BIMODAL COMBINATION OF SPEECH AND HANDWRITING FOR IMPROVED WORD RECOGNITION

Pascale Woodruff and Stéphane Dupont

TCTS lab - FPMs & MULTITEL
Avenue Copernic, B-7000, Mons, Belgium
woodruff@tcts.fpms.ac.be, dupont@multitel.be

ABSTRACT

This paper presents a multimodal interface combining the use of speech and handwriting for isolated word recognition. Automatic Speech Recognition accuracy decreases as the perplexity of the task increases with the vocabulary size and the level of noise. The combination of different input modalities can improve the recognition performance. Handwriting is a modality that is natural to use, and can replace a keyboard on small portable devices, like Tablet PC's or PDA's. However this input method can be quite slow by itself. The proposed method in this paper combines both modalities by using handwriting to input only the first letters of a word, and speech to complete the word. The platform used to test this combination was a Tablet PC, using the Windows XP Tablet PC integrated handwriting recognition engine. Experiments were done based on a vocabulary of 35000 words. Relative word recognition improvements as high as 53% were obtained.

1. INTRODUCTION

Recent developments in the design of human-computer interfaces (HCI) aim at integrating different input sources (modalities) into a single system, to make the interface more natural and closer to human-to-human communication. Human beings use a variety of signals to communicate with each other, such as speech, gesture, eye-contact, facial expressions, etc. It is the combination of these signals that make communication natural and effective, and in order to reproduce this in human-computer interaction, research has taken the direction of multimodal interfaces. Since Bolt's original "Put that there" in 1980 [1], various approaches to multimodal integration have been published, from generic techniques and architectures [2],[3], to individual modality studies: speech, three-dimensional gestures [4], two-dimensional gestures [5], lip reading [6], etc. The speech modality is present in almost all published work, because it has the advantage of being a hands- and eyes-free modality, and is more adapted to an association with another modality that uses hands or eyes.

The more recent tendency is to integrate these technologies into small embedded devices, such as Tablet PC's, PDA's, or even mobile phones. These portable devices are not equipped with keyboards, but instead have a sensitive touch screen, used with a stylus, or pen. Therefore, the favoured input modalities that tend to be combined by researchers for these devices are speech and pen ([7],[8]).

In speech-only based interfaces, the accuracy of current automatic speech recognition (ASR) systems falls drastically as the vocabulary size increases, and as the noise environment becomes different from training conditions. The pen can complement or substitute speech by providing an extra source of information to the system, which is insensitive to noise. An example of this is the basic "tap-and-talk" concept [7], where the user must tap with the pen on a button in order to talk, providing a discriminating information to the system about when to start speech recognition.

In this paper, we consider a particular aspect of the use of pen : handwriting, or more precisely, on-line handwriting, meaning the data is captured immediately while text is being written, as opposed to off-line handwriting, where data is obtained from a scanned document. On-line handwriting recognition, as well as speech recognition, has seen great development in the past years [9],[10]. This technology generally produces better recognition rates than speech on large vocabularies. But it has the inconvenience of being quite a bit slower. Indeed, on average, people write at a rate of 15 to 25 words per minute, while they speak at a rate of 120 to 170 words per minute.

The method presented in this paper combines both speech and handwriting to enter text into a system: by handwriting the first couple of letters of a pronounced word. The goal is to improve large-vocabulary word recognition in dictation-type applications. The proposed method benefits from the strengths of both modalities. In order to input some text, instead of writing the whole word, the user must only enter the first letters of the word, which does not take much longer than to pronounce the word. This way, the speed feature of speech input is preserved but at the same time extra information is entered into the system.

In Section 2, we will describe the system that was developed in order to test the efficiency of this combination. Section 3 explains the multimodal fusion method used, and Section 2 gives the results that were obtained.

