Interactive Speech Recognition System Research Postprint

Authors: Li Xinhui, Wang Xiangdong, Qian Yueliang, Lin Shouxun

Date: 2017-03-09T00:00:00+00:00

Abstract

To enable the practical application of large-vocabulary continuous speech recognition technology, this paper proposes the concept of interactive speech recognition and focuses on investigating its key technologies. Interactive speech recognition refers to equipping a speech recognition system with an operator who guides and supervises the system during recognition and corrects its output. Simultaneously, the recognition system learns from interaction information, adaptively adjusting its internal models based on the operator' s guidance and corrections to improve recognition performance. The research in this paper constitutes both development and innovation in the practical application of current large-vocabulary continuous speech recognition technology, with significant scientific and technological implications and promising industrial prospects. Moreover, it offers practical reference for speech recognition applications in other areas (e.g., real-time caption generation, library audio material organization).

Full Text

Preamble

Interactive Speech Recognition System Research Xinhui Li, Xiangdong Wang, Yueliang Qian, Shouxun Lin
Information Technology Letters, Vol. 8 No. 5

Abstract

To enable practical application of large vocabulary continuous speech recognition technology, this paper proposes the concept of interactive speech recognition and focuses on its key technologies. Interactive speech recognition involves equipping the speech recognition system with an operator who guides and supervises the recognition process and corrects recognition results. Simultaneously, the system learns from this interaction, adaptively adjusting its internal models

based on the operator' s guidance and corrections to improve recognition performance. This research represents a development and innovation in the practical application of current large vocabulary continuous speech recognition technology, holding significant scientific and technological importance and promising industrial applications. It also provides practical insights for speech recognition applications in other domains, such as real-time caption generation and audio archive organization in libraries.

Keywords: speech recognition, interactive speech recognition, speech utterance extraction, Chinese candidate generation, interactive acoustic model adaptation

1 Introduction

Speech is the most natural and important form of human communication [1]. Consequently, automatic speech recognition has long been a focus of significant attention from governments and researchers worldwide as a natural and efficient human-computer interaction modality. In recent years, speech recognition technology has made substantial progress. Medium and small vocabulary speech recognition technologies for specialized applications have matured [2, 3], leading to practical systems such as mobile phone voice dialing and telephone inquiry systems. However, due to limitations imposed by background noise, dialectal accents, colloquial natural speech, and semantic understanding, research on large vocabulary continuous speech recognition remains largely in the laboratory stage. The performance of large vocabulary continuous automatic speech recognition systems for real-world scenarios falls far short of practical application requirements.

While the concept of interactive speech recognition has not been explicitly proposed in existing research, some studies have introduced interaction into the speech recognition process. Early representative work came from IBM Corporation, Carnegie Mellon University (CMU), and the University of Michigan, focusing primarily on speech recognition error correction technology. In these systems, speakers correct errors in recognition results after each utterance is recognized. Multiple interaction modalities were provided, including word respeaking, spelling, keyboard input, handwriting input, pen-based device clicking, drag-and-drop input, and selection from N-best candidates [4-7]. A more recent representative work is the "speech repair" system from Japan' s National Institute of Advanced Industrial Science and Technology (AIST) [8]. This system provides multiple candidates for each word and offers a corresponding interface, allowing users to correct speech recognition results by selecting candidates during or after speech input. Targeting clean read speech, this system can achieve real-time application with over 96% accuracy after correction. However, it only provides a candidate selection interface without other interaction functions, and does not utilize user correction information for model adaptation. Consequently, performance degrades significantly in real natural speech scenarios such as meetings. Overall, research on interactive speech recognition

is limited, mostly concentrating on result correction, and lacks investigation into utilizing multiple interaction modalities and leveraging interactive information for acoustic model adaptation.

To advance large vocabulary continuous speech recognition technology toward practical application, this paper proposes the concept of interactive speech recognition, investigates its key technologies, and constructs a complete system. In this context, interactive speech recognition involves equipping the speech recognition system with an operator who interacts with the system during recognition. The interaction modes fall into two main categories: first, providing appropriate guidance to the system based on prior knowledge or characteristics of the current speaker' s voice, such as indicating speaker changes, topic switches, specifying speaker gender or dialectal accent type, or even inputting partial prior corpora into the system; second, manually correcting current speech recognition results based on auditory perception. Considering efficiency and interaction friendliness, this type of interaction primarily employs candidate selection, where after recognizing an utterance, multiple candidates are provided for each character. When the first candidate is incorrect, the operator can select the correct candidate from the alternatives or input the correct content to rectify recognition errors. In interactive speech recognition, the system not only corrects recognition errors through the operator' s rapid corrections but also selects and adaptively adjusts its internal models based on the operator' s guidance and interaction information. This makes the models better match the current speaker' s pronunciation characteristics and speech content, resulting in increasingly accurate candidate generation and higher correction efficiency for the operator, thereby meeting practical application requirements.

