J Inf Process Syst, Vol.12, No.3, pp.410~421, September 2016
http://dx.doi.org/10.3745/JIPS.03.0052

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Landmark-Guided Segmental Speech Decoding for Continuous Mandarin Speech Recognition

Hao Chao* and Cheng Song*

Abstract

In this paper, we propose a framework that attempts to incorporate landmarks into a segment-based Mandarin speech recognition system. In this method, landmarks provide boundary information and phonetic class information, and the information is used to direct the decoding process. To prove the validity of this method, two kinds of landmarks that can be reliably detected are used to direct the decoding process of a segment model (SM) based Mandarin LVCSR (large vocabulary continuous speech recognition) system. The results of our experiment show that about 30% decoding time can be saved without an obvious decrease in recognition accuracy. Thus, the potential of our method is demonstrated.

Keywords

Decoding, Landmark, Mandarin, Speech Recognition, Segment Model

1. Introduction

Currently, the hidden Markov model (HMM) is widely used in speech recognition systems and can achieve good performance. However, the assumptions of HMM, such as constant statistics within an HMM state and observations which are conditional independent, are not realistic for sequences of the speech spectra. To release from the limitations of HMM, some alternative models have been proposed. One of the models is the segment model (SM) [1].

Although SM can give competitive recognition results, the complexity of computing the joint likelihood and searching for the optimal segment boundaries is high. One of the reasons is that attention is distributed uniformly throughout the speech signal in the decoding process and thus, the blind search is used. So applications of SM are limited to small vocabulary tasks or re-scoring in large vocabulary continuous speech recognition (LVCSR). In previous works on SM [2,3], several methods that can speed up the likelihood computation and segment-based recognition have been proposed and achieve good performance. Nonetheless, the blind search is still used in these methods.

On the other hand, landmark-based speech recognition systems, which rely mainly on precise phonetic knowledge and that exploit distinctive features, have been recently given increasing importance [4-7]. From the perspective of speech perception, regions near landmarks contain more information than other regions in the speech signal. In this paper we conduct an attempt to speed up SM using the

Manuscript received November 4, 2014; first revision May 12, 2015; accepted June 17, 2015.

Corresponding Author: Hao Chao (chaohao@hpu.edu.cn)
* College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China ({chaohao, songcheng}@hpu.edu.cn)

information provided by landmarks. In our method, more attention is focused on the regions around landmarks other than all regions in speech signal. We employed landmarks to guide the search for the best path during decoding in an SM-based system. It should be noted that landmarks are only used to reduce the search space during decoding in our method, which is different from the landmark-based speech recognition systems using phonetic landmarks as a feature describing the signal.

The rest of this paper is organized as follows: in the next section we describe the details of the landmarks. In Section 3, the introduction of the stochastic segment model (SSM) is given. Section 4 introduces how the landmark information is integrated into the SM system. In Section 5, the results of our experiments are presented and discussed. The conclusions are described in Section 6.

2. Landmarks and Information Provided by Landmarks

Landmarks are defined as the time points of acoustic events that are consistently correlated to major articulatory movements [4,8]. According to Park [8], there are five different types of landmarks: the glottis landmark (g-landmark), sonorant landmark (s-landmark), burst landmark (b-landmark), vowel landmark (V-landmark), and glide landmark (G-landmark). Landmarks can identify articulator-free features of adjacent segments. For Mandarin, we sought to obtain the phonetic class information and phonetic boundary information according to landmarks.

2.1 Categories of Mandarin Initials and Finals

Mandarin is a syllabic centric language. Each Chinese character can be phonetically represented by a syllable and most Chinese syllables have the initial-final structure. Initial includes consonants and glides (semi-vowels). According to pronouncing methods, the consonants may be divided into five kinds (stops, fricatives, affricates, nasals, and laterals), and according to the vibration of vocal cords, the consonants may be divided into voiced consonants and unvoiced consonants (Table 1). Finals have three kinds of structures (Table 2): vowel, complex vowels (CV) and vowel or complex vowels ending with a nasal “n” and “ng” (CVN). Complex vowels are composed of two or more vowels. When the complex vowel sounds occur, the tongue position and lip shape change continuously and smoothly according to the order of the vowels. In the complex vowels, the tongue position and lip shape for two adjacent vowels interact with each other and eventually form a fixed pronunciation similar to a simple vowel [9].

