EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech

Madoka Miki[1], Norihide Kitaoka[1*], Chiyomi Miyajima[1], Takanori Nishino[2] and Kazuya Takeda[1]

Abstract

We propose an integrative method of recognizing gestures such as pointing, accompanying speech. Speech generated simultaneously with gestures can assist in the recognition of gestures, and since this occurs in a complementary manner, gestures can also assist in the recognition of speech. Our integrative recognition method uses a probability distribution which expresses the distribution of the time interval between the starting times of gestures and of the corresponding utterances. We evaluate the rate of improvement of the proposed integrative recognition method with a task involving the solution of a geometry problem.

1 Introduction

Multimodal interaction, where multiple modalities sometimes play complementary roles with one another, is likely to become more widespread in human-machine communication. The semantics expressed in a modality may be ambiguous, but another modality might be able to remove these ambiguities. Combining gestures and speech is a typical example of such multimodality.

When completing a task using an interface, as task difficulty increases, users often prefer multimodal interactions rather than unimodal ones, for example, when entering data in an interface system with speech and pen modalities [1]. This implies that the smooth completion of complex transactions is facilitated by multimodality, especially by the ability to select a method capable of expressing complex intentions.

In this paper, we propose a method for improving gesture and speech recognition and use a task involving the solution of a geometry problem to test it. When performing such tasks, verbal utterances are often accompanied by pointing because individual modalities are often ambiguous. For an automated system to understand such bimodal input, this kind of problem is generally divided into three sub-problems: independent recognition of speech and fingertip movements, matching up the utterances and fingertip movements, and simultaneous recognition (and understanding) of this bimodal input, taking into account both modalities. In this paper, we focus on the second and third issues, which, if successfully resolved, will result in what is known as 'modality fusion', which can be defined as the integration of the analysis of multiple modalities.

Although multiple feature streams from multiple modalities may be integrated and recognized simultaneously (using 'early integration' or 'data-level fusion') [2], as in bimodal audio-visual speech recognition, this approach is only successful when the modalities are well synchronized with each other. Therefore, it cannot be applied to the integration of speech and gestures. Thus, 'late integration' (or 'decision-level fusion') [2] is usually used, and thus all three of the sub-problems above need to be resolved.

To address the first issue of gesture recognition, methods using image processing have been proposed to recognize gestures, including fingertip movements. Head and hand positions have been tracked using video [3], fingertip

*Correspondence: kitaoka@nagoya-u.jp
[1] Department of Media Science, Nagoya University, Nagoya, Aichi Prefecture 464-8603, Japan
Full list of author information is available at the end of the article

position have been tracked using images captured by humans [4], position sensors have been used to acquire the position of a fingertip [5], and touch pens and panels have been used to interpret pointing [6,7]. In this paper, we used derivatives of position sensor data to recognize gestures. We may be able to use other methods as well to improve performance, but this is out of the scope of our current investigation.

After independently recognizing speech and gestures, correspondence must be found between them. Utterances and gestures which express identical meanings are paired. For such pairing, temporal order [6] and inclusion [8], semantic compatibility [6], and the relationship between prosodic features in speech and the speed of hand/finger movements [3] have been used. Utilizing prosodic features is an interesting approach, but extraction of F_0 features is not easy, and prosodic features include a wide range of individual variations. Thus, results using this method tend to vary widely in accuracy. Constraint by temporal order or inclusion (by overlapping the periods of modalities) is effective. However, the order constraint is relatively weak compared to the overlap constraint. On the other hand, the overlap constraint makes it difficult to determine correspondence, resulting in a lack of flexibility. We propose a soft decision method based on the statistics of the overlaps.

Finally, the information from the speech and gestures is used to construct an integrated representation. Integration/fusion methods of multimodal inputs have been well categorized, and the use of frame-based fusion has been proposed [9]. The concepts obtained from individual recognizers are put into semantic slots to represent an integrated meaning. These types of methods cannot consider temporal constraints directly, so temporal constraints are often combined, as in the method referred to above. The following schemes have also been proposed: a graph-based optimization method [10], a finite-state parsing method [11], a unification-based parsing method [12], the integration of multimodal posterior probabilities [13], and hidden Markov model-based multimodal fusion [2]. Some of these methods are able to take temporal constraints into account to some extent; however, these methods are not intended to improve single mode recognition performance as a result of the fusion.