2 Speech Utterance Extraction

In speech recognition, good results are typically obtained by recognizing a complete utterance before outputting results. Therefore, when recognizing a speech segment, it is necessary to first extract utterances from the segment before performing recognition. Currently, endpoint detection methods are primarily used for utterance extraction. Endpoint detection technology identifies the start and end points of speech within a signal containing speech. In speech recognition, effective utterance extraction not only reduces system processing time and improves real-time performance but also eliminates noise interference from silent segments, thereby significantly improving subsequent recognition performance.

In interactive speech recognition, the input speech can be either pre-recorded audio files or real-time speech. The utterance extraction module should be capable of extracting utterances in both scenarios. [Figure 2: see original paper] illustrates the speech utterance extraction process for interactive speech recognition.

For audio file input, the system directly applies endpoint detection to extract all utterances from the file. For real-time speech input, the system continu-

ously captures the speaker' s voice and performs endpoint detection on the captured speech to extract utterances. To enable real-time utterance extraction in the latter case, this paper employs a segmented collection and buffer pool approach, where each fixed-length audio segment is placed into a buffer pool. As long as the buffer pool is not empty, a segment is retrieved for endpoint detection, with audio collection and endpoint detection accessing the buffer pool synchronously. In this method, the selection of the fixed audio length is critical: excessive length increases endpoint detection waiting time and affects real-time performance, while insufficient length generates many unnecessary detections, reducing system resource utilization. This paper sets the length to 3 seconds, as experimental statistics show that most utterances are within 3 seconds.

3 Chinese Candidate Generation

In interactive speech recognition, the candidate generation method directly determines the quality of generated candidates, which in turn determines the operator' s workload and efficiency throughout the recognition process. Abroad, the primary method for candidate generation uses confusion networks, employing confusion network algorithms [10-12] to compress word lattices into confusion networks to obtain candidates. This method requires that each arc in the word lattice corresponds to an individual, indivisible word. In English word lattices, each arc corresponds to a single English word, making this method suitable for generating English candidates. However, in Chinese word lattices, each arc corresponds to a word composed of one or more Chinese characters, where each word may be split into two or more characters (e.g., "中国" can be split into "中" and "国"). Therefore, this method cannot be used to generate appropriate Chinese candidates.

Through analysis of requirements in interactive speech recognition systems, we propose that Chinese candidate generation should satisfy three constraints: (1) Competing candidates should belong to the same candidate column, enabling operators to search for the correct candidate within a single column. (2) All candidate columns should be arranged in chronological order of recognition, allowing users to traverse and correct all recognition errors in a single forward pass. (3) Within each candidate column, all candidates should be arranged in descending order of recognition scores. Higher scores indicate greater likelihood of being the correct word, making it easier for operators to find the correct candidate when searching from top to bottom.

3.1 Character-based Chinese Candidate Generation Method

To generate high-quality Chinese candidates that satisfy the above constraints, we propose a character-based Chinese candidate generation method [13]. In this method, the Chinese word lattice is first aligned to generate an alignment network, and then words are split into characters to generate candidates based on the alignment network. [Figure 3: see original paper] illustrates the character-based Chinese candidate generation process, where (a) shows the alignment of

the Chinese word lattice to generate an alignment network, and (b) shows the generation of character-based candidates by splitting the alignment network. This paper describes the algorithm in two parts: word lattice alignment and character candidate generation.

Before describing the algorithm, we provide several definitions:

(1) **Chinese Word Lattice**: A Chinese word lattice is represented by a tuple $\langle N, E\rangle$, where N is the set of nodes and E is the set of arcs in the lattice. For all $n \in N$, $t(n)$ represents the time corresponding to node n. For all $e \in E$, each arc is represented by a 5-tuple $\langle e_S, e_F, e_W, e_A, e_L\rangle$, where e_S denotes the start node of arc e, e_F denotes the end node of arc e, e_W denotes the Chinese word on arc e, e_A denotes the acoustic probability score of arc e, and e_L denotes the language probability score of arc e.

