An Integrative Recognition Method for Speech and Gestures

Madoka Miki
Graduate School of Information Science, Nagoya University
1 Furo-cho Chikusa-ku
Nagoya, Japan
miki@sp.m.is.nagoya-u.ac.jp

Chiyomi Miyajima
Graduate School of Information Science, Nagoya University
1 Furo-cho Chikusa-ku
Nagoya, Japan
miyajima@is.nagoya-u.ac.jp

Takanori Nishino
Center for Information Media Studies, Nagoya University
1 Furo-cho Chikusa-ku
Nagoya, Japan
nishino@media.nagoya-u.ac.jp

Norihide Kitaoka
Graduate School of Information Science, Nagoya University
1 Furo-cho Chikusa-ku
Nagoya, Japan
kitaoka@nagoya-u.jp

Kazuya Takeda
Graduate School of Information Science, Nagoya University
1 Furo-cho Chikusa-ku
Nagoya, Japan
kazuya.takeda@nagoya-u.jp

ABSTRACT

We propose an integrative recognition method of speech accompanied with gestures such as pointing. Simultaneously generated speech and pointing complementarily help the recognition of both, and thus the integration of these multiple modalities may improve recognition performance. As an example of such multimodal speech, we selected the explanation of a geometry problem. While the problem was being solved, speech and fingertip movements were recorded with a close-talking microphone and a 3D position sensor. To find the correspondence between utterance and gestures, we propose probability distribution of the time gap between the starting times of an utterance and gestures. We also propose an integrative recognition method using this distribution. We obtained approximately 3-point improvement for both speech and fingertip movement recognition performance with this method.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Experimentation, Human Factors

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

Keywords

multimodal interface, speech recognition, gesture recognition, integrative recognition

1. INTRODUCTION

In human-human communication, such multiple modalities as speech and gestures are often used. Such multimodal interaction is expected in human-machine communication where multiple modalities sometimes play complementary roles with each other. The semantics expressed in a modality may be ambiguous, but the other modality might be able to remove them. Thus multiple modalities must be recognized and understood integratively. Combining speech and gesture is a typical example of such multimodality. In speech, such demonstratives as "this" and "here" are often used because some correspondences cannot be solved only by context in sequences of utterances. Gestures such as pointing may increase the understanding of the meaning of demonstratives when they appear with gestures.

When the task difficulty increases, users often prefer multimodal interactions rather than unimodal ones in a system that has speech and pen modalities [1, 2]. This result implies that smooth completion of complex transactions needs multimodality that includes the selection of methods to express complex intentions.

In this paper, we consider bimodal input with speech and fingertip movement on a desk. To understand such bimodal input, we divide this problem into three subproblems: individual recognition of speech and fingertip movements; finding the correspondence between utterances and fingertip movements in sequences of utterances and movements; and simultaneous recognition and understanding of bimodal input considering both of these modalities.

Methods adopting image processing have been proposed to recognize gestures including fingertip movements. Head and hand positions were tracked using video in [3]. A fingertip was tracked using images captured from the side of a human in [4]. Some research uses position sensors to acquire a fingertip position [5]. Touch pens and panels may

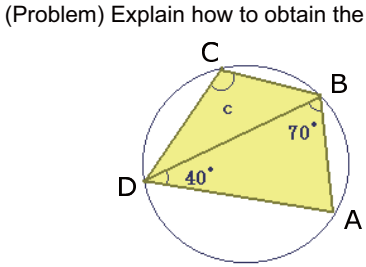

Figure 1: Mathematical problem: calculating an angle in a quadrilateral inscribed in a circle