Qu and Chai [7] proposed the use of information obtained from gestures to improve speech recognition performance. Our goal is to improve both speech and gesture recognition performance simultaneously through the modality fusion process.

In a previous study, we used the time interval between digit utterance in connected digits and accompanying finger tapping to improve digit recognition [14]. Synchronicity of speech and pen input has also been used for continuous speech recognition [15].

The rest of this paper is organized as follows. We first introduce the experimental task and explain the method of recording the multimodal inputs in Section 2. We then explain our gesture and speech recognition methods in Sections 3 and 4, respectively, and propose an integrative recognition method using multimodal time alignment in Section 5 [16]. We discuss our experimental results in Section 6 and conclude the paper in Section 7.

2 Experimental task and recording methods

An illustration of the geometry problem used to collect data for the multimodal input task is shown in Figure 1. The speech and pointing gestures of the subjects were recorded with a close-talk microphone and a 3D position sensor attached to the tip of the index finger, as shown in Figures 2 and 3, at sampling frequencies of 100 Hz and 48 kHz, respectively. Six subjects (four males and two females, all 23- to 27-year-old graduate or undergraduate students) performed a total of eight trials in a laboratory environment. Before recording, we told the subjects that they could use demonstratives such as ‘this’ and ‘here’ and point at the figure, instead of using precise explanations such as ‘angle ABC’. Subjects pushed a button to start and stop the synchronized recording. The total length of the recorded data was 249.0 s (31.1 s/trial on average).

3 Gesture recognition method

In order to recognize gestures, the automated system must be able to differentiate when subjects are pointing at items such as angles, segments, vertices, etc. from the movement of their fingertips.

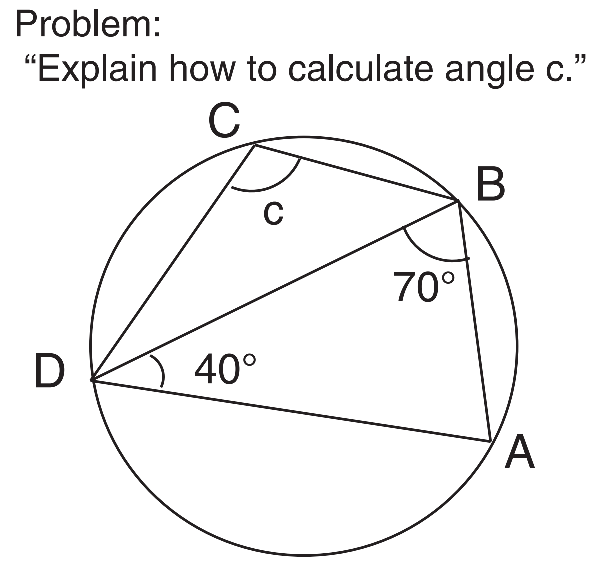

Figure 1 **Mathematical problem: calculating an angle in a quadrilateral inscribed in a circle.**

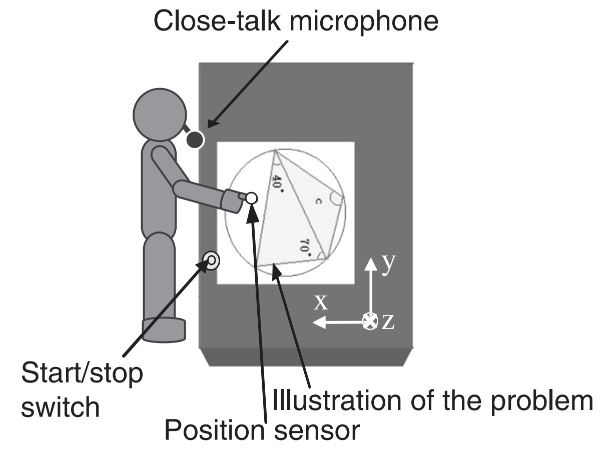

Figure 2 Recording setup.