(2) **Alignment Network**: An alignment network is represented by a tuple $\langle N, AE\rangle$, where AE is the set of all alignment classes. $AE = \{E'_1, E'_2, \ldots, E'_k\}$, where E'_k denotes the set of arcs at the k-th alignment position.

(3) **Character Candidate**: A character candidate is represented by a tuple $\langle C, L\rangle$, where C is the set of all candidate sets. $C = \{c_1, c_2, \ldots, c_m\}$, where c_l denotes the set of all candidates in the l-th candidate column. Each candidate c is represented by a binary tuple $\langle c_W, c_P\rangle$, where c_W denotes the candidate word corresponding to candidate c, and c_P denotes the score corresponding to candidate c.

3.1.1 Alignment Network Generation We can align a Chinese word lattice by clustering arcs to form an alignment network. Arcs clustered into the same class should satisfy two conditions: (1) The last characters of the word hypotheses corresponding to each arc exhibit phonetic similarity. (2) The arcs have temporal overlap.

The alignment network generation algorithm is described as follows:

Step 1: Use the forward-backward algorithm [10] to compute the posterior probability $p(e)$ for each arc e in the word lattice.

Step 2: Sort all arcs in set E by their end time $t_F(e)$ in ascending order. For arcs with equal end times, sort by their start time $t_S(e)$ in ascending order.

Step 3: Initialize $E'_0 = \emptyset$. For each arc $e \in E$, if $p(e) = \max_{e' \in E} p(e')$, then $E'_0 = E'_0 \cup \{e\}$.

Step 4: For each arc $e_i \in E$, where $i = 0, 1, ...$, assume $e_i \in E'_j$:

(a) If $t_S(e_i) - t_F(e_i) < \delta$, then $E'_{j+1} = E'_{j+1} \cup \{e_i\}$.

(b) If $\exists e' \in E'_j$ such that $\mathrm{SIM}(e_i, e') > \theta$, then $\exists e'' \in E'_j$ such that $\mathrm{SIM}(e_i, e'') > \theta$. If $\mathrm{SIM}(e_i, e') > \mathrm{SIM}(e_i, e'')$, then $E'_{j+1} = E'_{j+1} \cup \{e_i\}$. Here, $\mathrm{sim}(c(e), c(e'))$ represents the acoustic similarity between the

last characters of the words corresponding to arcs e and e', calculated using the most appropriate phonetic similarity formula. overlap(e, e') represents the smoothed temporal overlap degree between arcs e and e'.

(c) If $\exists e' \in E'_j$ and $K < I$ such that $\min\{\text{SIM}(e_i, e')\} > \theta$, then $E'_{j+1} = E'_{j+1} \cup \{e_i\}$. Here, $u(e)$ denotes the number of Chinese characters contained in the word corresponding to arc e.

(d) If $\exists e' \in E'_j$ and $K < I$ such that $\min\{\text{SIM}(e_i, e')\} > \theta$, and $\exists e'' \in E'_j$ and $K < I$ such that $\min\{\text{SIM}(e_i, e'')\} > \theta$, then $E'_{j+1} = E'_{j+1} \cup \{e_i\}$.

(e) If $\min\{\text{SIM}(e_i, e')\} > \theta$, $K \leq H \leq I$, where H is determined by: $H = \arg\max_{K \leq h \leq I}\{\text{SIM}(e_i, e')\}$. Here, $w(E')$ denotes the number of arcs contained in E', and $\text{SIM}(e_i, e')$ is defined as above.

Step 5: For each alignment class in E'_k, merge arcs with the same Chinese word into a single arc, with its probability value equal to the sum of the posterior probabilities of the merged arcs.

3.1.2 Character Candidate Generation Based on the alignment network, Chinese words are split to generate character candidates, and candidates in each column are sorted in descending order of probability scores.

The character candidate generation algorithm is described as follows:

Step 1: Let $n = 0$, $m = 0$.

Step 2: Let $\min\{u(e')\}$ be defined as before. For all arcs $e'_i \in E'_n$, where $i = 1, 2, 3, ...$, process as follows:

(a) If $u(e'_i) = 1$, let candidate c_j' s candidate word $c_{jW} = Q(W_{e'_i}, j)$, where $j = 0, 1, ...$, and $Q(W_{e'_i}, j)$ denotes taking the j-th Chinese character of the word corresponding to arc e'_i. Let candidate probability $c_{jP} = e'_i Q(W_{e'_i}, j)$.

