An Initial Study on A Segmental Probability Model Approach to Large-Vocabulary Continuous Mandarin Speech Recognition

Jia-lin Shen, Hsin-min Wang, Bo-ren Bai and Lin-shan Lee
Department of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

ABSTRACT

This paper presents an initial study to perform large-vocabulary continuous Mandarin speech recognition based on a Segmental Probability Model(SPM) approach. SPM was first proposed for recognition of isolated Mandarin syllables, in which every syllable must be equally segmented before recognition. Therefore, A concatenated syllable matching algorithm in place of the conventional Viterbi search algorithm is therefore introduced to perform the recognition process based on SPM. In addition, a training procedure is also proposed to reestimate the SPM parameters for continuous speech. Preliminary simulation results indicate that significant improvements in both recognition rates and speed can be achieved as compared to the conventional HMM-based Viterbi search approaches.

1 INTRODUCTION

Mandarin Chinese is a tonal language. Each Mandarin syllable is assigned a tone, and there exists 4 lexical tones and 1 neutral tone. When the differences in tones are disregarded, the total of 1333 different Mandarin syllables can be reduced to only 408 base syllables. In this research, the recognition of the total of 408 base syllables disregarding the tones for large-vocabulary continuous Mandarin speech is considered based on a Segmental Probability Model(SPM) approach. The Segmental Probability Model(SPM) was first proposed for recognition of isolated Mandarin syllables considering the monosyllablic structure of Chinese language[1]. This model is very similar to continuous hidden markov model(CHMM), except that the state transition probabilities are deleted and the N states equally segment the syllable. In other words, the stochastic state transition behavior in CHMM is replaced by a deterministic process, while the stochastic observation behavior remains unchanged, represented by Gaussian mixtures. This model was found to be very suitable for Mandarin syllables due to their relatively simple phonetic structures, with improved recognition rates achievable at much higher speed[1]. This is why we try to extend the application of this model to continuous Mandarin speech recognition. The Viterbi search algorithm is applied to the HMM-based continuous speech recognition.[2][3][4] Howerer, it is impossible to directly implement any Viterbi search algorithm in a SPM-based continuous speech recognition task, because using SPM every syllable in an utterance must be equally segmented and thus the beginning and ending point of a syllable must be known. We therefore introduce a concatenated syllable matching algorithm based on dynamic programming concept[5] such that the recognition process can be performed syllable by syllable based on SPM, instead of frame by frame as in the HMM-based approaches. Furthermore, a training procedure is also introduced to segment an utterance into syllables and reestimate the SPM parameters by interpolation.

This paper is organized as follows. In section 2, we introduce the concatenated syllable matching algorithm. In section 3, the training procedure is discussed based on the concatenated syllable matching concept. The experimental results are presented in section 4. Section 5 finally gives the concluding remarks.

0-7803-1775-0/94 $3.00

2 THE CONCATENATED SYLLABLE MATCHING ALGORITHM

The concatenated syllable matching algorithm for continuous speech recognition using SPM is first presented below. For a given test utterance, all possible syllable beginning frames can be first obtained by picking up all the dips in the energy contour, such as x, y, z in Fig.1. The possible ending frames such as y-1 and z-1 corresponding to each beginning frame such as x in Fig.1 can then be found using estimated minimum and maximum duration of a syllable D_{min} and D_{max} as in Fig.1.

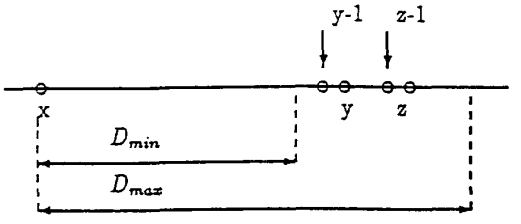

Figure 1: A part of an example utterance. x,y,and z are possible beginning points, where y-1 and z-1 are ending points corresponding to x.

With beginning and ending points of a syllable estimated as above, the accumulated score at an ending point y-1 is then determined by the dynamic programming approach:

$$T[y-1] = T[x-1] + \max_{1 \le i \le 408} [S_i(x, y-1)]$$

where T[u] is the accumulated score at a point u, and S_i(u,v) is the score when the SPM for the syllable i was matched with utterance section (u,v), and the total number of different Mandarin syllables is 408 when the tone is disregarded. At the end of the utterance, the optimal syllable sequence can then be easily obtained by backtracking the entire utterance. In this way, the SPM can be easily applied and the recognition can be performed syllable by syllable. This is similar to a previously reported method[5][6] except that the recognition process starts from the possible beginning points instead of ending points. In this way, it is more easier to implement on a real-time system because we can start the recognition process right at the beginning of an utterance. In addition, when a Viterbi search operation is performed, the computation for all possible ending points for a given possible beginning point need not be repeated because their starting point is exactly the same.