2. SYSTEM DESCRIPTION

A multimodal interface was developed using a Tablet PC as platform. It consists of a blank area where a user can freely write, using a stylus, with no constraints of direction or size of characters. The virtual "ink" captured is then analysed by Windows XP Tablet PC's integrated handwriting recognition engine. A close-talk microphone is used to capture the user's voice, which is then processed by our speech recognition system, on the basis of a large vocabulary of 35000 French words. On the handwriting side, three dictionaries were created. The first one contained all the letters of the French alphabet, the second one contained all the possible pairs of letters that exist at the beginning of French words, and the third one contained all the possible triplets of letters

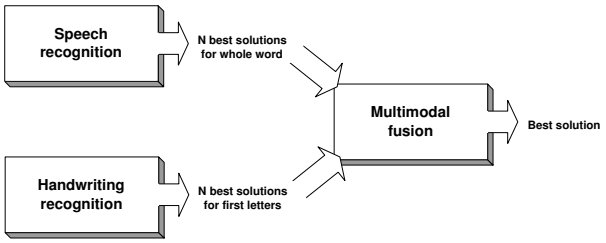

Figure 1: The multimodal fusion system

at the beginning of French words. These were used to test handwriting recognition of one, two, and three first letters of words.

As described in Figure 1, once the speech and handwriting signals are processed through their respective recognition modules, the N best recognition results of each modality are outputted. These results are then sent to a *fusion* module, which calculates and returns the best solution by combining information from both modalities.

In order to test the efficiency of the system, a database of nearly two thousand samples of isolated words were collected from 25 speakers. Words were randomly chosen from the 35 thousand word French vocabulary. Speakers pronounced each word while writing one, two, or three letters. Recording conditions were from a working office environment. It was interesting to notice that some people had difficulty writing and speaking simultaneously, and preferred writing the letters first and then pronouncing the word, while others could perform the task with no trouble at all. The database was separated into training and testing subsets.

After recognition of the audio and ink samples, the lists of N best solutions from both modalities were saved into files for subsequent analysis by the multimodal fusion module. The next section will explain exactly how that fusion module combines those lists and decides on the best solution.

3. MULTIMODAL FUSION METHOD

Bayesian theory

The goal of the multimodal fusion module is to determine which word w_i has the maximum probability of being the correct word, given the observation vectors O_1 of the first modality, and O_2 of the second modality. Thus, the best word must maximize the probability below, called *a posteriori* probability:

$$w_{best} = \arg\max_{w_i} P(w_i|O_1,O_2)$$

According to Bayes's theorem, this probability is equivalent to:

$$w_{best} = \arg\max_{w_i} \frac{P(O_1,O_2|w_i)P(w_i)}{P(O_1,O_2)}$$

Assuming that the observation vectors of both modalities are independent (which is not exactly the case, but is a usual hypothesis for sensor fusion [11]), it becomes:

$$w_{best} = \arg\max_{w_i} \frac{P(O_1|w_i)P(O_2|w_i)P(w_i)}{P(O_1)P(O_2)} \quad (1)$$

Where:

- $P(O_1)$ and $P(O_2)$ are called *a priori* probabilities of observing vectors O_1 and O_2. These probabilities do not depend on w_i and will not influence the choice of the best word w_{best}.
- $P(w_i)$ is the *a priori* probability of the word w_i. It is a constant in our case because it is assumed that all words in the task have the same chance to be pronounced.
- $P(O_1|w_i)$ and $P(O_2|w_i)$ are the *likelihoods* of observing respectively vectors O_1 and O_2 considering the word w_i has been inputted in the system.

By removing from expression 1 the terms that are constant or independent of w_i, we get:

$$w_{best} = \arg\max_{w_i} P(O_1|w_i)P(O_2|w_i) \quad (2)$$

In our case, O_1 is the observation vector of the speech modality, and O_2 is the observation vector of the handwriting modality. For the former, w_i represents the whole word, while for the latter, it represents just the first letters of the word. Below are described what these vectors represent, and the methods used to determine the terms of equation 2 for each modality.

