state duration is a variable that affects the path extension decision.

Let (t, v, j) denote the optimal partial path that extends to state *j* of model *v* at frame *t*, and $L(t, v, j)$ be the corresponding accumulated path score. Accordingly, the path extension algorithms are modified as follows:

1) When the path gets to the first state of an HMM, its predecessor can be the last state of any other HMM. For each possible predecessor $(t-d, u, N)$, the duration score $D_{u,N}(d)$ is computed, where d is the duration of staying at state N. $D_{u,N}(d)$ is incorporated into the path extension decision as

$$L(t,v,1) = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-d,u,N) \times a_{N,N+1} \times \prod_{t-d<\tau<t} b_N(o_\tau) \times \left[D_{u,N}(d) \right]^w \right\} \times b_1(o_t), \qquad (3)$$

where d_{max} and d_{min} are the upper and lower bounds, respectively, of the state duration value, and *w* is an empirically determined weighting factor that controls the relative contribution of the duration scores.

2) For the path extension from state $j-1$ to state j of an HMM, where $j \neq 1$, we have

$$L(t,v,j) = \max_{d_{\min} \le d \le d_{\max}} \left\{ L(t-d,v,j-1) \times a_{j-1,j} \times \prod_{t-d<\tau<t} b_{j-1}(o_\tau) \times \left[D_{u,j-1}(d) \right]^w \right\} \times b_j(o_t), \qquad (4)$$

In this case, all competing path extensions are from state $j-1$ to state j. They differ from each other in terms of the time instant at which the extension occurs, which is specified by the value of d.

The above formulation is referred to as the 3-dimensional optimal decoder, because the token (t, v, j) has three elements. As seen in Eqs. (3) and (4), each possible path extension involves the computation of $\prod_{t-d<\tau<t} b_i(o_\tau)$. If the paths are evaluated individually, there are a lot of duplicated computations. To alleviate this problem, the search algorithm is

re-formulated. A new dimension "*d*" is introduced into the path token. The token (t, v, j, d) refers to a path that has stayed at state *j* in HMM *v* for *d* frames at frame *t*. Equations (3) and (4) can be written as

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

$$L(t,v,j,1) = \max_{d_{\min} \le d \le d_{\max}} \left\{ L(t-1,v,j-1,d) \times a_{j-1,j} \times \left[D_{u,j-1}(d) \right]^{w} \right\} \times b_j(o_t), \tag{6}$$

$$L(t,v,j,d) = L(t-1,v,j,d-1) \times b_j(o_t). \tag{7}$$

Such a 4-dimensional formulation is equivalent to the decoding framework in [Gu *et al.* 1991]. The computation cost of this decoder is d_{max} times that of the baseline decoder.

4.2 Incorporation of Word-Level Duration Models

To incorporate word-level duration scores, the step size of a path extension is defined to be a word (an HMM). A path extension is from the beginning frame of one word to that of another word. Let (t,v) denote the optimal partial path that extends to HMM *v* at frame *t*, and let $L(t,v)$ be its path score. The path extension decision is obtained as follows:

$$L(t,v) = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-d,u) \times warp(u,t-d,t-1) \times \left[D_u(d) \right]^{w} \right\}, \tag{8}$$

where $warp(u,t-d,t-1)$ is the probability that the sub-sequence of feature vectors from *t-d* to *t-1* is generated by HMM *u*, and d_{max} and d_{min} are the upper and lower bounds, respectively, of a word duration. $D_u(d)$ is the word-level duration score given by HMM *u*. It can be contributed by one or more duration features, including AD, RS, and TR as described in Section 3.1. For RS, it is assumed that the relative durations of individual states are independent of each other and the overall duration score is given by the multiplication of the probabilities obtained at all states.

Similar to the state-level case, the 4-dimensional formulation of the above algorithm is given as

$$L(t,v,1,1) = \max_{\substack{u \\ d_{\min} \le d \le d_{\max}}} \left\{ L(t-1,u,N,d) \times a_{N,N+1} \times \left[D_u(d) \right]^{w} \right\} \times b_1(o_t), \tag{9}$$

$$L(t,v,j,d) = \max_{\substack{i=j \text{ or} \\ i=j-1}} \left\{ L(t-1,v,i,d-1) \times a_{ij} \right\} \times b_j(o_t), \tag{10}$$

where (t,v,j,d) refers to a path that has stayed at state j of HMM v for d frames. The computation cost of this decoder is d_{max} times that of the baseline. Since word duration is much larger than state duration, the computation load of integrating word-level duration features is much heavier than that with state-level features. Such a 4-dimensional formulation is equivalent to the decoding framework in [Kwon and Un 1996].

5. Experimental Results and Discussion

5.1 Effectiveness of Different Duration Features

Experiments on Cantonese connected-digit recognition were carried out to evaluate the use of different duration features and their combinations. In all the experiments, the acoustic models were the same as those in the baseline system. The features and weights in the experiments are listed in Table. It is observed that the acoustic scores produced by the HMMs have a much wider dynamic range than the duration scores. 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. For each of them, the weighting factor *w* for the duration scores was empirically determined from the development data (see Section 2.2). Different values of weights were tested and the one with the best results are shown as in Table 4. The values of d_{max} are 15 and 80 for state-level and word-level models, respectively.

