Improved syllable-based continuous Mandarin speech recognition using intersyllable boundary models

Saga Chang and Sin-Horng Chen

Indexing term: Speech recognition

A novel approach to compensate for the intersyllable coarticulation effect on continuous Mandarin speech recognition by supplementing the syllable HMM models with intersyllable boundary models is proposed. Experimental results show that the syllable recognition errors of a speaker-dependent recognition task are reduced by 24%.

Introduction: One of the most important problems in continuous speech recognition is how to accurately model the acoustic variabilities caused by coarticulation to obtain high recognition accuracy. For English speech recognition, it has been found that using context-dependent phone models is an effective way to solve the problem [1]. For Mandarin speech recognition, a simpler approach can be adopted to take advantage of the simple phonetic structure of Mandarin syllables. Because there are only 411 phonetically distinguishable base syllables, it is wise to adopt the context-independent (CI) syllable-based speech recognition approach so that only intersyllable coarticulation is needed to be additionally considered. We therefore propose in this Letter a novel approach for improving the CI syllable-based continuous Mandarin speech recognition method by incorporating some intersyllable boundary models to compensate for the intersyllable coarticulation effect.

Intersyllable boundary models: For each pair of consecutive syllables, a fixed-length feature vector sequence located symmetrically across the intersyllable boundary is extracted for constructing the corresponding intersyllable boundary model. Since the acoustic characteristics in the transition region between two syllables are, in general, not stationary, a segment-based statistical model [2, 3] is adopted to deal with the nonstationarity of the feature vector sequence. Specifically, instead of modelling the feature vector of each frame independently as a Gaussian distribution, we represent, in an intersyllable boundary model, the feature vector sequence of the whole segment by a reference template with the residual vectors treated as independently and identically distributed white Gaussian noise with zero mean vector and diagonal covariance matrix **R**. In this model, feature vectors are regarded as a nonstationary Gaussian vector process with the mean vector sequence equal to a reference template and the common diagonal covariance matrix **R** estimated from residual vectors.

Two methods for generating reference templates of intersyllable boundary models are proposed. One is to employ matrix quantisation to find some representative reference templates for each syllable pair by taking each feature vector sequence as a matrix. The other uses contour quantisation [4] to first take the time sequence of each component of the feature vector sequence as a contour and represent it by the first few lower-order coefficients of all orthonormal polynomial expansion, and then collect coefficents of all components to form a new feature vector and apply vector quantisation to find some codewords. Reference templates for each syllable pair are generated from these codewords by orthonormal polynomial expansions. Here, the first three discrete Legendre polynomials were chosen as the basic functions [2] of the orthonormal polynomial expansion. These basis functions are normalised in length to [0,1] and expressed as

$$\Phi_0\left(\frac{t}{\tau}\right) = 1$$

$$\Phi_1\left(\frac{t}{\tau}\right) = \left[\frac{12(\tau-1)}{\tau+1}\right]^{1/2}\left[\left(\frac{t-1}{\tau-1}\right) - \frac{1}{2}\right]$$

$$\Phi_2\left(\frac{t}{\tau}\right) = \left[\frac{180(\tau-1)^2}{(\tau-2)(\tau+1)(\tau+2)}\right]^{1/2} \times \left[\left(\frac{t-1}{\tau-1}\right)^2 - \left(\frac{t-1}{\tau-1}\right) + \frac{\tau-2}{6(\tau-1)}\right] \quad (1)$$

for $t = 1, 2, \ldots \tau$, where τ is the length of the feature vector sequence.

If we construct one model for each syllable pair, there will be 411 × 411 intersyllable boundary models. However, as the training data are usually very limited in practical applications, the number of intersyllable boundary models should somehow be greatly reduced. Here we solve the problem by simply assuming that the acoustic variability of an intersyllable boundary segment results solely from the two phones located nearest to the boundary. According to the phonetic structure of Mandarin syllables, there are in total 11 types of ending phones and 27 types of starting phones. Hence, only 297 intersyllable boundary models are needed. Furthermore, silence must be considered separately. By treating silence as an ending phone for the following syllable and a starting phone for the preceding syllable, we construct 38 additional models.