Table 1. Categories of Mandarin initials

	Unvoiced	Voiced
Consonant		
Stop	b, p, d, t, g, k	
Fricative	f, s, sh, x, h	r
Affricate	z, zh, j, c, ch, q	
Nasal		m, n
Lateral		l
Glide		w, y

Table 2. Categories of Mandarin finals

Vowel	a, o, e, i, u, ü, er
Complex vowels	ai, ao, ei, ia, iao, ie, iu, ua, ui, ue, uo, uai, ou
Vowel or complex vowels ended with nasal	an, ian, uan, üan, en, in, uen, ün, ang, iang, uang, eng eng, ing, ueng, ong, iong

2.2 Information Provided by Landmarks

Broad phonetic classes and phonetic boundary information can be determined from individual landmark types (Fig. 1). A g-landmark marks a time when vocal folds start vibrating freely or when the vibration stops vibrating freely.

A +g landmark corresponds to the onset of vocal fold vibration and a –g landmark corresponds to the offset of the vocal fold vibration. According to the definition of g-landmark and the syllabic structure of Mandarin, the phonetic class information of segments adjacent to g-landmarks can be obtained. A +g landmark is located at the boundary between an unvoiced consonant initial and a final ("z" and "eng1" in Fig. 2), or at the boundary between a silent interval and a glide initial or a voiced consonant initial. A –g landmark is located at the boundary between a final and a silent interval, or at the boundary between a final and an unvoiced consonant initial ("eng1" and "j" in Fig. 2).

An s-landmark marks a time when the velopharyngeal port opens or closes during a sonorant sound. A +s landmark corresponds to the closing of the velopharyngeal port, and is located phonetically at the boundary between a voiced consonant initial and a final ("l" and "e0" in Fig. 2). A –s landmark corresponds to the opening of the velopharyngeal port. It is located at the boundary between a vowel final or a CV final and a voiced consonant initial ("ia1" and "l" in Fig. 2), or it is located inside a CVN final ("eng1" in Fig. 2) because the head of a CVN is a vowel and the tail of CVN is nasal.

A b-landmark marks a time when a stop or an affricate bursts (+b landmark), or a time when a turbulence noise ends in regions without glottal vibrations (–b landmark). A +b landmark is set at the boundary between a silent interval and a stop burst, or at the boundary between a silent interval and an affricate burst ("z" in Fig. 2). Phonetically, the b-landmark is located inside a stop or affricate. In Mandarin, only the +b landmarks exist. The reason is that stops, fricatives, and affricates are initials and all of them must be followed closely by finals other than a silent interval. Consequently, –b landmarks, which correspond to the end of turbulence noises, will not appear.

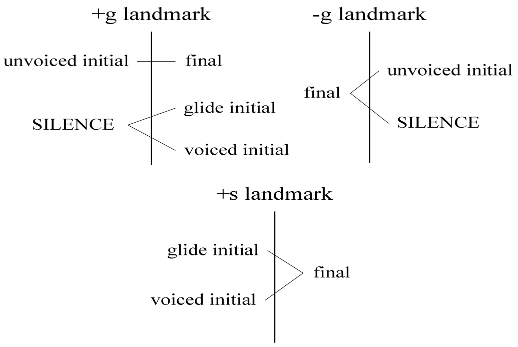


Fig. 1. Broad phonetic classes which can be identified from individual landmark types.

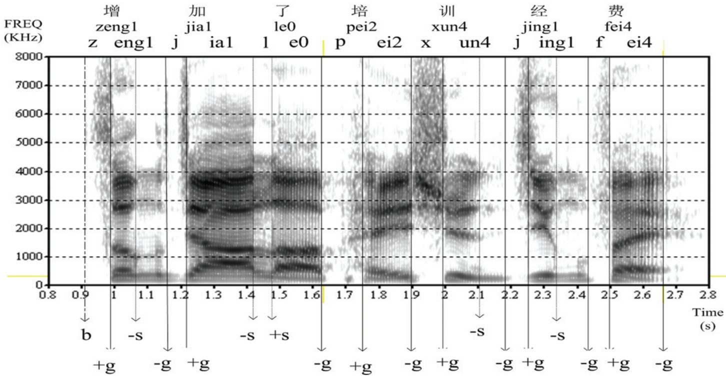


Fig. 2. Consonantal landmarks.