Table 4. List of duration features and the respective weights for duration scores

Duration features		*w*
State-level	AS	3
Word-level	AD	6
	RS	4
	TR	4
	AD+RS	6, 2
	AD+TR	6, 4

In addition, an experiment was performed using the word insertion penalty method, which is commonly used to reduce insertions [Huang *et al.* 2001]. The penalty value was also determined empirically from the development data.

The experimental results are given in Table 5. In all cases, the recognition accuracy is improved compared with the baseline system. The most significant improvement is 2.36% in terms of digit accuracy, which is attained by using the absolute state duration. The performance improvement results mostly from the reduction in insertion errors, and the

substitution errors also decreased. Meanwhile, more deletion errors are produced. The use of the word insertion penalty method can also improve recognition accuracy. However, it is not as effective as the explicit duration models.

Table 5. Recognition performance with different duration features

Method of duration control		Accuracy	Deletions	Substitutions	Insertions
Baseline		95.09%	82	116	418
State-level	AS	97.45%	105	88	127
Word-level	AD	96.70%	132	100	182
	RS	96.74%	98	108	203
	TR	96.11%	81	100	308
	AD+RS	97.22%	142	90	116
	AD+TR	97.24%	133	90	124
Insertion penalty		96.37%	124	117	215

The absolute state duration (AS) gives a better recognition performance than any of the word-level features. Since the incorporation of a state-level duration model requires much less computation, it is more preferable than the word-level duration models.

Among the three word-level features, the relative state duration (RS) is the most effective, while the tail part ratio (TR) gives little improvement. The combined use of word-level features, e.g., AD+RS and AD+TR, attains a similar performance to AS. This implies that RS and TR carry certain complementary information to AD.

5.2 The Effect of the Speaking Rate

It is obvious that duration features depend greatly on the speaking rate. We divided the evaluation utterances evenly into three categories based on their speaking rates. The speaking rate was defined based on normalized word duration as described in [Lee *et al.* 1998b]. For each category, a set of speaking-rate dependent duration models were built.

Table 6 shows the recognition performance for each speaking rate category. It is noted that the use of duration models is most effective for slow utterances, though improvement is observed in all categories.

Table 6. Recognition accuracy (%) for different speaking rates

Method of duration control		Fast	Medium	Slow
Baseline		96.19%	94.79%	93.57%
State-level	AS	96.77%	97.96%	97.40%
Word-level	AD+RS	96.45%	97.70%	97.40%
	AD+TR	96.42%	97.89%	96.92%

6. Conclusions

HMM does not give effective control over duration. For speech recognition tasks in which high-level linguistic constraints are not applicable, the duration of speech segments is a useful cue that supplements the conventional spectral features. In this work, we have shown how duration features can be used to improve the accuracy of Cantonese connected-digit recognition.

Among all of the duration features investigated, the absolute state duration gave the most noticeable performance improvement. A similar level of performance was also achieved with the combined use of absolute digit duration and relative state duration. With the use of duration information, insertion errors were much reduced, while deletion errors increased slightly. The reduction in insertion errors is particularly critical for Cantonese speech recognition because many of the short syllables in Cantonese are likely to be inserted if there is no duration control. Our experimental results also revealed that explicit duration models were more effective for slow speech than fast speech.

To incorporate duration models into the speech recognition process, the standard Viterbi search algorithm has to be modified. To ensure that the search is optimal, a larger step size for path extension is needed so as to accommodate the long time-span required for computing the duration scores. This leads to a significant increase in the computation load. To reduce the computation load, a sub-optimal search can be considered.

Acknowledgement

This research was partially supported by a research grant from the Hong Kong Research Grants Council (Ref: CUHK4206/01E).

References

Burshtein, D., "Robust parametric modeling of durations in Hidden Markov Models," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1995, pp 548-551.

Dong, R. and J. Zhu, "On use of duration modeling for continuous digits speech recognition," In *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 385-388.

Gadde, V. R. R., "Modeling word durations," In *Proceedings of the International Conference on Spoken Language Processing*, 2000, pp. 601-604.

Gu, H.Y., C.Y. Tseng and L.S. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with bounded State Duration," *IEEE Trans. Signal Processing*, 39(8), pp. 1743-1751, Aug. 1991.

Huang, X. D., A. Acero and H. W. Hon, Spoken language processing: A Guide to Theory, Algorithm and system development. Carnegie Mellon University, 2001.

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

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

Lee, T., W.K. Lo and P.C. Ching, "Development of Cantonese spoken language corpora for speech applications," In *Proceedings of the International Symposium on Chinese Spoken Language Processing*, 1998, pp. 102-107.

Lee, T., R. Carlson and B. Granström, "Context-dependent duration modeling for continuous speech recognition," In *Proceedings of the International Conference on Spoken Language Processing*, 1998, pp.2955-2958.

Levinson, S.E., "Continuously Variable Duration Hidden Markov Models for Speech Analysis," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1986, 2, pp. 1241-1244.

Ney, H., "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.

Power, K., "Duration modeling for improved connected digit recognition," In *Proceedings of the International Conference on Spoken Language Processing*, 1996, pp. 885-888.

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

Russell, M. and R. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 2376-2379.

Yang, C., "On the robustness of static and dynamic spectral information for speech recognition in noise," PhD dissertation, The Chinese University of Hong Kong, 2004.

Zhu, Y., "Using Duration Information in HMM-based Automatic Speech Recognition" MPhil Thesis, The Chinese University of Hong Kong, 2005.