To account for individual differences in the sizes of gestures, we used the differentials of subjects' fingertip positions in the X-Y plane as recognition features:

$$\begin{pmatrix} \Delta x[n] \\ \Delta y[n] \end{pmatrix} = \begin{pmatrix} x[n] - x[n-1] \\ y[n] - y[n-1] \end{pmatrix}, \tag{1}$$

where n indicates time, and $x[n]$ and $y[n]$ describe the fingertip position in the X-Y plane, respectively. The graph at the bottom of Figure 4 shows an example of time sequence $(\Delta x, \Delta y)^T$, indicated by arrows.

A subject's finger position in the z-axis is also important for recognizing gestures because meaningful movements can occur when a fingertip is resting on a desk, for example, so we also used absolute position in the z-axis as a feature. Additionally, we used the first derivatives of the features, resulting in six-dimensional features consisting of Δx, Δy, z, $\Delta\Delta x$, $\Delta\Delta y$, and Δz.

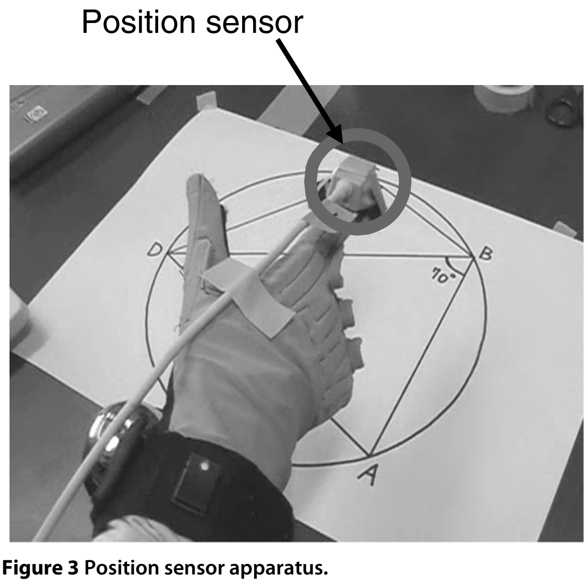

Figure 3 Position sensor apparatus.

We used three-state HMMs with a single mixture to model 21 finger movements. A total of 18 of the 21 gestures corresponded to pointing at one of the 11 segments, 4 vertices, or 3 arcs between segments in the figure shown in Figure 1. The three remaining finger movements consisted of gestures which occurred during intervals between pointing gestures, pushing the start/stop switch, and touching the desk without pointing at any of the items.

We evaluated the system's gesture recognition performance using eightfold cross validation of the data recorded in Section 2 and obtained a 91.0% correct rate and 64.7% accuracy, which were defined as:

$$\mathrm{Corr} = \frac{C_g}{N_g}, \tag{2}$$

$$\mathrm{Acc} = \frac{C_g}{R_g}, \tag{3}$$

where Corr and Acc describe correct rate and accuracy, respectively, and C_g, N_g, and R_g represent the number of correctly recognized gestures, the number of gestures included in the test data, and the number of recognized gestures, respectively.

As mentioned in our introduction, we could have adopted other features and/or methods to improve recognition performance. We understand that using HMMs with the features Δx, Δy, and z may not be the best choice for gesture recognition. We did so, however, because our proposed method involves an integration, which will be described in Section 5, and each recognition method should be kept separate from the integration[a]. Improvement of the performance of individual modality recognition rates is a subject of future work, and we believe improved individual recognition methods will increase the benefits of our integrative method.

4 Speech recognition method

We also performed speech recognition experiments using the recorded explanation utterances. The Julius decoder was used for speech recognition [17]. We used a network grammar that accepted a sequence of elements, such as the expression 'angle ADB equals angle ACB', etc. Since subjects were often explaining how to solve the problem while they were still thinking about the solution, they often used fillers and disfluencies; therefore, the grammar was set up to accept fillers between any words. No other methods were used to deal with out-of-vocabulary words. The size of vocabulary was 77 words. These words and the grammar were predefined empirically and thus they could