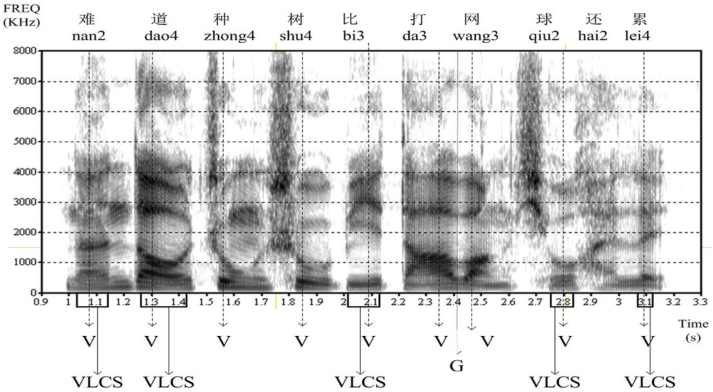


Fig. 3. Vowel and glide landmarks.

V-landmarks or G-landmarks mark the time when the effect of the vowel or glide on the acoustic signal is the most dominant. Thus V-landmarks are located within finals and G-landmarks are located within glides (Fig. 3).

A vowel is produced with a maximum opening in the vocal tract, and a V-landmark can be located in the vowel where there is maximum amplitude in the low frequency range. In the pronouncing process of vowels, the amplitude in the low frequency range changes slowly and smoothly, so the V-landmark-centered speech segment (VLCS) can be obtained according to the amplitude change. For a V-landmark, its corresponding VLCS (Fig. 3) is a speech segment that contains this V-landmark, and this speech segment is within the final that the V-landmark belongs to. Meanwhile, the amplitude in the low frequency range changes slowly and smoothly in this speech segment.

3. Landmark Detected

In this paper, the +g landmark, –g landmark, +s landmark, and VLCS are detected to guide the decoding process of SSM. We adopted S. Liu's basic framework of general processing for landmark detection in [4] to find the candidates of landmarks. Then, the final landmarks were picked out at a selecting stage according to acoustic rules, where the energy criteria and duration criteria are applied.

3.1 General Processing

In general processing, a spectrogram is computed with a 6-ms Hanning window every 1 ms, and two frequency bands were defined, as shown in Table 3. Then, the smoothing energy and energy ROR (Rate of Rise) data for each band were calculated in a coarse-processing (CP mean) pass and a fine-processing (FP) pass. In the next step, the peaks for each CP and FP ROR waveform with a local maximum or minimum value were found. The peaks had the same polarity with its ROR value. In the end, the final peaks were picked out from FP peaks using CP peaks as guides.

Table 3. Two energy bands and their frequency range

Band	Frequency (kHz)
1	0.0–0.8
2	0.8–2.0

A new peak-picking method is proposed in our approach. After ROR data *Ror*$_i$ [*t*] for frame *t* in band 1 is calculated, where i = 1, and t = 0,1,…,T for the frames whose absolute ROR value is greater than 9 dB and that satisfy the requirement shown below were picked out first.

$$[Ror_i(t-1)-Ror_i(t)][Ror_i(t+1)-Ror_i(t)]>0.5 \tag{1}$$

Then, we looked at two segments [*t*–8, *t*] and [*t*, *t*+8] for frame *t*. If both pieces had more than half a point, except *t* whose absolute ROR values are all greater or smaller than *t*'s, we kept the frame *t* down Lastly, just one peak with a maximum absolute ROR value was held and others were deleted in [*t*–10, *t*]. This method was executed in both coarse and fine processing.

3.2 G-Landmark Detection

The output of general processing peaks represent time points of abrupt energy changes in two bands. To locate g-landmarks, the peaks in band 1 are candidates because the energy in band 1 monitors the presence or absence of glottal vibration. Several energy and duration rules, designed from observation and acoustic knowledge, were used to select reliable unvoiced landmarks from peaks in band 1 and to remove others.