(b) If $u(e'_i) > 1$, let candidate c_j' s candidate word $c_{jW} = Q(W_{e'_i}, j)$, where $j = 0, 1, ...$, and $Q(W_{e'_i}, j)$ denotes taking the j-th Chinese character of the word corresponding to arc e'_i. Let candidate probability $c_{jP} = e'_i Q(W_{e'_i}, j)$. Then $m = m + \text{num}(u(e'_i))$, where $\text{num}(u(e'_i))$ denotes the number of characters in the word corresponding to arc e'_i.

Step 3: $n = n + 1$, $m = m + \text{num}(n, w(E'))$. Return to Step 2 unless finished.

Step 4: For candidates in C'_k corresponding to the same candidate word, merge them into a single candidate with probability equal to the sum of the merged candidates' probabilities. If $c_W = \text{null}$, let candidate c'' s candidate word $c'_W = \text{null}$ and candidate probability $c'_P = -\sum c_P$.

Step 5: For the merged candidates, sort them in descending order of probability values.

3.2 Experiments and Results Analysis

In this experiment, we performed speech recognition on 278 self-recorded test utterances and generated candidates using the Chinese candidate generation method described above. The acoustic model used in the experiment was trained on over 40,000 sentences from the 863 corpus and over 70,000 sentences from northern dialect corpora. The language model was a binary language model trained on over 600 MB of text corpora. The evaluation metrics employed were: 1-Best accuracy, 10-Best coverage, average candidate rank, and candidate redundancy, calculated as follows:

- **1-Best Accuracy** = (Number of correct characters in 1-Best results) / (Total number of characters in reference)
- **10-Best Coverage** = (Number of correct characters in top 10 candidates) / (Total number of characters in reference)
- **Average Candidate Rank** = Average position of correct characters among candidates
- **Candidate Redundancy** = (Sum of all character candidates after the correct character) / (Total number of candidates)

Among these metrics, 1-Best accuracy reflects the inherent recognition performance of the speech recognition system (i.e., accuracy without candidate generation), while 10-Best coverage reflects how many correct words are included in the candidates, indicating how many recognition errors can be corrected through candidate selection. Average candidate rank and candidate redundancy evaluate candidate quality from the operator' s perspective. A lower average rank and lower redundancy enable faster location of correct words by the operator.

presents the experimental results evaluating the generated Chinese candidates using the above metrics.

Table 1: Chinese Candidate Generation Experimental Results

Metric	Result
1-Best Accuracy	76.848%
10-Best Coverage	92.468%
Average Candidate Rank	77.331%
Candidate Redundancy	

The results demonstrate that the proposed Chinese candidate generation method can correct most recognition errors. In this experiment, the generated Chinese candidates could correct over 15% of recognition errors. Moreover, based on the average candidate rank, most correct characters can be found within the first and second candidates.

4 Interactive Acoustic Model Adaptation

In interactive speech recognition, the quality of generated Chinese candidates is influenced not only by the candidate generation method itself but also by automatic speech recognition performance. This paper proposes an accent and gender-based acoustic model selection method and a supervised acoustic model adaptation method based on interactive information, both leveraging the operator' s guidance and corrective interaction information. The accent and gender-based method trains multiple acoustic models beforehand and selects the most appropriate model for each speaker based on operator-input information before recognition begins. The supervised adaptation method uses corrected recognition results and corresponding speaker speech for supervised acoustic model adaptation. Experimental results show that both methods can improve automatic speech recognition performance and consequently enhance candidate quality.

4.1 Accent and Gender-based Acoustic Model Selection

To improve speech recognition performance and the quality of generated Chinese candidates, this paper proposes an accent and gender-based acoustic model selection method utilizing operator guidance. In this approach, multiple models are trained based on accent and gender differences, and the acoustic model most similar to the recognition target' s pronunciation characteristics is selected before recognition begins. In China, people from different regions may pronounce the same character differently. For example, in Hunan province, people tend to pronounce "hu" as "fu" . Additionally, gender differences cause pronunciation variations, with female voices typically having higher pitch (i.e., higher frequency) compared to males. Therefore, this paper trains multiple acoustic models based on regional accent and gender, selecting the appropriate model for each recognition target based on their accent and gender information, which substantially improves speech recognition performance. [Figure 4: see original paper] illustrates the flowchart of accent and gender-based acoustic model selection.