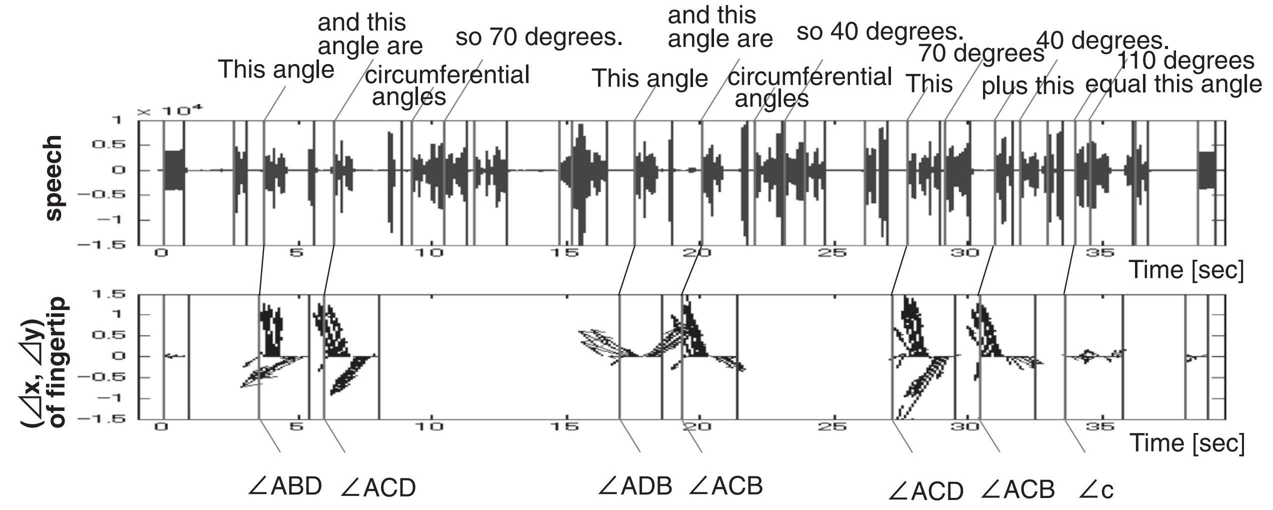

Figure 4 Examples of data for speech (top) and gestures (bottom). Red lines indicate starting times of utterances and gestures. Arrows in gesture graphs indicate $(\Delta x, \Delta y)^T$.

be used for all of the test data. Triphone HMMs were used as the acoustic models, and they were trained using the Corpus of Spontaneous Japanese (CSJ) [18], which is suitable for spontaneous speech. Each HMM had three states with output probabilities. The sampling frequency was 16 kHz, frame length and shift were 25 and 10 ms, respectively, and a 12-dimensional MFCC and its delta with delta log power were used as features. These acoustic models were also trained in advance, not using part of the test set; thus, we used the models for all of the test data. For this reason, we did not need to perform *n*-fold cross validation. We obtained a 75.0% speech recognition rate with a 66.7% accuracy.

5 Integrative recognition method

5.1 Relationship between speech and gestures

Some utterances could be easily paired with simultaneous gestures. However, speech, and the gestures which accompanied it, often did not occur simultaneously, as in the example given in Figure 4. In such cases, the utterances tended to begin after the corresponding gestures occurred. This was especially true at the beginning of the recordings. Figure 5 shows a histogram of the time differences, which was calculated as follows:

$$\tau = t_s - t_g, \tag{4}$$

where t_s and t_g indicate the starting time of an utterance and a gesture, respectively. We used this probabilistic tendency to match utterances and gestures. From the histogram, we can observe a symmetrical tendency towards decay to both sides. So first, we express this histogram using the Gaussian distribution of τ:

$$p_d(\tau) = \frac{1}{\sqrt{2\pi}\sigma_\tau} \exp\left\{-\frac{(\tau-\mu_\tau)^2}{2\sigma_\tau^2}\right\}, \tag{5}$$

where μ_τ and σ_τ^2 are the mean and the variance of time difference τ, respectively. Utterances are paired with gestures with maximal probabilities of corresponding starting time differences. We could have used the discrete distribution derived directly from the histogram, but we decided to fit a parametric distribution to the histogram instead, for the purpose of generalization[b].