First of all, for two adjacent peaks within 20 ms with the same sign, the one with the maximum absolute ROR value was saved. After this processing, the peaks with small energy fluctuations caused by a bigger one were removed.

Second, a sentence must begin with a +peak, which labels the point at which the vocal folds start vibrating, and it must end with a –peak, which locates the stop of vibration.

The first valid +peak was found where the total energy of the segment between the +peak and the next -peak was larger than 40 dB, which made sure that this region was a speech segment. At the end of speech, if the last peak is a +peak, we deleted it until a valid –peak is the last one.

The +peaks and –peaks must be paired in speech corresponding to the beginning and termination of vocal folds' freely vibrating. For each +peak, the –peak before it locate the unvoiced speech segment. The –peak behind it labels the end of the voiced segment started at this +peak and a new unvoiced segment's start. In our system, for two adjacent +peaks, if they were within 200 ms, which is the maximum duration of syllable, the latter one was deleted because it is an energy jump for a new syllable's sonorant consonant. Otherwise, the one with a minimum absolute ROR value as deleted. For more than one –peak behind a +peak, the duration between the +peak and –peak must be longer than 80 ms, which is the minimum voiced segment duration. For connected –peaks, only the one with the maximum ROR value was kept. The energy rule, in which the average total energy in +/–peak duration must be larger than 40 dB, was used in here to remove creaky voiced speech segments. The +peak and –peak before it were removed if the energy rule was not satisfied.

Another energy a criterion was imposed to unvoiced segments where the average band 1 energy must be less than 45 dB, making sure that the vocal folds do not freely vibrate.

Finally, about 20% of the peaks were deleted, and the peaks left are the g-landmarks we looked for.

3.3 +s Landmark Detection

Based on the g-landmarks found previously, +s landmarks were then detected.

First, the ROR data $Ror_i\,[t]$ for frame t in band 2 as calculated, where $i = 2$, and $t = 0,1,\ldots,T$, and the frames whose ROR value was greater than 6 dB and satisfied the Eq. (1) were picked out first. Second, we looked at two segments [*t*–8, *t*] and [*t*, *t*+8] for frame *t*. If both pieces had more than half a point, except *t* whose absolute ROR values are all greater or smaller than *t*'s, we kept the frame *t* down. Finally, just one peak with a maximum absolute ROR value was held and the others were deleted in [*t*–10, *t*].

After the previous process, if a reserved peak is between a +g/–g landmark pair, then the peak is identified as a +s landmark.

3.4 VLCS Detection

The detection method of V-landmarks has been proposed based on the algorithm proposed by Howitt [10]. Instead of using a recursive convex hull method, for every +g/–g landmark pair, two frames between them were found: +peak and –peak. +peak is the frame that has the greatest ROR value in band 1 and band 2, and the –peak is the frame that has the lowest ROR value in band 1 and band 2. If the ROR value of the +peak is below 5 dB, then the +g landmark replaces the current frame of the +peak. If the ROR value of the –peak is greater than −5 dB, then the current frame of the –g landmark replaces the –peak. Then, the speech segment between the +peak and –peak is considered to be VLCS.

4. Landmark-Driven SM Decoding

4.1 Stochastic Segment Model

The stochastic segment model (SSM) is one type of segment model. SSM represents the variable length observation sequence by a fixed length region sequence. So, a resample function is needed to map the observation segment $x_1^N = \{x_1, x_2, \cdots, x_N\}$ to the fixed length frame sequence y_1^L . The re-sampled frame is measured by "region" model, which is similar to conception of the state in HMM, and L is the fixed length of the region sequence in each SSM.

$$y_i = x_{\left\lfloor \frac{i}{L} N \right\rfloor},\ 0 < i \le L \qquad (2)$$

where, $\lfloor z \rfloor$ is the maximum integer no longer than *z*. The log-likelihood of a segment x_1^N given model *α*:

$$\log\left[p\left(x_1^N \mid \alpha\right)\right] = \sum_{i=1}^{L} \log\left[p\left(y_i \mid \alpha, r_i\right)\right] \tag{3}$$

where, r_i is the *i*-th region model in the segment model *α*. Usually, each region consists of Gaussian mixture models.

