IMPROVED CHINESE CHARACTER INPUT BY MERGING SPEECH AND HANDWRITING RECOGNITION HYPOTHESES

Xi Zhou[1], *Ye Tian*[3], *Jian-lai Zhou*[2], *Frank K. Soong*[2], *Bei-qian Dai*[1]

[1]MOE-MS Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China

[2]Microsoft Research Asia, Beijing, China

[3]Microsoft Corporation, One Microsoft Way, Redmond, WA, USA
zhouxi@mail.ustc.edu.cn, { ytian, jlzhou, frankkps }@microsoft.com, bqdai@ustc.edu.cn

ABSTRACT

In this paper we propose to merge speech and handwriting recognition hypotheses together for improving the performance of Chinese character input. The recognition result of handwriting character input can be reliable when the character is written rather squarely. However, more legible of square handwriting tends to slow down the input (stroke writing) speed. On the other hand, speech input is fairly efficient but a large number of homonyms and its vulnerability to adverse environment prevent speech from being used as a robust Chinese character input method. The handwriting stroke information and acoustic speech information, in many cases, are complementary to each other. In this study we use independent, statistically trained HMMs for recognizing each input mode individually but merge recognition hypotheses from the two recognizers. Generalized posterior probabilities are used to synchronize, compare and merge hypotheses appropriately. Experimental results have shown that significant input speedup can be obtained while maintaining the same recognition performance.

1. INTRODUCTION

Handwriting based character input, as a natural human-machine user interface, has been widely used on many handheld devices. State-of-the-art handwriting recognition system has high accuracy, i.e, close to 99%, if the characters are written rather squarely. However, the speed of handwriting is one of the main concerns of the system because on the average many strokes are needed to input a character. This is especially true for Chinese.

To improve the speed of handwriting input, we take advantage of other input mode like speech. Speech input of a Chinese character is fast due to its monosyllable nature. However, speech input may not yield a high recognition performance especially in a noisy environment. For a Chinese character, its handwriting strokes and acoustic pronunciation form rather independent information. In most cases characters with a similar stroke structure are pronounced differently. Also homonyms are usually written rather differently. This motivates us to merge speech and handwriting recognition hypotheses for Chinese character input. The hypotheses merge is promising because: 1) The Chinese character input can be sped up; 2) The errors made by the handwriting and speech input are in general rather independent and they can work together to improve recognition accuracy.

There are many papers talking about combining different information sources into one pattern recognition system, such as audio-visual fusion [1] [2] which are fusing audio information and visual information for a robust speech recognition system. Experiments show that this is effective and significant improvement is achieved in noise environment.

To merge speech and handwriting recognition hypotheses for improving the performance of Chinese character input, we should solve the problem that the speech input is not necessarily synchronized with the handwriting input. Moreover, a unified framework is needed for calculating the confidence score for hypotheses of two channels. In our algorithm, state-of-the-art Hidden Markov model (HMM) based systems are used for both speech and handwriting recognition. Combination module is based on generalized posterior probability (GPP) which could synchronize and merge hypotheses coming from different sources. The experiment shows that Chinese character input efficiency is significantly improved while maintaining the same recognition performance.

The paper is organized as follows: In section 2 we describe the basic framework to combine speech and

handwriting inputs. In section 3, we introduce the combination system structure. In section 4, the experiments and results are shown. Finally conclusion and discussions are presented.

2. BASIC FRAMEWORK

In a dual input mode, the user can say a Chinese character while he is writing it. Time wise, the speech input needs not to be synchronized with the pen input. The user can say the character at any time in the (time) neighborhood of handwriting input; or without saying it. Thus hypothesis level combination is more appropriate than feature level combination. Each input mode should have its own recognizer. HMM based recognizer has been successfully applied to both automatic speech recognition [4][5] and handwriting recognition [6][7]. We follow this approach to facilitate a unified maximum likelihood based training and decoding framework. The unified framework is also essential to merge hypotheses from the two recognizers. Then Generalized Posterior Probability (GPP) based combination module is used in our framework to synchronize, compare and merge hypotheses.