Speech

For the speech modality, the observation vector O_1 is the sequence of acoustic features extracted from the voice audio signal every 10 *ms*, hence $O_1 = [o_{11}, o_{12}, ...o_{1T}]$, where T is the number of feature frames. The ASR recognizer used is a hybrid HMM/ANN system. As its name implies, this type of system combines Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) [12]. It uses HMM's to model each phoneme of the language (in this case, French), and the phoneme models are joined together to form word models, such as in Figure 2. The system is based on the feature extraction and modelling setups described in [13]. It has been imported for better modelling of inter-frame correlation.

An HMM consists of a series of states q_j to which are associated *emission* probabilities $P(o_{1t}|q_j)$ and state *transition* probabilities $P(q_j|q_i)$ [14], such as in Figure 2. In our case, o_{1t} represents the acoustic features extracted from the voice audio signal at time t.

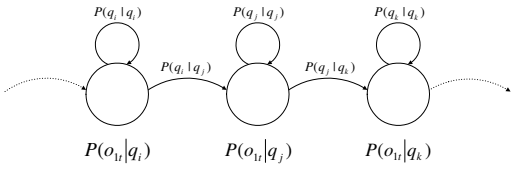

Figure 2: Hidden Markov Model (HMM) of a word

In our hybrid system, the HMM's are reduced to a simple form, by considering equal state transition probabilities between phonemes. After assembling the HMM's of every phoneme constituting a word w_i, we can calculate the probability of the whole sequence of states by multiplying the emission probabilities :

$$P(O_1|w_i) = \prod_{n,j} P(o_{1t}|q_j) \quad (3)$$

Table 1: Part of the confusion matrix of rank 1 of the N best hypothesis list of handwriting recognition system

	a	b	c	d	e	f	g	h	i	j	k	l	
a	56	0	4	0	0	0	0	0	0	0	0	0	60
b	4	51	0	1	0	2	0	9	0	1	1	1	70
c	4	1	50	1	9	0	0	0	0	0	0	1	66
d	1	0	0	61	0	0	0	0	0	0	2	0	64
e	3	0	0	0	26	0	0	0	0	0	0	2	31
f	4	4	0	0	1	53	0	2	0	0	2	1	67
g	0	0	0	0	1	4	57	1	0	1	1	0	65
h	1	2	0	1	0	4	0	54	0	0	5	1	68
i	0	0	0	1	0	0	0	0	69	0	0	0	70
j	1	0	0	0	0	0	0	0	2	58	0	0	61
k	0	0	0	0	0	0	0	5	0	0	62	0	67
l	0	0	0	3	0	0	0	0	0	0	0	65	68

The individual probabilities $P(o_{1t}|q_j)$ of each phoneme are calculated by the ANN, that takes as input the acoustic features o_{1t}. Thus, given an input acoustic signal, the probability that a certain word was pronounced can be calculated by Equation 3. This is done for every word in the given vocabulary, which are then listed by probability order in an N-best list. Equation 3 gives us the first *likelihood* needed in equation 2.

Handwriting

The handwriting recognition system outputs the N best hypothesis list for every written group of letters. This list is what we called the observation vector $O_2 = [o_{21}, o_{22}, ...o_{2N}]$, where o_{2n} is the outputted hypothesis at "rank" n of the list.

However, the corresponding probabilities are not accessible through Microsoft Tablet PC's SDK. Therefore, we used the training subset of the database to estimate them statistically. A *confusion matrix* was computed, by counting the number of times the recognizer confused a letter with another. This was done for every rank of the N best list. For example, the sample matrix of Table 1 was computed for the first rank, or the best solution given by the handwriting recognizer. The first value of the matrix means that 56 times, the recognizer outputted the letter 'a' as best solution when 'a' was actually written. On the same row, the value "4" means that 4 times the recognizer outputted 'c' as best solution when 'a' was written, etc. In total, the letter 'a' was written 60 times. Thus, when 'a' was written, the probability that it will be recognized as an 'a' at the first position of the N best list is estimated by $P(o_{21}|a) = \frac{56}{60}$.