The decoding of SSM is a two-level process [2]. The first level is segment classification. For a segment of speech x_τ^m, which starts at *τ* and ends at *m*, the segment model *α* with the highest likelihood score is found first:

$$D_m(\tau) = \max_{\alpha}\left\{\log\left[p\left(x_\tau^m \mid \alpha\right)\right](m-\tau) + \log[p(\alpha)] + \log\left[p_s\left(x_\tau^m \mid \alpha\right)\right]\right\}$$
$$0 \le \tau < m < T \tag{4}$$

where, $D_m(\tau)$ is the highest likelihood score for feature segment x_τ^m, $P(\alpha)$ is the language score, and $P_s\left(x_\tau^m \mid \alpha\right)$ is the segmental score (duration, etc).

Then, for each end point from 0 to T, the most suitable starting point *τ* is selected to construct a segment with the highest probability:

$$J^*(m) = \max_{\tau}\left\{J^*(\tau) + D_m(\tau) + C\right\},\ J^*(0) = 0$$
$$\max\{m - L_{ext}, 0\} \le \tau \le m \tag{5}$$

where, $J^*(m)$ is the accumulated score of the best acoustic model sequence at end point *m*, *c* is the insertion factor for each segment, and L_{ext} is the allowed maximum segment duration.

4.2 Decoding

In the first level process of SM decoding, the likelihood score of all candidate models must be calculated to get the most suitable model. Similarly, in the second level process of SM decoding, the likelihood score of all possible speech segments with different starting points must be calculated. Thus, a lot of data and computed resources are needed.

Landmarks can provide the phonetic class information of the segments adjacent to landmarks. According to the information, some segment models can be penalized or be directly removed without calculating their probability in the first level process of decoding. At the same time, landmarks can also provide the boundary information of initials and finals. Therefore, in the second level process of SM decoding, some speech segments can be directly removed without calculating their probability according to the boundary information. Thus, the decoding of SM can be accelerated.

In the second level process of SM decoding (Eq. (5)) for each end point, it is obvious that the most suitable starting point *t* is located at the boundary between the initial and final. If all the boundary information between initials and finals can be obtained according to landmarks, then only the speech segments whose start points are near the landmarks are left. And the probabilities of those speech segments are calculated to get the most suitable starting point *τ*. Eq. (4) can be modified as:

$$J^{*}(m)=\max_{\tau}\left\{J^{*}(\tau)+D_{m}(\tau)+C\right\},\ J^{*}(0)=0$$
$$\tau \le m\ ,\ \tau \in [L_{near}-d, L_{near}+d] \tag{6}$$

where, L_{near} represents the landmark that not only can provide the boundary information, but is also nearest to the end point m. d is the offset value. However, in practice, it is impossible to identify all the boundary information only according to landmarks. For example, it is difficult to set a clear boundary between a glide initial and a final, and the V-landmarks and G-landmarks cannot provide this boundary information. According to Section 2, both the g-landmark and s-landmark can provide the boundary information of initials and finals, but the –s landmark may also be inside finals. So, only the g-landmark and +s landmark must be located at the boundary between the initial and final. In the case where speech segments whose start points are behind the landmark are left, Eq. (5) is modified as:

$$J^{*}(m)=\max_{\tau}\left\{J^{*}(\tau)+D_{m}(\tau)+C\right\},J^{*}(0)=0$$
$$\max\{L_{near}-d, m-L_{ext}\} \le \tau \le m \tag{7}$$

In practice, the detection result of s-landmarks is unreliable [4,5]. Therefore, in the following experiment only the boundary information provided by g-landmarks is used.

The idea described above is how to guide the decoding of SM using the boundary information provided by landmarks. On the other hand, the non-boundary region information provided by landmarks can also be integrated into the process of decoding.