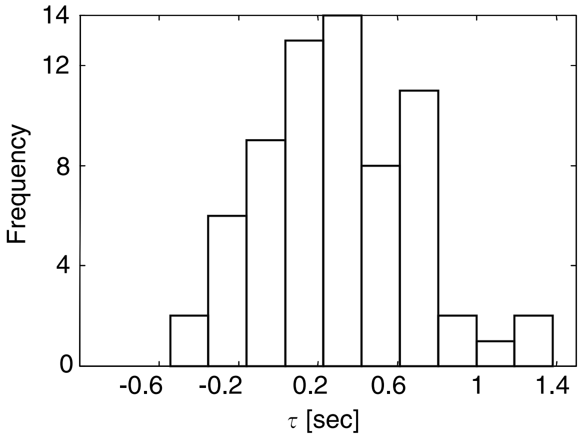

Figure 5 Histogram of differences between starting times of utterances and gestures.

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 its maximum probability calculated using Equation 5. Key phrases included demonstratives ('here', 'this', etc.) and parts of the figure ('angle ADB', '70 degrees', 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 from the seven trials, not including each test trial (that is, using eightfold cross validation). Matches were considered to be 'correct' when utterances were associated with the correct gestures, and utterances without any accompanying gestures were considered 'correct' when no gestures were associated with them. Of the utterances, 93.8% was correctly associated with gestures. The nearest matching starting time strategy and longest overlapping time strategy obtained 89.7% and 83.5% association rates, respectively, and thus, our method was proven to function effectively.

5.2 Integration algorithm

We performed multimodal integration of the N-best rescorings of individual recognition results. First, we obtained the N-best lists of both speech and gesture recognition results. Each candidate in the lists was a sequence of utterances (for speech) or fingertip positions (for gestures). Then, we obtained the combined scores for all of the speech and gesture candidate pairs using dynamic programming. Local score L between utterance u_i and gesture g_j was calculated as follows using dynamic programming:

$$L(u_i, g_j) = \begin{cases} \alpha L_s(u_i) + \beta L_g(g_j) + \gamma \log p_d(t_{s_i} - t_{g_j}), & \text{if } M(u_i, g_j) = 1, \\ -\infty, & \text{if } M(u_i, g_j) = 0 \end{cases} \tag{6}$$

where $L_s(u_i)$ and $L_g(g_j)$ are the recognition scores for u_i and g_j, respectively, $p_d(\tau)$ is the probability of the time difference as defined by Equation 5, and $M(u_i, g_j)$ takes 1 or 0 as an indicator of the possibility of an association between u_i and g_j, based on Table 1. $L_s(u_i)$ and $L_g(g_i)$ are segment log-likelihoods for u_i and g_j, respectively, obtained using the Viterbi alignments of the HMMs. These segment log-likelihoods are not normalized using word/gesture durations. When an utterance is not associated with any gesture, it is associated with an interval between gestures, and time difference score p_d is not considered. Using local score $L(u_i, g_j)$, a candidate pair is globally aligned and scored. The candidate pair with the maximum global score in all $N \times N$ pairs[c] is selected as the final result.

Table 1 Table of possible associations between utterances and gestures (examples are excerpted)

Keyword/phrase in utterance	Example utterance(s)	Possible gesture
General 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/segment	'Angle ADB'	Pointing at/tracing a specific angle
	'70 degrees'	

6 Experiment

6.1 Experimental setup

We conducted an experiment to evaluate the improvement in speech and gesture recognition using the proposed integrative recognition method. We used the eight trials described in Section 2 as the test set and obtained the N-best results using both the speech and gesture recognition methods introduced in Sections 3 and 4. Each candidate included a word sequence and the time alignment data, i.e., the start and ending time of each word, to determine the correspondence of utterances and gestures. We set both values of N (of the N-best candidates for speech and gestures) at 20. This means that the system compared a maximum of $N \times N(400)$ pairs of speech and gesture recognition candidates per trial[d]. As for gaps between corresponding utterances and gestures, we approximated the statistics using a Gaussian distribution with the same μ_τ and σ_τ^2 used in Eqn. (5) in Section 3. To allow for dynamic ranges of likelihood for speech, gestures, and time gaps between utterances and gestures, we set α, β, and γ in Equation 6 appropriately, as the result of a preliminary experiment.