The accent and gender-based acoustic model selection method consists of the following steps:

(1) **Train Multiple Acoustic Models Based on Accent and Gender**: Classify speech corpora by regional accent and gender, and train an acoustic model for each class. In this paper, we classified our research group' s accumulated northern accent Mandarin corpus and southern accent Mandarin corpus by north-south region and male-female gender, training four acoustic models (northern male, northern female, southern male, southern female).

(2) **Select Appropriate Acoustic Model Before Recognition**: Before recognition begins, the operator inputs information about the recognition target (primarily regional accent and gender), and the system selects the

most suitable acoustic model for each target and launches the corresponding recognition service process.

(3) **Real-time Switching During Recognition**: When the speaker changes during recognition, the operator marks the current speaker in the system, which then routes the current speaker' s utterances to the corresponding recognition service process.

4.2 Supervised Acoustic Model Adaptation Based on Interactive Information

In interactive speech recognition, every recognition result produced by the system is corrected by the operator. Leveraging these corrective interactions, this paper proposes a supervised acoustic model adaptation method based on interactive information. In this method, recognized speech and corresponding corrected results serve as adaptation training corpora for supervised acoustic model adaptation. [Figure 5: see original paper] illustrates the flowchart of supervised acoustic model adaptation based on interactive information.

The acoustic model adaptation method based on interactive information consists of the following steps:

(1) **Adaptation Corpus Collection**: During recognition, for each recognition target, we collect the extracted speech utterances and the corresponding text information corrected by the operator.

(2) **Supervised Acoustic Model Adaptation**: Using the collected speech corpora and corresponding text information, we perform supervised acoustic model adaptation for each recognition target' s acoustic model. This adaptation can be performed in two ways: (a) **Online adaptation**—when the collected speech corpus for any recognition target exceeds a certain threshold (measured in utterances and configurable), we perform supervised adaptation for its corresponding acoustic model; (b) **Offline adaptation**—after the entire recognition process is completed, we perform supervised acoustic model adaptation for each recognition target' s acoustic model, with the adapted models saved for future use.

(3) **Acoustic Model Switching**: This step primarily addresses online adaptation. To enable the adapted acoustic model to be quickly applied to subsequent speech recognition and improve subsequent system performance, we launch a recognition service process for the online-adapted acoustic model and close the recognition service process corresponding to the pre-adaptation model after successful launch.

4.3.1 Acoustic Model Selection

To verify the impact of accent and gender-based acoustic model selection on speech recognition performance and candidate quality, we pre-trained six acoustic models in this experiment: northern accent male, northern accent female,

northern accent mixed, southern accent male, southern accent female, and southern accent mixed. All six acoustic models were trained on uniformly sized corpora of 35,750 utterances, with mixed models using equal proportions of male and female speech. The language model was a binary language model trained on over 600 MB of text corpora. The test corpus consisted of 278 northern accent male utterances. Experimental results are shown in .

Table 2: Experimental Results of Model Selection on Northern Male Test Corpus

Model	1-Best Accuracy	10-Best Coverage	Average Candidate Rank	Candidate Redundancy
Northern Male Model	77.25%	89.35%		
Northern Female Model	57.39%			
Northern Mixed Model	75.07%	88.40%	67.52%	85.70%
Southern Male Model	73.09%	86.98%	60.64%	59.98%
Southern Female Model	49.49%		36.96%	57.48%
Southern Mixed Model	68.40%	74.46%	71.76%	56.40%

The bolded column in the results indicates the best experimental performance. For northern accent male test corpora, the northern accent male acoustic model achieved the best results. Moreover, northern accent acoustic models (including northern male, northern female, and northern mixed) outperformed southern accent acoustic models. Male acoustic models (including northern male and southern male) outperformed female acoustic models. These results demonstrate that accent and gender-based acoustic model selection can improve speech recognition performance and candidate generation quality.

4.3.2 Supervised Acoustic Model Adaptation

In this experiment, we split the previous test corpus in half. One half was used for recognition and correction, with the corrected text and corresponding

speech used to adapt the northern male acoustic model. The other half was tested using both the pre-adaptation and post-adaptation northern male acoustic models to obtain comparative results. The language model used throughout the experiment was a binary language model trained on over 600 MB of text corpora. presents the experimental results for the northern male test corpus.

Table 3: Comparison of Results Before and After Adaptation for Northern Male Test Corpus

Model	1-Best Accuracy	10-Best Coverage	Average Candidate Rank	Candidate Redundancy
Before Adaptation	77.52%	84.37%	62.89%	
After Adaptation	89.67%	95.41%	60.92%	

The experimental results indicate that using corrected information for acoustic model adaptation and continuing recognition with the adapted model yields better results than using the pre-adaptation model. Therefore, the results demonstrate that supervised acoustic model adaptation based on interactive information can improve speech recognition performance and candidate generation quality.