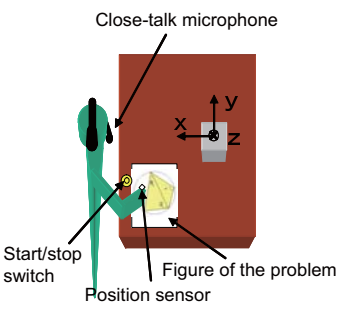

Figure 2: Recording setup

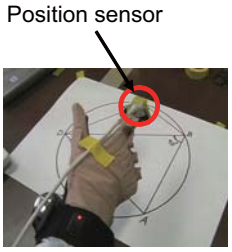

Figure 3: Position sensor settings

Figure 4: Examples of finger tip movement indicating an angle (Size of movements may be quite different as shown by solid arrows.)

help obtain pointing [6, 7]. After individually recognizing speech and gestures, correspondence between them must be found. An utterance and a gesture that express identical contents are paired. For such pairing, temporal inclusion and order [6, 8], semantical compatibility [6], and the relation between the prosodic features in speech and the speed of hand/finger movements [3] have been used. Finally, the information from speech and gestures is interpreted to an integrated representation. To achieve this integration, the following schemes have been proposed: a graph-base optimization method [9], a finite-state parsing method [10], a unification-based parsing method [11], and the integration of multimodal posterior probabilities [12]. Qu et al. [7] proposed the use of information obtained from gestures to improve speech recognition performance. Although multiple feature streams from multiple modalities may be integrated and recognized simultaneously, as in bimodal audio-visual speech recognition, since this approach only succeeds when the modalities are well synchronized with each other, it cannot be applied to integrate speech and gestures.

To understand speech with gestures, we propose an integrative recognition method with which we investigate an explanation task of a geometry problem. In this task, pointing is often accompanied with utterances because individual modalities are ambiguous in many cases.

In this paper, we first introduce the task setting and the recording of the multimodal explanations in Section 2. Then we explain our speech and gesture recognition methods in Section 3. In Section 4, we explain how to find multimodal alignment, and we propose an integrative recognition method in Section 5.

2. TASK SETTING

The problem we used as the task is shown in Figure 1. Subjects explained how to solve the problem while pointing at the figure. Speech and pointing were recorded with a close-talking microphone and a 3D position sensor attached to the tip of the index finger (Figures 2 and 3) at sampling frequencies of 48 kHz and 100 Hz, respectively. Six subjects (four males and two females) performed eight trials. Before recording, we explained that the subjects could use such demonstratives as 'this' and 'here' by pointing at the figure instead of such precise terms as 'angle ABC.' Subjects pushed a button to start/stop the synchronized recording at the start/end of recording. The total length of the recorded data was 417.8 sec (26.1 sec/trial).

3. SPEECH RECOGNITION AND GESTURE RECOGNITION METHODS

3.1 Speech recognition method

We performed speech recognition experiments for recorded explanation utterances. We used a network grammar that accepted a sequence of such elements as expressing "∠ADB= ∠ACB" etc. Since subjects explained while thinking and thus often used fillers and disfluences, the grammar accepted fillers between any words. The vocabulary size was 77 words. Triphone HMMs were used as the acoustic models trained from the CSJ corpus [13], which is suitable for spontaneous speech. Each HMM had three states with output probabilities, and the sampling frequency was 16 kHz. Frame length and shift were 25 and 10 ms, respectively, and 12-dimensional MFCC and its delta with delta log power were used as features.

We obtained a 75.0% recognition rate and 66.7% accuracy.

3.2 Gesture recognition

We also performed gesture recognition. In our task, the system must recognize such gesture pointing items as ∠ABC, segment AB, and vertex A from the fingertip movements.

Although remarkable individual differences were observed in the gestures and their size varied by individual, the direction of the fingertip movements was almost identical when pointing to the same item, as shown in Figure 4. So we used the differentials of the fingertip position in the X-Y plane as recognition features:

$$\boldsymbol{v}[n] = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} x[n] - x[n-1] \\ y[n] - y[n-1] \end{pmatrix}, \qquad (1)$$

where n indicates the time and $x[n]$ and $y[n]$ describe the fingertip position in the X-Y plane obtained by a 3D position sensor. The graph at the bottom of Figure 5 shows an example of the time sequence of $(\Delta x, \Delta y)$ indicated by arrows.