By taking into account all the ranks of the N best list, it is possible to calculate the probability that the group of letters w_i was written, given that list:

$$P(O_2|w_i) \simeq \prod_n P(o_{2n}|w_i), \quad (4)$$

assuming the observations at different ranks are independent. This, of course, is an approximation.

When we have several written letters, the probabilities of each rank $P(o_{2n}|w_i)$ will be estimated as a product of probabilities of the individual letters o^j_{2n}:

$$P(o_{2n}|w_i) \simeq \prod_j P(o^j_{2n}|w^j_i)$$

Equation 4 gives us the second *likelihood* needed in equation 2.

Multimodal fusion

Given the observations of both modalities, we can now calculate the terms of equation 2, which we want to maximize for a certain word w_i.

However, both information sources should not influence equation 2 in the same way. The handwriting recognition is somewhat more reliable than the speech recognition. Hence, it should have a greater impact on the estimation. To take this into account, similarly to [6], we will weight the probability of the first modality with an exponent , and the probability of the second modality with $(1 - \ )$. The optimal "reliability weight" will be determined to produce the best combined results. Hence,

$$w_{best} = \arg\max_{w_i} P(O_1|w_i) \ \ P(O_2|w_i)^{1-} \quad (5)$$

In the log domain, this gives:

$$w_{best} = \arg\max_{w_i} [\ \cdot log(P(O_1|w_i)) + (1 - \ ) \cdot log(P(O_2|w_i))]$$

4. RESULTS

After the samples were collected of combined pen/voice entry of isolated words, they were processed through the speech and handwriting recognition systems. Then the results were combined using the method described in Section 3. The terms in equation 5 were computed for every word of the speech N best list, and the maximum value determined the choice of the best recognition hypothesis. This was done for different values of the reliability weight , and results were plotted on Figure 3 to determine the optimal value. The three curves represent the multimodal recognition scores for 1, 2 and 3 handwritten letters. The curves show that the peak moves towards the left when the number of handwritten letters increases. This shows that, when more letters are written, more credit should be given to the handwriting probability.

It is also interesting to calculate the fusion scores by considering that the handwriting recognizer is always right (instead of outputting a list of N best answers, it outputs only one, which is correct). In that case, the probability $P(w_i|O_2)$ in equation 5 will either be equal to one (for the outputted solution) or zero (for the other solutions). This means that, in the N best list given by the speech recognizer, we select only the words beginning with the correct letters (those for which $P(w_i|O_2)$ is equal to one), and we choose the best one. By doing this, we calculate the maximum performance that would be attained by using this method coupled with a deterministic character input mechanism. It simulates for instance the use of a virtual keyboard, instead of handwriting, for entering some letters of the word.

The recognition rates we obtained are shown in Table 2. The columns contain recognition rates of speech only, combined speech and handwritten first letters, limit of performance attainable with a perfect handwriting recognizer, absolute and relative improvements. Relative improvements represent the error reduction after combination. Error was reduced by respectively 26.78%, 33.62%, and 53.22%, by adding one, two and three written letters. The performance would continue to improve by adding more handwritten letters, but then the system would loose in speed.

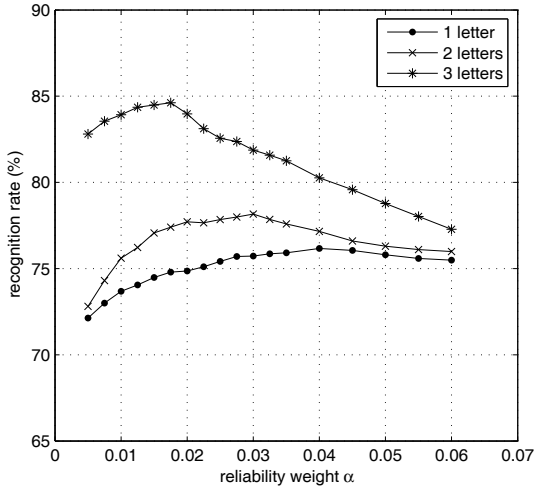

Figure 3: Determination of the optimal weight for the multimodal combination of one, two or three handwritten letters (task = isolated words dictionary query; perplexity = 35000)