Figure 1 shows an example of the basic framework of our system. User has provided handwriting input and speech input. Handwriting input includes only two strokes, which are the first two for Chinese character 我. User also provided speech data, which represents the character 我. Handwriting input is sent to a HMM based handwriting recognizer to hypothesize potential character candidates by using only two input strokes. Likewise, speech data is sent to HMM based speech recognizer. These hypotheses from two channels are sent to a GPP based combination module to generate a final output.

2.1. Speech recognizer

The speech recognition system uses mixture Gaussian densities with tied HMM states determined by standard decision tree. Cross word triphone models that ignore the word boundaries in the context are used. The system is produced by standard iterative mixture splitting using four embedded training per mixture configuration. 16 mixture components with diagonal covariance are trained for every tied-triphone state. The features used were 13 Mel frequency cepstral coefficients (MFCC) with delta and acceleration coefficients.

2.2. Handwriting recognizer

We need an on-line handwriting recognizer in order to incorporate speech recognition result as early as possible. Therefore, stroke is more appropriate than whole character as the basic model unit [7]. Beside that, other advantages of stroke-based HMM can be summarized as follows.

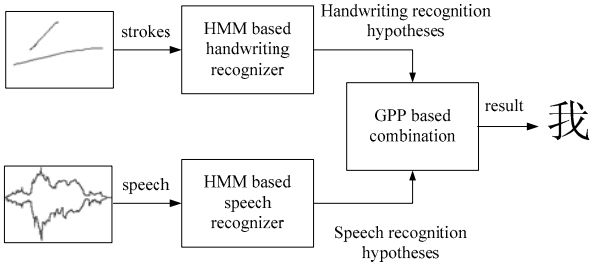

Figure 1 Basic framework

1. The memory required for storing stroke based HMM models and a dictionary is small and recognition speed is improved in an efficient stroke network.
2. Untrained characters can be recognized by just adding stroke sequence information to the dictionary.
3. Characters written with different stroke orders can be recognized by using multiple definitions of strokes per character in the dictionary.

In our handwriting stroke HMMs, 5 basic strokes (horizontal, vertical, left slant, right slant and other) are defined.

For each stroke, the original data is a point sequence from the input device (tablet). First, interpolation and resampling are performed to reduce distance variation between two adjacent online points and the variance of number of points in a stroke. Then tangent slope of each point is chosen as the handwriting feature. The tangent slope of a special point is approximated by averaging the slope between the point and its N neighbors.

2.3. Combination module

GPP is used to merge speech and handwriting hypotheses in the combination module. In our algorithm, we assume that speech and handwriting features are statistical independent.

$$P\left(o_{h,}o_s \mid H_i\right) = P\left(o_h \mid H_i\right) P\left(o_s \mid H_i\right) \qquad (1)$$

Where H_i are the hypotheses and o_h , o_s are the handwriting and speech observations. Then Posterior Probability (PP) can be represented as

$$P\left(H_i \mid o_{h,}o_s\right) = \frac{P\left(o_h \mid H_i\right) P\left(o_s \mid H_i\right) P\left(H_i\right)}{P\left(o_h, o_s\right)} \qquad (2)$$

Where $P\left(o_h \mid H_i\right)$ is handwriting model probability, $P\left(o_s \mid H_i\right)$ is acoustic model probability and $P\left(H_i\right)$ is language model probability, respectively.

As introduced in [3], GPP is a generalization of PP, we reweigh handwriting, speech model likelihood and

prior probability of character, so expression (2) is represent as

$$P(H_i \mid o_{h,} o_s) = \frac{P(o_h \mid H_i)^{\alpha} P(o_s \mid H_i)^{\beta} P(H_i)^{\gamma}}{P(o_h, o_s)} \qquad (3)$$

The exponential weights of the handwriting, acoustic and language models are labeled as α, β and γ, respectively.

Expression (3) is then used to rescore and to rerank all promising candidates. The top rank hypothesis corresponds to the one that yields the maximal score of $P(H_i \mid o_{h,} o_s)$.