In order to reduce the computation due to the overlap of some possible syllable sections, a table of size D_{max} is dynamically used to store the likelihood scores of SPM. In addition, for those possible syllable sections with the same starting point, the likelihood scores of the same segment of SPM would not be calculated repeatedly. Therefore the recognition speed can be improved significantly.

3 THE TRAINING PROCEDURE

In the training process, a similar iterative reestimation procedure based on the above mentioned concatenated syllable matching concept was developed. For the j-th training sentence with n syllables $(l_1, l_2, ..., l_n)_j$, where l_k is the k-th syllable in the sentence, we can first find the average syllable duration for this sentence and then use some estimated parameters to determine the minimum and maximum syllable duration, $D_{min,j}$ and $D_{max,j}$, of this sentence. With the help of the dips in short-time energy contour, the possible beginning and ending points can be similarly found just as above in Fig.1. For an utterance section (x,y-1), the possible syllable occupying this section is estimated as l_k, where the range of possible values of k is determined by x, $D_{min,j}$ and $D_{max,j}$. The rest of the training process is very similar to that in recognition discussed above, except that only the SPM for the few possible syllables l_k in the sentence estimated in the above need to be matched. The dynamic programming approach just as above can be used:

$$T[y-1, l_k] = T[x-1, l_{k-1}] + S_{l_k}(x, y-1)$$

where T[u,l_k] is the accumulated score at a point u when the last syllable matched is l_k. When the end of the sentence is reached, backtracking through the utterance gives the optimal syllable segmentation, and the optimally segmented syllables for all the training utterances are used to re-train the SPM for all the syllables.

The parameters of each segment in SPM are estimated by a multivariate Gaussian distribution. In this study, only diagonal covariance matrices are considered for the multivariate Gaussian distributions. Therefore the interpolation for different statistical parameters is relatively easy:

$$\mu_{jnew} = w_1\mu_{ji} + w_2\mu_{jc}$$

$$\sigma^2_{jnew} = w_1(\mu^2_{ji} + \sigma^2_{ji}) + w_2(\mu^2_{jc} + \sigma^2_{jc}) - \mu^2_{jnew}$$

where μ_{ji} and σ^2_{ji} are the mean and variance of the j-th mixture component of the isolated model for a certain syllable, while μ_{jc} and σ^2_{jc} are the mean and variance of training samples observed in continuous speech for that mixture j. w1 and w2 are the interpolation weights for the prior mean and the observed sample mean respectively. The obtained models defined by the new parameters $(\mu_{jnew}, \sigma^2_{jnew})$ are then taken as new initial models and the above training procedure can be performed iteratively.

4 EXPERIMENTS AND DISCUSSION

In the preliminary simulation study for speaker dependent task, a speech database produced by two male speakers is used. Each speaker produced one isolated utterance for each of the 1333 syllables, a set of 112 phonetically balanced continuous sentences including a total of 891 syllables covering all the 408 base syllables, and one continuous utterance each for a set of 200 sentences taken from primary school Chinese textbook with a total of 1289 syllables. The average speaking rates of isolated and continuous utterances are 0.5 and 0.3 *sec/syllable* respectively. Cepstral coefficients of order 14 and the corresponding 14 delta cepstral coefficients are derived from the LPC coefficients. The 1333 isolated syllables are used in training initial models, the 112 phonetically balanced continuous sentences are used in re-estimating the continuous model parameters, and the rest of 200 continuous textbook sentences are used in testing. The recognition rates in the following experiments are evaluated as the percentages of correctly recognized syllables minus insertion rates. The average error rates when the continuous training utterances were used in testing after the iterative training procedure are listed in Figure 2. It is noted that the training procedure improved the model performance significantly and the error rate is almost unchanged after 2 iterations. Table 2 lists the average recognition rates of the test utterances after the iterative training procedure. The experimental results for HMM-based Viterbi search are also listed in Table 2 for comparison where syllable is also used as the recognition unit. In table 2, all the HMM models are trained using the isolated utterance as initial models and reestimated by the 112 phonetically balanced continuous utterance using segmental k-means algorithm with 10 iterations. It can be found that for HMM-based approaches DHMM has the fastest speed even in full search, while the highest achievable rate is 63% for CHMM with 7 states and 3 mixtures per state. But when SPM-based approach is used, more than 73% recognition rate can be easily obtained with smaller number of states(segments) using less than a quarter of time.