We used 3-state HMMs with single mixtures to model the finger movements. The 3D fingertip positions were recorded with a sampling frequency of 100 Hz. 19 of the 21 modeled gestures must be recognized, including eleven tracing gestures for segments, four gestures for vertices, one gesture describing an arc between two segments, and three gestures without pointing at an item (pointing for angles were expressed by one or combinations of these gestures). Gestures without pointing included between-pointing gestures, pushing the start/stop switch, and touching the desk without pointing at items. In our task, the finger position in the z-axis is important to recognize gestures because meaningful movements occur when a finger tip is on the table.

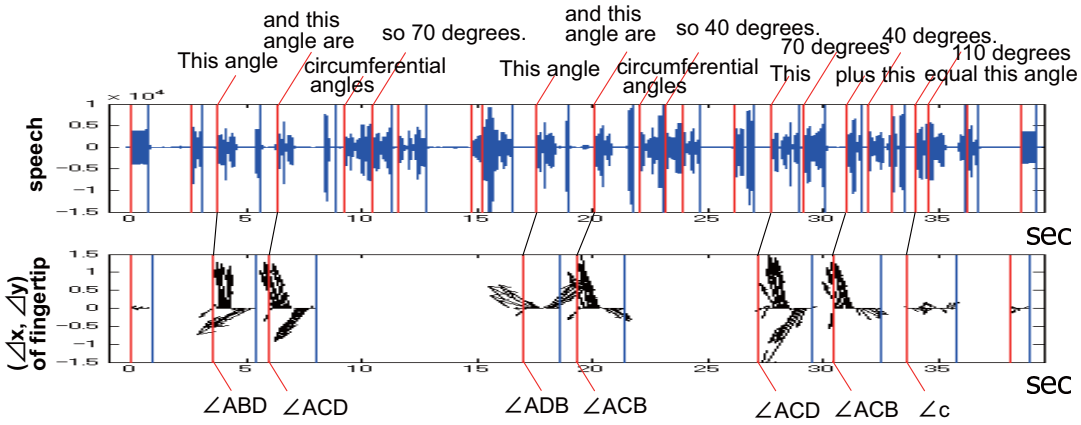

Figure 5: Speech (top) and gesture (bottom) examples. Red lines indicate starting time of utterances and gestures. Arrows in gesture graphs describe (Δx, Δy).

So we used absolute position in z-axis as a feature. We also used the first derivatives of the features, resulting in 6-dimensional features including $\Delta x, \Delta y, z, \Delta\Delta x, \Delta\Delta y, \Delta z$.

We evaluated the gesture recognition performance with 8-fold cross variation on the data recorded in Section 2 and obtained a 91.0% recognition rate with 64.7% accuracy.

4. FINDING CORRESPONDENCE BETWEEN SPEECH AND GESTURES

After individually recognizing the speech and gestures, recognition results should be aligned. Some utterances are paired with gestures.

Such time constraints as overlapping were often used [12]. Generally, the beginning times of the speech and gestures were not identical. Figure 5 shows the speech and gesture signals, and the utterances tend to begin after the corresponding gestures. Figure 6 shows a histogram of the time differences:

τ = (Starting time of utterances) - (Starting time of gestures). (2)

We use this probabilistic tendency to find the correspondence between utterances and gestures. First we express this histogram by a Gaussian distribution of τ:

$$p_d(\tau) = \frac{1}{\sqrt{2\pi}\sigma_\tau} exp\left\{-\frac{(\tau-\mu_\tau)^2}{2\sigma_\tau^2}\right\}, \quad (3)$$

where μ_τ, σ_τ^2 are the mean and the variance of time difference τ. Utterances are paired with gestures with maximal probabilities of starting time differences.