3. SYSTEM DESCRIPTION

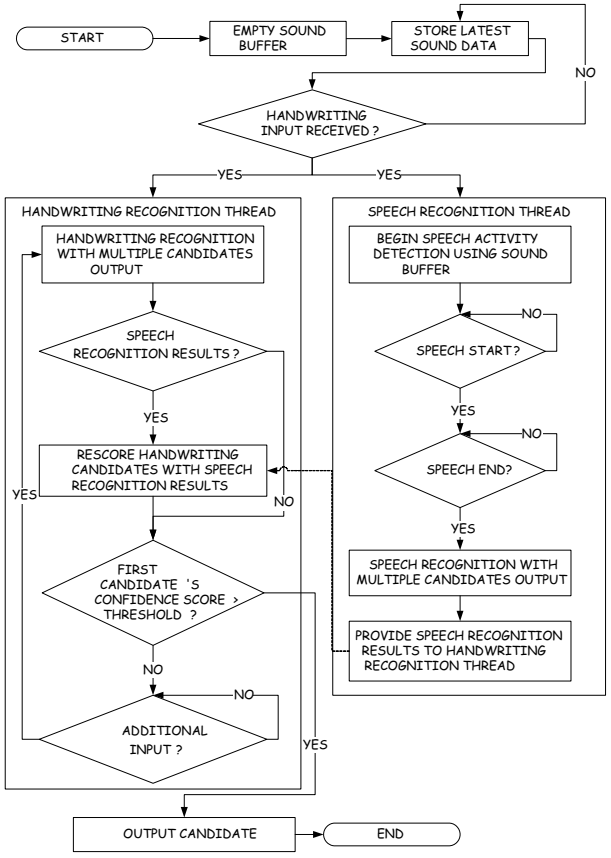

Figure 2 . Combination system structure

Figure 2 is a flow diagram of our system. The system contains one handwriting recognition thread and one speech recognition thread. The handwriting recognition thread is the main thread, and the handwriting thread reads the speech recognition thread's results.

The system starts with pen input detection and in the mean time buffers the microphone wave data. When handwriting (pen) input is active, the system will activate handwriting recognition decoding. At the same time, the speech recognition thread is started for detecting speech activity (endpoints) in the neighborhood when handwriting input becomes active. The starting point of speech detection process can be placed some interval before the pen activation time because input speech is stored in a wave data buffer.

When speech input is detected, the system will activate speech recognition decoding and save the multiple candidates in a word-graph or in an N-Best list and corresponding confidence measures.

Every time there is a new handwriting stroke input, the handwriting will revise recognition and output multiple candidates. If speech recognition results are available, the handwriting recognition results are rescored with speech recognition results. Once the system has enough confidence of the best candidate (i.e., its confidence score is better than a threshold), the input of the character is completed and the final result is output.

4. EXPERIMENT

In our experiment, we use a speech database and handwriting database which are collected independently to simulate simultaneous input. Speech data in this evaluation comes from MSRA's mandarin speech corpora [8], which is collected by 100 male speakers. Handwriting data used in this evaluation is collected from 5 writers, each wrote 100 most common Chinese characters. Each handwriting character is combined to the corresponding pronunciation spoken by randomly chosen 10 speakers. The lexicon in the experiment is 3,000 most common Chinese characters.

In Figure 3, we compare the efficiency of using handwriting input only with merging speech and handwriting input. X-axis is the number of strokes which have been input for a special character, while y-axis is the rank of the target character in the lexicon (3000 characters). Dashed line represents the average rank when we only consider handwriting input information, while solid line represents the average rank when considering the combined information for handwriting input and speech input. From the figure, we can see that fewer strokes are needed when we combine speech and handwriting input. Traditional handwriting needs 5 strokes to promote the correct word to top 10 candidates, while the combined approach only needs 3 strokes. We save 40% strokes input.