Table 2: Recognition rates for the multimodal combination of speech and one, two, or three handwritten letters (n = number of handwritten letters)

n	speech	combined	limit	abs. imp.	rel. imp.
1	67.10%	75.91%	80.13%	8.81%	26.78%
2	67.10%	78.16%	84.48%	11.06%	33.62%
3	67.10%	84.61%	87.77%	17.51%	53.22%

5. CONCLUSION AND FUTURE WORK

This work has presented a multimodal text input approach using a combination of speech and handwriting to input words into a system. Although the combination rules that have been applied are fairly simple, the preliminary results obtained show a considerable improvement compared to speech alone in terms of acuracy, and to handwriting alone in terms of text-entry speed. Future work could be undertaken to extend the system to continuous speech, by adding grammar rules to allow only certain sequences of word types. This would involve studying user behaviour as well as temporal synchronization between both modalities. Another improvement of the system would be to allow any number of handwritten letters, which would give more flexibility to the system, but would introduce a new uncertainty. It should also allow the user to completely skip writing letters for shorter words, like prepositions and articles. Finally, further research could be done for fusion methods at earlier stages of the recognition chain, based for instance on Graphical Models or Multistream Models [6].

ACKNOWLEDGEMENTS

This work has been partly supported by the DGTRE division of the Walloon Region, under project F3M. The work on ASR has been partly supported by the EU 6th Framework Programme, under contract number IST-2002-002034 (project DIVINES). We also thank L. Couvreur, J.M. Boite and C. Ris for their contributions to the ASR system development, and T.Dutoit and S. Deketelaere for useful discussions about the approach. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] Richard A. Bolt. 'put-that-there': Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270. ACM Press, 1980.

[2] Lizhong Wu, Sharon L. Oviatt, and Philip R. Cohen. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, 1999.

[3] H. Djenidi, A. Ramdane-Cherif, C. Tadj, and N. Levy. Generic multi-agent architectures for multimedia multimodal dialogs. In *Proceedings of the Second International Workshop on Modelling of Objects, Components, and Agents (MOCA'02)*, pages 29–46, 2002.

[4] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma. Toward natural gesture/speech HCI: A case study of weather narration. In *Proc. Workshops on Perceptual User Interfaces*, pages 1–6, November 1998.

[5] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of ACM Multimedia 1997, 31-40*, 1997.

[6] S. Dupont and J. Luettin. Audio-visual speech modelling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2:141–151, September 2000.

[7] X. Huang, A. Acero, C. Chelba, L. Deng, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Wang, and Y. Wang. MIPAD: A next generation pda prototype. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

[8] S. Dusan, G.J. Gadbois, and J. Flanagan. Multimodal interaction on pda's integrating speech and pen inputs. In *EUROSPEECH-2003*, pages 2225–2228, 2003.

[9] T. Starner, J. Makhoul, R. Schwartz, and G. Chou. Online cursive handwriting recognition using speech recognition techniques. In *International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 125–128, 1994.

[10] S. Manke, M. Finke, and A. Waibel. NPen++: A writer independent, large vocabulary on-line cursive handwriting Recognition System. In *Proceedings of the International Conference on Document Analysis and Recognition*. Montreal, 1995.

[11] John J. Sudano. Equivalence between belief theories and naïve bayesian Fusion for systems with independent evidential data: Part I, the theory. In *Applications of Plausible, Paradoxical, and Neutrosophical Reasoning for Information Fusion*, pages 1239–1243, Cairns, Queensland, Australia, 2003.

[12] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[13] S. Dupont and C. Ris. Robust feature extraction and acoustic modeling at Multitel: Experiments on the Aurora databases. In *Proc. Eurospeech 2003*, pages 1789–1792, Genève, 2003.

[14] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257–285, 1989.