A vowel is produced with a maximum opening in the vocal tract, and a V-landmark can be located in the vowel where there is a maximum amplitude in the low frequency range. In the pronouncing process of vowels, the amplitude in the low frequency range changes slowly and smoothly, so the VLCS can be obtained according to the amplitude change. For a V-landmark, its corresponding VLCS (Fig. 2) is a speech segment that contains this V-landmark, and this speech segment is within the final that the V-landmark belongs to. Meanwhile, the amplitude in the low frequency range changes slowly and smoothly in this speech segment.

Because VLCS is located within finals other than at the boundary, in the second level process of SM decoding (Eq. (5)), any of speech segments whose start and end points are both located within the VLCS will be directly removed without calculating their probability and Eq. (6) is modified as:

$$J^{*}(m)=\max_{\tau}\left\{J^{*}(\tau)+D_{m}(\tau)+C\right\},J^{*}(0)=0$$
$$\max\{L_{near}-d, m-L_{ext}\} \le \tau \le m\ , \tau \notin [V_s, V_e] \text{ or } m \notin [V_s, V_e] \tag{8}$$

where, V_s is the start point of a VLCS and V_e is the end point.

In the first level process of SM decoding (Eq. (4)), if the starting point of current speech segment is near a landmark that can provide boundary information, some candidate models that are inconsistent with phonetic class information provided by the landmark will be directly removed without calculating their probability. For example, if the start point of the current speech segment is near a +g landmark, the candidate models must represent finals or glides, and their previous models must represent unvoiced consonants or silent intervals. If the start point of current speech segment is near a –g landmark, the candidate models must represent unvoiced consonants or a silent interval, and their previous models must represent finals.

Not all boundaries can be determined by landmarks. Therefore, the method only applies to the situation where the landmark used is close enough from the end point to ensure there are no other boundaries between the landmark and the end point.

5. Experiments and Analysis

The Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development provided the data corpus applied in our experiments. Eighty-three male speakers' data was employed for training (48,373 sentences, 55.6 hours) and six male speakers' for testing (240 sentences, 17.1 minutes). There are 24 syllable initials and 37 syllable finals in our Mandarin phonetic set. Each syllable final has five tones. The baseline was a context-dependent triphone SM recognition system built by Tang et al. [2]. To make the experiments comparable, a continuous density HMM (CDHMM) was developed as the baseline of HMM by HTK V3.2.1 [11]. The structure of HMM is left to right with 5 states, 3 emitting distributions and no state skipping, except the "sp" (short pause) model with 3 states and 1 emitting distribution. Sixteen Gaussian mixtures modeled each emitting distribution. A bigram language model with 48,188 words was used in both HMM and SM systems.

5.1 Landmark Detection

The results on the test set showed that most g-landmarks are near the corresponding boundaries (Table 4). Criterion 1 in Table 4 represents that the current landmark is before the corresponding boundary and the distance between the landmark and its corresponding boundary is less than 60 ms. Criterion 2 represents that the current landmark is after the corresponding boundary and that the distance between the landmark and its corresponding boundary is less than 60 ms. The boundary information was provided by forced alignment. "Corr" in Table 4 represents a proportion of +g/+s landmarks that met Criterion 1 or a proportion of –g landmarks that met Criterion 2.

Table 4. Detection result of g-landmarks

Kind of landmark	Kind of criterion	Corr (%)
+g landmark	Criterion 1	99.2
–g landmark	Criterion 2	98.9
+s landmark	Criterion 1	98.6

The detection result of VLCS on the test set is shown in Table 5. The third column is the sum of frames contained by all VLCSs, and the fourth column is the time of these frames. "Corr" in Table 5 represents a proportion of VLCS that are located within their corresponding finals.

Table 5. The detection result of VLCS

	Corr (%)	Frame no.	Time (min)
VLCS	95.4	45995	7.7

5.2 Experimental Setup and Result Analysis

System1 integrated the boundary information provided by g-landmarks into the process of decoding. Because most +g landmarks were located before the corresponding boundaries provided by forced alignment, the value of L_{ext} in Eq. (7) was set to 0 when the current landmark was a +g landmark. Most –g landmarks are located behind the corresponding boundaries and the distance between a –g landmark and its corresponding boundary (Distance2) was also less than 60 ms, so the value of L_{ext} in Eq. (7) was set to 60 ms when the current landmark was a –g landmark. The recognition results and decoding time of the HMM system and the baseline system are shown in the second and third rows of Table 6, and the recognition results and decoding time of the system by using g-landmarks are shown in the fourth row. Compared with the baseline system, the decoding time of system1 was reduced by 3.8 minutes. Meanwhile, the performance was slightly worse than the baseline system because of a few recognition errors by g-landmarks.