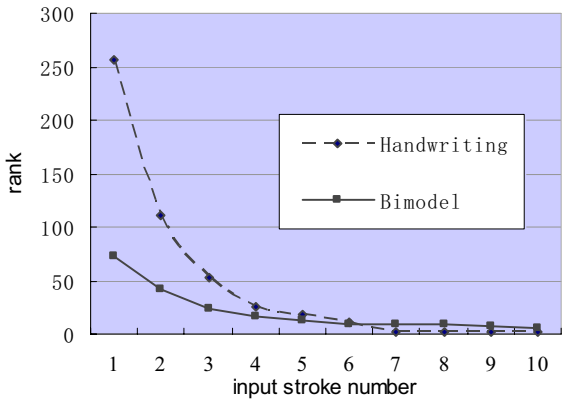

Figure 3. Recognition performance comparison between handwriting and bimodel input

5. CONCLUSION

Handwriting input may take a user long time to input and speech recognition input can suffer from high recognition error rate in noisy environment. In this paper, we propose to merge speech and handwriting recognition hypotheses together for improving the performance of Chinese character input. Our framework is expected to be:

1) Fast: user needs only to input a few handwriting strokes to get the final result.
2) Reliable: performance is greatly improved because we have information from two channels.
3) Flexible: User can speak while he is writing the character. User can also only write without speaking.
4) Robust: in our system, the speech input needs not to be exactly synchronized with the pen input.

6. DISCUSSION

Besides presented above, merging speech and handwriting recognition hypotheses also enables the following features:

1. Joint speech and handwriting model adaptation. We can use the final (better) results for unsupervised or lightly supervised adaptation of both the speech and handwriting models.
2. Improving continuous speech/handwriting recognition by merging multiple hypotheses (summarized in word-graph / N-best lists) of two input modes. In the current continuous speech and handwriting recognition, performance is not robust and many errors are caused by insertions and deletions besides substitution errors. By aligning the two sets of multiple hypotheses, we should be able to get better recognition performance.
3. Post-input editing help for correcting misrecognized characters. For example, n-best candidates of a character can be available on demand and the quality of n-best candidates can be adaptively improved on-line during the editing process when more misrecognized characters are corrected.
4. Better speech activity (endpoint) detection in a noisy environment with the "pen active" information.
5. Robust speech recognition with a better Signal-to-Noise-Ratio(SNR) estimate. With the "pen active" information, speech activity boundaries can be more reliably marked and SNR estimation can be improved.

Online demo of our system can be downloaded at http://mail.ustc.edu.cn/~zhouxi/multimodel.avi

7. REFERENCES

[1] M. Brooke and E. D. Petajan, "Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics," in *Proc. Int. Conf. Speech Input/Output, Techniques and Applications,* London, U.K., pp. 104–109, 1986,.

[2] Silsbee, P.L, "Sensory Integration in Audiovisual Automatic Speech Recognition," in *28th Annual Asilomar Conference on Signals, Systems, and Computers*, vol. 1, pp. 561-565. 1994.

[3] Wai Kit LO, Frank K. SOONG, Satoshi NAKAMURA, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels", in *Chinese Spoken Language Processing, International Symposium* 2004, pp.13 – 16, 2004.

[4] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition" in *Proceedings of the IEEE*, Volume 77, Issue 2, pp. 257 – 286, Feb. 1989.

[5] SJ Young, DJ Kershaw, JJ Odell, D Ollason, V Valtchev, and PC Woodland, "The HTK Book Version 3.2.1," Cambridge University, Cambridge, England, 2002.

[6] Thad Starnert , John Makhoul, Richard Schwartz, and George Chou, "On-line cursive handwriting recognition using speech recognition methods," *Proc. ICASSP* 94., vol 5, pp. 125-128, 1994.

[7] Mitsuru NAKAI, Naoto AKIRA, Hiroshi SHIMODAIRA and Shigeki SAGAYAMA， "Substroke Approach to HMM-based On-line Kanji Handwriting Recognition," *ICDAR* 2001: pp. 491-495, 2001.

[8] Eric Chang, Yu Shi, Jianlai Zhou, and Chao Huang, "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research," *Eurospeech 2001*, Denmark, 2001.