6.2 Results

The results of our experiment are shown in Table 2. '1-best' describes the ordinary 1-best recognition rate, and '20-best' describes the rate when the best candidates were selected from the 20-best candidate lists (which is the upper limit of our proposed method).

The proposed integration method achieved a 3.4% point improvement in speech recognition performance and a 3.7% percentage point improvement in gesture recognition performance. The speech recognition performance of the proposed method was near the upper bounds, and its gesture recognition performance was at the upper bounds.

Table 2 Recognition results using an 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 and gesture	-	*78.4*	*94.7*

There were many speech and gesture recognition errors. By aligning the corresponding words in utterances with gestures using dynamic programming (DP), we were able to reject pairs with low DP scores. Although this strategy was effective in our proposed method, it only aligned speech and gestures in order, and thus, its rejection ability was weak. Semantically inconsistent alignments were rejected by $M(u_i, g_j) = 0$ in Equation 6. This was a strong constraint, and some incorrect alignments were rejected, but because there were so many ambiguous words among the utterances, such as 'here' and 'this', which had many possible corresponding gestures, it was not highly effective. The distribution of time differences, however, was an effective constraint of the DP path. The start times of corresponding speech and gesture pairs should not differ greatly, and the correspondences were better identified using this strategy than by the 'nearest matching' and 'longest overlapping time' strategies described in Section 5.1. The distribution in Equation 6 worked as a 'soft' path limitation, and this may be a reason why this strategy worked so well.

Overall, this is how we obtained the abovementioned improvements, but we likely could have achieved the same performance using only a simple framework based on Equation 6.

We also evaluated the proposed method using the identification rate of the referents. The items in the figure cannot be identified using only the speech from the recordings, but gesture integration clarifies the ambiguities:

$$I = \frac{C}{C_s}, \tag{7}$$

where I is the identification rate, and C and C_s are the number of utterances with correctly identified referents, and the total number of utterances accompanied with gestures, respectively. The identification rate using the integrated recognition results was 91.7%, while the identification rate using only the speech portion of the integrated recognition results was 20.0%, thus a 71.7% point improvement was achieved through integration.

7 Conclusions

In this paper, we introduced an integrative recognition method using accompanying speech to recognize gestures. First, we proposed a probability density of the differences in starting times between speech and the corresponding gestures to align the two. Then, we incorporated this probability into an integrative recognition method, which scored sequenced pairs of utterances and gestures using dynamic programming. This multimodal recognition method achieved more than 3% points of improvement in both speech and gesture recognition.

Note that our method could also possibly be used with other types of multimodalities, although currently, this method is specialized to the task which we have selected. A speaker-dependent, large-vocabulary, continuous speech recognizer could be used without any specific training, but a task-specific gesture recognizer would need to be constructed because there are no universal primitive units for gesture recognition corresponding to the phonemes and syllables used for speech recognition. The correspondence between modalities should also be defined for the task *a priori*. Even so, we believe that we can apply our method to any task which meets the following conditions: each of the modalities can be recognized using methods such as HMMs, the relationship between modalities can be described by constraint rules, and the timing difference between modalities can be described as a probability density. The larger the task becomes, the more difficult it is to construct such a framework, but once this is achieved, our proposed method can be applied. Application of this method to larger scale tasks is one of our future goals.

Although so far, we have only used N-best lists as intermediate expressions for our integrative method, other expressions with less information loss could also be used, such as word graphs or HMM trellises.

Endnotes

[a] Another reason to use HMMs is that the score obtained from an HMM is based on a probability, and thus the integration explained in Section 5 becomes theoretically correct.

[b] Of course, we can use other parametric discrete/continuous distributions, and one of them may achieve better performance, but pursuing such distributions is a task for future work.

[c] The N values of utterances and gestures can differ. In this paper, however, we used the same $N(20)$ for both values, as described in Section 6. This was decided through preliminary experiments.