After careful examination on the test results, it was found that one major reason that SPM-based approach can outperform HMM-based approach is the significant reduction of the insertion rates, which are also listed in Table 2. Recall that SPM approach replaces the stochastic state transition process by deterministic segment transition process, uses extra knowledge such as energy contour dips to detect syllable end-points, and perform the recognition operation syllable by syllable instead of frame by frame. All these characteristics of SPM approach apparently provides extra advantages in recognizing an unknown utterance with unknown number of syllables. Note that the results here are obtained from acoustic information only. With the help of knowledge in tone, lexicon, context and assisted by language modeling, it is expected that much higher recognition rates can be achieved.

iteration number	recognition rate(%)
0	47.61
1	73.06
2	73.40
3	73.51

Table 1: the recognition results after iterative training procedure

5 CONCLUSION

In this paper, we applied Segmental Probability Model(SPM) to large-vocabulary continuous Mandarin speech recognition and achieved very good performance. Because the conventional Viterbi search algorithm can not be applied in a SPM-based continuous speech recognition task due to the unknown beginnig and ending points of the syllable in an utterance, we first introduced the concatenated syllable matching algorithm for continuous speech recognition using SPM. Then a training procedure is proposed to reestimate the SPM parameters. The experimental results show that SPM is very useful not only in isolated syllable recognition, but also for continuous speech in Mandarin Chinese.

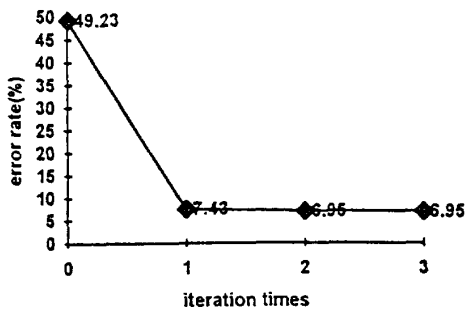

Figure 2: The error rates of continuous training utterance after iterative training procedure

References

[1] Lin-Shan Lee,*et al*, "Golden Mandarin(II)-An improved Single-chip Real-time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary" *ICASSP 1993, pp.503-506.*

[2] Picone,J. , "Continuous Speech recognition Using Hidden Markov Model" *IEEE ASSP Magazine, pp.26-41, July 1990.*

[3] Lee,K., Hon,H.and Reddy,R. , "An Overview of the SPHINX Speech Recognition System" *IEEE Trans. on ASSP, Jan. pp.35-45.*

[4] L.R.Rabiner, B.H.Juang, S.E.Levinson and M.M.Sondhi, " Recognition of isolated digits using Hidden Markov Models with continuous mixture densities" *AT&T Technical Journal,vol 64,No 6, pp.1211-1234, 1985.*

[5] T.Ukita, E.Saito, T.Nitta, and S.Watanabe, "A speaker-independent connected digit recognition system concatenating statistically discriminated words" *IEEE Trans. on Signal Processing vol.40 No.10 1992.*

[6] Mari Ostendorf, Salim Roukos, "A stochastic segment model for phoneme-based continuous speech recognition" *IEEE Trans. on ASSP vol.37 No.12 1989.*

models	schemes	recognition rate(%)	insertion rate(%)	speed (sec/syllable)
DHMM	5 states,codebook size=128 full search	59.25	4.5	22.63
CHMM	4 states,2 mixtures beam width = 400	55.18	9.3	29.24
CHMM	5 states,3 mixtures beam width = 500	61.29	6.5	51.26
CHMM	7 states,3 mixtures beam width = 500	62.82	4.5	53.27
SPM	cep:4 segments,2 mixtures dcep:1 segments,6 mixtures	73.51	0.78	12.31

Table 2: continuous speech recognition rate of base syllables for HMM and SPM (all experiments are conducted in a Sun Sparc 2 station)