To verify the effectiveness of this method, we performed a preliminary experiment in which utterances and gestures were manually segmented *a priori*. Then each gesture was associated with an utterance including a key phrase that had maximum probability calculated from Equation 3. Key phrases included demonstratives ('here' etc.) and parts of the figure ('angle ADB', '70o, etc.). Some utterances were not associated with any gestures. The eight trials described in Section 2 were used as the test set, and μ_τ and σ_τ^2 were estimated from the data of seven trials other than the test sets. Utterances with accompanied gestures were considered 'correct' when associated with correct gestures, and those without any accompanied gestures were considered 'correct' when any gestures were associated with them. 93.8% of the utterances were correctly associated with gestures with our method. The nearest starting time strategy and longest overlapping time strategy obtained 89.7% and 83.5% association rates, respectively, and thus our method proved to work effectively.

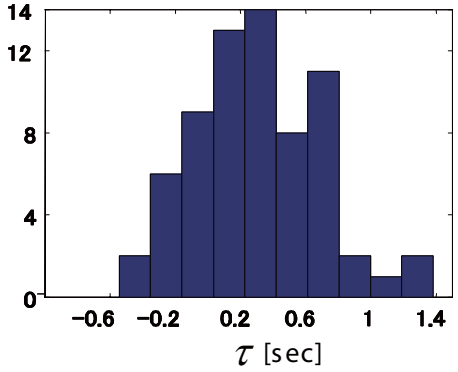

Figure 6: Histogram of differences between starting times of utterances and gestures

5. INTEGRATIVE RECOGNITION OF SPEECH AND GESTURES

Speech and gestures play complementary roles in multimodal communication. If multiple modalities are recognized simultaneously and integratively, recognition performance may improve. In this section, we investigate a method to integrate speech recognition, gesture recognition, and the correspondence between speech and gestures.

5.1 Semantic constraints between speech and gestures

To constrain the alignment between speech and gestures, we made a possibility table of the associations between an utterance and a gesture in Table 1 that defines the possible combinations.

5.2 Integrative recognition method

We adopt multimodal integration in the N-best rescoring of individual recognition results. First we obtain N-best lists as speech and gesture recognition results. Each candidate is a sequence of utterances (for speech)/pointings (for gestures). Then we calculated the combination scores for all the hypothesis pairs between speech and gestures using dynamic programming. Local scores between utterance u_i and gesture g_j in the dynamic programming are calculated as

Table 1: Possibility table of associations between utterances and gestures (Examples are excerpted)

Keyword/phrase in utterance	Examples	Possible gesture
Demonstratives	'Here'	Pointing at an angle, a segment, or a vertex
Demonstratives for angles	'This angle'	Pointing at/tracing an angle
Demonstratives for segments	'This segment'	Tracing a segment
Demonstratives for points	'This point'	Pointing at a vertex
Expressions for angles	'Angle ADB'	Pointing at/tracing a specific angle

Table 2: Recognition results by integration of multiple modalities: recognition rate [%]

Modality		Recognition rate	
		Speech	Gesture
Speech	1-best	75.0	—
	20-best	80.0	—
Gesture	1-best	—	91.0
	20-best	—	94.7
Speech & Gesture	—	**78.4**	**94.7**

follows:

$$Score(u_i, g_j) \tag{4}$$

$$= \begin{cases} \alpha L_s(u_i) + \beta L_l(g_j) + \gamma p_d(t_{u_i} - t_{g_j}), \\ \qquad\qquad M(u_i, g_j) = 1 \\ -\infty, \qquad\qquad M(u_i, g_j) = 0 \end{cases}, \tag{5}$$

where $L_s(u_i)$ and $L_l(g_j)$ are the recognition scores for u_i and g_j, respectively, $p_d(\tau)$ is the probability of the time difference defined by Equation (3), and $M(u_i, g_j)$ takes 1 or 0 as an indicator of the association possibility of u_i and g_j based on Table 1. When an utterance is associated with no gesture, it is associated with a between-pointing gesture and the time difference score is not considered. Using local score $Score(u_i, g_j)$, a hypothesis pair is globally aligned and scored. The hypothesis pair with a maximum global score is selected as the final result.