[d] We used K-fold cross validation because we tested the HMM parameters for gesture recognition under an open data condition. This setting is different from that used for speech recognition, in which we prepared training and test data separately. Under both conditions, however, no data were used for both training and test data, and thus, the difference in the experimental setup for gestures and speech did not affect the results.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Media Science, Nagoya University, Nagoya, Aichi Prefecture 464-8603, Japan. [2] Department of Information Engineering, Mie University, Mie Prefecture 514-8507, Japan.

Received: 28 February 2013 Accepted: 3 January 2014
Published: 13 January 2014

References

1. S Oviatt, R Coulston, R Lundsford, When do we interact multimodally? Cognitive load and multimodal communication patterns, in *Proceedings of ICMI* (ACM, New York, 2004), pp. 129–136
2. B Dumas, B Signer, D Lalanne, Fusion in multimodal interactive systems: an HMM-Based algorithm for user-induced adaptation, in *Proceedings of 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (ACM, New York, 2012), pp. 15–24
3. S Kettebekov, M Yeasin, R Sharma, Prosody based co-analysis for continuous recognition of co-verbal gestures, in *Proceedings of ICME* (IEEE Computer Society, Washington DC, 2002), pp. 161–166
4. M Fukumoto, Y Suenaga, K Mase, Finger-pointer: pointing interface by image processing. ACM Comput. Graph. **18**(5), 633–642 (1994)
5. RA Bolt, Put-that-there: voice and gesture at the graphics interface. ACM Comput. Graph. **14**(3), 262–270 (1980)
6. P Hui, H Meng, Joint interpretation of input speech and pen gestures for multimodal human computer interaction, in *INTERSPEECH* (ISCA, Pittsburgh, 2006), pp. 1197–1200
7. S Qu, JY Chai, Salience modeling based on non-verbal modalities for spoken language understanding, in *Proceedings of ICMI* (ACM, New York, 2006), pp. 193–200
8. N Krahnstoever, S Kettebekov, M Yeasin, R Sharma, A real-time framework for natural multimodal interaction with large screen displays, in *Proceedings of ICMI* (IEEE, Piscataway, 2002), pp. 349–354
9. D Lalanne, L Nigay, P Palanque, P Robinson, J Vanderdonckt, J Ladry, Fusion engines for multimodal input: a survey, in *Proceedings of ICMI-MLMI* (ACM, New York, 2009), pp. 153–160
10. J Chai, P Hong, M Zhou, Z Prasov, Optimization in multimodal interpretation, in *Proceedings of ACL* (Association for Computational Linguistics, Stroudsburg, 2004), pp. 1–8
11. M Johnston, Finite-state multimodal parsing and understanding, in *Proceedings of COLING* (Association for Computational Linguistics, Stroudsburg, 2000), pp. 369–375
12. M Johnston, Unification-based multimodal parsing, in *Proceedings of COLING-ACL* (Association for Computational Linguistics, Stroudsburg, 1998), pp. 624–630
13. L Wu, L Oviatt, PR Cohen, Multimodal integration - a statistical view. Trans. Multimedia **1**(4), 334–341 (1999)
14. H Ban, C Miyajima, K Itou, K Takeda, F Itakura, Speech recognition using synchronization between speech and finger tapping, in *Proceedings of ICSLP* (ISCA, Pittsburgh, 2004), pp. 943–946
15. K Shinoda, Y Watanabe, K Iwata, Y Liang, R Nakagawa, S Furui, Semi-synchronous speech and pen input for mobile user interfaces. Speech Commun. **53**(3), 283–291 (2011)
16. M Miki, C Miyajima, T Nishino, N Kitaoka, K Takeda, An integrative recognition method for speech and gestures, in *Proceedings of ICMI* (ACM, New York, 2008), pp. 93–96
17. A Lee, T Kawahara, K Shikano, Julius — an open source real-time large vocabulary recognition engine, in *Proceedings of EUROSPEECH* (ISCA, Aalborg, 2001), pp. 1691–1694
18. K Maekawa, Corpus of spontaneous Japanese: its design and evaluation, in *Proceedings of SSPR* (ISCA and IEEE, Tokyo, 2003), pp. 7–12

doi:10.1186/1687-4722-2014-2
Cite this article as: Miki *et al.*: Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:2.