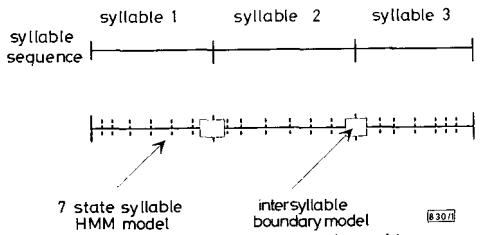

Fig. 1 *Schematic diagram for representing speech signal by sequence of syllable HMM models overlaid with sequence of intersyllable boundary models*

Proposed speech recognition approach: By supplementing these 335 intersyllable boundary models to a set of 411 CI syllable HMM models, we can represent a speech signal by a sequence of syllable HMM models overlaid with a sequence of intersyllable boundary models. Fig. 1 shows a schematic diagram of the representation. Using this representation of the speech signal, the task of continuous Mandarin speech recognition is then to find the best syllable string for a testing utterance based on the given sets of CI syllable HMM models and intersyllable boundary models.

In the training phase, a set of 411 initial CI syllable HMM models is first generated by the segmental k-means training algorithm. We then use the results of syllable segmentations of all training utterances to extract intersyllable boundary segments for training a set of 335 initial intersyllable boundary models. An iterative training procedure is then applied to alternatively segmenting all training utterances into syllable sequences by using a modified Viterbi algorithm and updating both sets of CI syllable HMM models and intersyllable boundary models. The only modification of the Viterbi algorithm is to add an additional score at the instant of the decision to make a syllable transition. This is the score of matching the input vector sequence across the boundary with an intersyllable boundary model of the syllable pair involving the syllable transition. A normalisation of the local score is then applied in order not to favour the path with syllable transitions. Repeatedly applying the procedure of segmentation and model updating, we can obtain two well-trained sets of 411 CI syllable HMM models and 335 intersyllable boundary models.

In the testing phase, a one-stage search algorithm is applied to find the best sequences of CI syllable HMM models and the overlaid intersyllable boundary models for the input utterance. As with the training algorithm, the one-stage search algorithm is also modified by considering the effect of the additional intersyllable boundary models. The syllable sequence associated with the best sequence of CI syllable HMM models is taken as the recognition output.

Table 1: Recognition result using CI syllable HMM models

acc	corr	sub	del	ins
%	%	%	%	%
87.1	80.6	18.9	0.5	2.5

Experiments: The effectiveness of the proposed method was tested with simulations using a continuous Mandarin speech database uttered by a single male speaker. The database contains 35231 syl-

lables in total including 28197 training syllables and 7034 testing syllables. All speech signals were sampled at a rate of 10kHz and pre-emphasised with digital filter with function $1 - 0.95z^{-1}$. The signals were then analysed for each Hamming-windowed frame of 20ms with 10 msframe shift. The recognition features consist of 12 mel cepstrums, 12 delta cepstrums, and the delta energy. First, the conventional HMM method based on 411 7state CI syllable HMM models and a 1 state silence model was tested. The number of Gaussian mixtures in each state of a syllable HMM model varies from one to five depending on the number of training data. Table 1 shows the recognition results. A syllable accuracy rate of 77.1% was achieved. Then, the proposed recognition method was tested.

Table 2: Recognition results using CI syllable HMM models and intersyllable boundary models trained by matrix quantisation

Length of boundary segment	acc	corr	sub	del	ins
	%	%	%	%	%
4	83.1	83.9	15.1	1.0	0.8
6	83.4	84.7	14.5	0.8	1.3

Table 3: Recognition results using CI syllable HMM models and intersyllable boundary models trained by contour quantisation

Length of boundary segment	acc	corr	sub	del	ins
	%	%	%	%	%
4	82.9	83.8	15.3	0.9	0.9
6	83.2	84.6	14.5	0.9	1.4

A set of 283 intersyllable boundary models for all syllable pairs existing in the database was then generated and supplemented to the set of 411 CI syllable HMM models to assist the recognition. The number of reference templates in each intersyllable boundary model was varied from one to five depending also on the number of training data. The recognition results for the case of using matrix quantisation to generate reference templates of the intersyllable boundary model are shown in Table 2. Syllable accuracy rates of 83.1 and 83.4% were achieved as the length of intersyllable boundary segments was set to four and six, respectively. Table 3 shows the recognition results when using contour quantisation to generate reference templates of the intersyllable boundary models. Syllable accuracy rates of 82.9 and 83.2% were obtained when the length of intersyllable boundary segments was set to four and six, respectively. By comparing Tables 2 and 3, we find that the performances for the two cases of generating reference templates are comparable to each other. However, comparing the results with those shown in Table 1, we find that the syllable accuracy rate for the proposed method was raised by ~5%, which corresponds to a reduction in recognition errors of 24%. Therefore, we conclude that supplementation with intersyllable boundary models is a valid way to improve the conventional HMM method which uses only CI syllable HMM models.