Table 6. Experiment results of HMM and SM based systems

System	WER (%)	Time (min)
HMM	15.60	-
Baseline	13.67	45.0
System1	13.86	41.2
System2	14.01	38.1
System3	14.58	31.6

System2 incorporates the VLCSs to system1 to guide the decoding process. The result can be seen in the fifth row of Table 6. Compared with system1, the decoding time of system2 was reduced by 3.1 minutes, and the recognition accuracy slightly declined.

Finally, the phonetic class information provided by the b-landmarks was used for pruning the candidate models in system3. If the current landmark was a +g landmark, the triphone models in which the left phonetic was an unvoiced consonant or silent interval and the middle phonetic was a final or glide were left. The rest of the models in the candidate set were directly removed. If the current landmark was a –g landmark, the triphone models in which the left phonetic was a final and the middle phonetic was an unvoiced consonant or silent interval were left. The results of the method are shown in the sixth row of Table 6. Compared with system2, although the WER (%) increased by 0.57, the decoding time of system3 was reduced by 6.5 minutes (17%).

Compared with the baseline system, the recognition accuracy of system3 decreased slightly. The reason is that we reduced the search space using information provided by landmarks, but not all landmarks can be reliably detected. Thus, a few correct paths were be removed because of the incorrect detecting of landmarks. Nonetheless, the recognition performance of system3 still outperforms the HMM system. Since the running time of HMM with a language model built by HTK is much slower than real systems, we did not compare it with the SSM system.

6. Conclusions

This paper proposes a preliminary framework to integrate phonetic landmarks into the segment-based Mandarin speech recognition system. The results showed that about 30% decoding time can be saved without there being an obvious influence on recognition accuracy and have demonstrated the potential of the framework.

Acknowledgement

This work was supported in part by the China National Nature Science Foundation (No. 61300124).

References

[1] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360-378, 1996.

[2] Y. Tang, W. J. Liu, H. Zhang, B. Xu, and G. H. Ding, "One-pass coarse-to-fine segmental speech decoding algorithm," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. 441-444.

[3] Y. Tang, W. Liu, Y. Zhang, and B. Xu, "A fast framework for the constrained mean trajectory model by avoidance of redundant computation on segment," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 11, no. 1, pp. 73-86, 2006.

[4] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417-3430, 1996.

[5] Z. Yang, W. Liu, and H. Chao, "An improved steady segment based decoding algorithm by using response probability for LVCSR," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, Hong Kong, 2012, pp. 306-310.

[6] Z. Yang and W. Liu, "A novel path extension framework using steady segment detection for Mandarin speech recognition," in *Proceedings of 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, 2010, pp. 226-229.

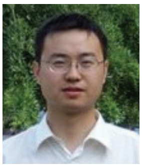
[7] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, et al., "Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, 2005, pp. 213-216.

[8] C. Y. Park, "Consonant landmark detection for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2008.

[9] J. Q. Han, L. Zhang, and T. R. Zheng, *Speech Signal Processing*, 1st ed. Beijing: Tsinghua University, 2005, pp. 20-23.

[10] W. Howitt, "Vowel landmark detection," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2279-2279, 2002.

[11] S. Young, G. Evermann, and M. Gales, *The HTK Book (Version 3.0)*. Cambridge: Microsoft, 2000, pp. 49-192.



Hao Chao

He received the B.S. degree in School of Computer and Information Engineering from HeNan University in 2002. He obtained his Ph.D. in Institute of Automation, Chinese Academy of Sciences in 2012. His current research interests consist of Speech Recognition, Pattern Recognition and Intelligent Information Processing.

Cheng Song

He received the M.S. degree in College of Computer Science and Technology, Henan Polytechnic University in 2002. He obtained his Ph.D. from Beijing University of Posts and Telecommunications in 2011. His current research interests are Trusted Computing and Intelligent Information Processing.