6. EXPERIMENT

We evaluated our integrative recognition method by using the eight trials described in Section 2 as the test set. We set both Ns of the N-best candidates for speech and gestures at 20 and obtained the N-best results using the recognition methods introduced in Sections 3.1 and 3.2. The same μ_τ and σ_τ^2 as Section 3.2 were used. We set α, β, and γ experimentally. Table 2 shows the result. '1-best' describes the ordinal 1-best recognition rate and '20-best' describes the rate when the best candidates were selected from the 20-best candidate lists (that is, the upper bound of the performance).

This integration method obtained 3.4- and 3.7-point recognition performance improvements for speech and gestures, respectively, and this performance is near the upper bounds.

We also evaluated the method by the identification rate of the referents. The reference words cannot identify the items in the figure, but gesture integration clarifies their ambiguities.

$$\text{Identification rate} = \frac{\text{\# utterances with correctly identified items}}{\text{\# all utterance accompanied with gestures}} \tag{6}$$

The identification rate of the integrative recognition results was 91.7%. The speech recognition results of the integrative recognition was 20.0%, and thus 71.7 points of improvement were achieved by integration.

7. CONCLUSION

In this paper, we proposed an integrative recognition method of speech accompanied with gestures. First we proposed a probability density of the starting time differences between speech and corresponding gestures to align them. Then we incorporated this probability for an integrative recognition method with which a sequence pair of utterances and gestures was scored by dynamic programming. This multimodal recognition method achieved more than three points of improvement for both speech and gesture recognition. Although so far we have only used N-best lists as intermediate expressions, other expressions with less information loss can be used, such as word graphs or HMM trellises.

8. REFERENCES

[1] S. Oviatt, R. Coulston, and R. Lundsford, "When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns," Proc. IEEE International Conference on Multimodal Interfaces (ICMI'04), 2004.

[2] S. Oviatt, R. Lunsford, and R. Coulston, "Individual Differences in Multimodal Integration Patterns: What Are They and Why Do They Exist?" CHI 2005, 2005.

[3] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody based co-analysis for continuous recognition of co-verbal gestures," Proc. ICME, 2002.

[4] M. Fukumoto, Y. Suenaga, and K. Mase, "Finger-pointer: pointing interface by image processing," Comput. Graph., Vol. 18, No. 5, pp. 633–642, 1994.

[5] R. A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," ACM Computer Graphics, Vol. 14, No. 3, pp. 262–270, 1980.

[6] P. Hui and H. M. Meng, "Joint interpretation of input speech and pen gestures for multimodal human computer interaction," Proc. INTERSPEECH-2006, pp. 1197–1200, Sept. 2006.

[7] S. Qu and J. Y. Chai, "Salience Modeling based on Non-Verbal Modalities for Spoken Language Understanding," Proc. ICMI'06, pp. 193–200, 2006.

[8] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma, "A real-time framework for natural multimodal interaction with large screen displays," Proc. ICMI 2002, Oct. 2002.

[9] J. Chai, P. Hong, M. Zhou, and Z. Prasov, "Optimization in multimodal interpretation," Proc. 42nd Annual Meeting of Association for Computational Linguistics (ACL), 2004.

[10] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," Proc. COLING 2000, 2000.

[11] M. Johnston, "Unification-based multimodal parsing," Proc. COLING-ACL'98, 1998.

[12] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration — A statistical view," IEEE Trans. Multimedia, Vol. 1, No. 4, pp. 334–341, 1999.

[13] K. Maekawa, 2003, Corpus of Spontaneous Japanese: Its design and evaluation, Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), pp. 7–12.