Conclusion: A novel approach for improving the conventional syllable-based HMM speech recognition method for continuous Mandarin speech has been studied. Experimental results show that ~24% of syllable recognition errors were corrected in a speaker-dependent speech recognition task.

4 April 1995
Electronics Letters Online No: 19950629

Saga Chang and Sin-Horng Chen (*Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*)

References

1 LEE, C.H., RABINER, L.R., PIERACCINI, R., and WILPON, J.G.: 'Acoustic modeling for large vocabulary speech recognition', *Comput. Speech Lang.*, 1990, **4**, pp. 127–165
2 CHANG, S., and CHEN, S.H.: 'A modified hidden semi-Markov model for multi-speaker Mandarin syllable recognition', *J. Chin. Inst. Electrical Engineering*, 1994, **1**, (2), pp. 95–104
3 DENG, L.: 'A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal', *Signal Process.*, 1992, **27**, pp. 65–78
4 CHANG, S., CHEN, S.H., CHUNG, C.J., and HONG, V.: 'A low data rate LPC Vocoder using contour quantization'. Proc. EUSIPCO-92: Sixth European Signal Processing Conf., 1992, pp. 459–462

Improving the input-queueing switch under bursty traffic

Jiunn-Jian Li

Indexing terms: Queueing theory, Switching theory

A new window scheme for improving the performance of nonblocking switches with input queueing under bursty traffic is proposed. The maximum throughput for the proposed scheme is measured and the scheme is shown to alleviate the effect of burst traffic on input queueing.

Introduction: The maximum throughput of a nonblocking switch with FIFO buffers on the input ports is limited by the head of line (HOL) blocking and approaches ~0.60 for large switch size N with random traffic. The window policy [1] can be used to reduce the HOL blocking by allowing the non-HOL cells to contend for the outputs. In the beginning of a time slot, the input port not selected to transmit its HOL cell has its second cell contending for the remaining idle outputs. The contending process repeats up to w times until a cell wins the contention, where w is referred to as the window size. In the window policy, each input buffer must be a first-in random-out (FIRO) queue.

A window size $w = 1$ corresponds to input queueing with FIFO buffers. As w increases, the throughput performance improves on a random traffic assumption. Note that the improvement diminishes quickly under bursty traffic. This is because the bursty traffic conditions make it likely that the first w cells in each input queue have the same destination.

In this Letter, an approach to alleviate the influence of bursty traffic on the window policy is proposed. Each input port employs an FIRO queue and b registers, called burst head pointers (BHPs). If there are k bursts in an input queue, BHP i ($1 \le i \le k$) points to the leading cell of the ith burst and the remaining $(b - k)$ BHPs remain idle. The cell pointed to by the ith BHP is labelled [BHP$_i$]. Fig. 1 illustrates the proposed scheme, where $b = 3$ and BHPs point to the leading cells of the first three bursts (whose destinations are y, z and x, respectively).

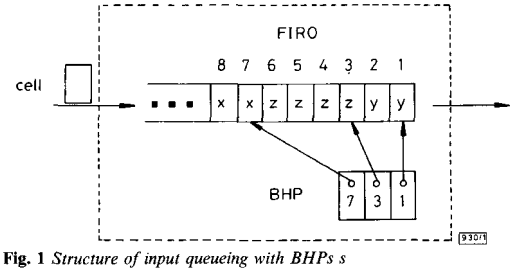

Fig. 1 *Structure of input queueing with BHPs s*

Operation: The proposed scheme is very similar in operation to the window policy [1]. At the beginning of each time slot, if the cell [BHP$_i$], $1 \le i \le b$, is blocked, a chance is given to the cell [BHP$_{i+1}$], until either a cell is selected or the cell [BHP$_b$] is reached, whichever comes first. Thus, the b cells pointed to by the BHPs form a 'window' in an input queue.

Maintaining the BHPs is easy. The contents of BHPs do not change except in the cases described below. Suppose there are k bursts in the target input queue.