5 Conclusion and Outlook

Given that current large vocabulary continuous speech recognition cannot meet practical application requirements, interactive speech recognition opens a new application paradigm for speech recognition. In interactive speech recognition, operator guidance and interaction information should be fully utilized to improve speech recognition performance and candidate quality. Future work in interactive speech recognition includes:

(1) **Language Model Adaptation**: Language model adaptation significantly improves speech recognition performance. In interactive speech recognition, topic-related corpora can be collected before recognition based on the discussion topic to perform offline adaptation of the language model. Subsequently, during recognition, the language model can be adapted online based on operator corrections. Therefore, utilizing guidance and interaction information for language model adaptation holds promising prospects.

(2) **Training More Regional Acoustic Models**: This paper mentions selecting acoustic models similar to the recognition target' s pronunciation before recognition based on accent and gender. China has vast territory with distinct regional pronunciations, and almost all provinces have different Mandarin accents. Therefore, to improve speech recognition performance and candidate quality, future work can involve training more acoustic models based on pronunciation differences across regions.

In summary, interactive speech recognition represents a new application approach for current speech recognition technology and can be extended to other applications.

References

[1] Juang, B.H. and S. Furui, "Automatic Recognition and Understanding of Spoken Language - A First Step toward Natural Human-machine Communication" , *Proceedings of IEEE*, vol. 88(8): pp. 1142-1165, 2000.

[2] He Xiangzhi, "Research and Development of Speech Recognition" , *Computer and Modernization*, Vol. 79(3), pp.3-6, 2002.

[3] Yao Wenbing, Yao Tianren, "Robust Speech Recognition Technology Research" , *Computer Engineering and Applications*, vol.7, pp.69-71, 2002.

[4] Sharon Oviatt, Phil Cohen, et. al., "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions" , *Human-Computer Interaction*, vol.15 (4), pp.263 –322, 2000.

[5] Suhm, B., Myers, B., Waibel, A., "Designing Interactive error Recovery Methods for Speech Interfaces" , *Proceedings of ACM CHI 1996*, Workshop on Designing the User interface for Speech Recognition applications,1996.

[6] Bernhard Suhm, "Empirical Evaluation of Interactive Multimodal Error Correction" , *Proc. IEEE Workshop on Speech recognition and Understanding*, pp.583-590, 1997.

[7] Karat, C., Halverson, C., Horn, D., and Karat, "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems", *Proc. CHI*, pp.568-575, 1999.

[8] Jun Ogata, Masataka Goto, "Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces" , *Proc. EuroSpeech*, pp.133-136, 2005.

[9] S.Young, J.Jansen, J.Odell, D.Ollason and P.Woodland: *The HTK Book*, In Entropic Cambridge Research Lab., 1995.

[10] L. Mangu, E. Brill and A. Stolcke, "Finding consensus in speech recognition: word error minization and other application of confusion network," *Computer Speech and Language*, vol.14 (4), pp. 373-400, 2000.

[11] L. Mangu, *Finding Consensus in Speech Recognition*, PhD Thesis, Johns Hopkins University, 2000.

[12] J. Xue and Y.-X. Zhao, "Improved confusion network algorithm and shortest path search from word lattice," *ICASSP 2005*, vol.1, pp.853-856, 2005.

[13] Xinhui Li, Xiangdong Wang, Yueliang Qian and Shouxun Lin. Candidate Generation for Interactive Chinese Speech Recognition. *Proc. Joint Conferences on Pervasive Computing (JCPC)*, 2009, 583 - 588.

Author Biographies:

Xinhui Li: Graduate student, Pervasive Computing Center, Institute of Computing Technology, Chinese Academy of Sciences

Xiangdong Wang: Ph.D., Assistant Researcher, Pervasive Computing Center, Institute of Computing Technology, Chinese Academy of Sciences, xdwang@ict.ac.cn

Yueliang Qian: Senior Researcher, Director of Pervasive Computing Center, Institute of Computing Technology, Chinese Academy of Sciences

Shouxun Lin: Ph.D., Professor, Pervasive Computing Center, Institute of Computing Technology, Chinese Academy of Sciences

Note: Figure translations are in progress. See original paper for figures.