Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Gestural Cue Analysis in Automated Semantic Miscommunication Annotation

Masashi Inoue · Mitsunori Ogihara · Ryoko Hanada · Nobuhiro Furuyama

Received: date / Accepted: date

Abstract Automated annotation of conversational video using semantic labels is a challenging topic. We investigated miscommunication as an example of a higher-level semantic concept in conversations. Miscommunication is an important obstacle in solving psychological problems such as those in psychotherapeutic interviews. Although miscommunications are often obvious to the speakers as well as observers, it is difficult for machines to detect them from low-level features. This difficulty is due to the lack of understanding on which cues contribute to miscommunication. Miscommunications are mainly associated with spoken content, though non-verbal elements may play a role. Among various non-verbal features, we investigate gestural cues. Various features are taken from gesture data, and both simple and complex classifiers are constructed using machine learning. The experimental results suggest that there is no single gestural feature that can predict or explain the occurrence of semantic miscommunication.

Keywords Semantic indexing · Gesture · Psychotherapy · Face-to-face

M. Inoue
Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
Collaborative Research Unit, National Institute of Informatics, Tokyo, Japan

M. Ogihara
Department of Computer Science / Center for Computational Science, The University of Miami, Miami, USA

S. Author
Department of Clinical Psychology and Training, Kyoto University of Education, Kyoto, Japan

S. Author
Information and Society Research Division, National Institute of Informatics, Tokyo, Japan

1 Introduction

1.1 Semantic Annotation of Conversational Video

Conversations used to be recorded either as text transcripts or speech sounds for archiving and post-analysis purposes. Todays, they are stored as video data since recording devices are readily available. Although the collection of video data is easy, their analysis still relies on the manual inspection. Video data usually lack an essential piece of information, the semantic annotations. As the amount of recorded conversations increases, there is a growing needs for computationally assigning semantic annotation to video data. Among semantic information, spoken word extraction might be relatively straightforward as there is a long tradition of automatic speech recognition. With videos, the challenge is in annotation based on non-verbal behavior. We are particularly interested in the potential of gestural cues in annotating semantic information. We used actual psychotherapy dialog data where miscommunication must be avoided, but which often occur. That is, we investigated the possibility of identifying a miscommunication label based on low-level gestural signals.

1.2 Miscommunication and Gesture in Psychotherapy

One of the basic tasks of psychotherapists is understanding their clients' mental problems and the contexts in which these problems arose. However, miscommunication often occurs during the interview sessions, and therapists need to engage with the clients' problem [2]. A miscommunication segment contains an utterance that suggests that the receiver does not understand the message. Typical examples are questions asking for clarification, e.g., "What do you mean by ...?" or "Could you explain it?". Therefore, even if the utterance of the message sender is ambiguous or irrational, but the receiver can understand it, such an interaction is not categorized as a miscommunication segment. We did not differentiate how miscommunications are brought about and what the outcomes are for this research.

Miscommunications often lead to more time dedicated to clarify the message, as above examples suggests; sometimes they even develop into some conflicts between therapists and client, making a solution of client's problem difficult. Psychotherapists can intentionally introduce miscommunication caused by their statements as a means of intervention, but they must avoid unintended miscommunication. Miscommunication caused by the statements by clients occurs frequently because of the following two reasons. First, compared to other types of conversation, therapeutic conversations do not have a predefined topic or standard interaction format. Second, since the clients usually have thought about their problems for a long time and thus those problems are highly evident to them, they cannot understand why the therapists fail to immediately recognize the problems.

Mining semantic knowledge, problematic events in particular, from conversations using computers is a relatively new research topic. There have been observational studies on this topic. For example, miscommunication patterns in survey interviews have been studied [9]. The main concern in this study is the awkwardness resulting from adherence to strictly pre-defined formats for questions. Also, emotional conflicts in face-to-face conversation have been qualitatively categorized [8]. A computational example is the analysis of telephone conversations at a contact center of a rent-a-car

business [10]. The goal was to identify successful conversational state transitions rather than failures for booking as many car rentals as possible.

Another aspect that has not been examined computationally is the role of gestural cues. Although gestures, or hand gestures in particular, have been studied qualitatively in professional conversations, such as in medical counseling [6], there have been few computational investigations on the relationship between particular conversational events and gestures. One exception was the attempt to computationally detect deception from non-verbal cues including gestures [1]. In this paper we are making an initial step toward a data-driven understanding of high-level semantic events and gestures in psychotherapeutic interviews.

Below, we explain how we represent conversations in a machine-readable format, the methodology we use to detect miscommunications from data, the properties of the data and features to be used, and the experimental results.

2 Methodology

2.1 Data representation

The conversation data used in our analysis were captured as video files depicting two speakers sitting facing each other. The nth video has a length of T_n. All videos were segmented into S_n time slots of size W ($S_n = T_n/W$) and they were treated as discrete time slots. After a manual inspection of the videos, we assigned two types of labels to each video segment of the conversation: gestures and miscommunication.

Gestures are specific movements of hands and arms. We investigated two classes of gestures: *communicative* (i.e., conveying messages) and *non-communicative*. The first class consists of *iconic*, *metaphoric*, and *deictic* gestures, and the second of *beat* gestures and *adapters*. McNeill defined four of these gesture types [7]. *Iconic* gestures are those that bear a close formal relationship to the semantic content of speech. *Metaphoric* gestures are like iconic gestures in that they are pictorial, but the pictorial content presents an abstract idea rather than a concrete object or event. *Deictic* (i.e., pointing) gestures indicate objects and events in the concrete world or abstract space. *Beat* gestures are those that look like beats in musical timing. *Adapters* are self-touching hand movements. Freedman [4] suggests that adapters represent mental status such as conflicts between speakers. McNeill does not consider them to be gestures; however, since conflicts are of interest in psychotherapy, we decided to include this additional gesture type. It should be noted that these gesture types are not exclusive. For example, certain hand movements can be understood as both iconic and deictic [1].

In each segment, we identify every gesture, regardless of which hand the speaker uses, that falls into one of the two classes, and measure its duration in that segment. Sometimes hands seamlessly transit from one gesture type to another. In such cases, instead of trying to divide the multi-gesture sequence into sub-sequences with unique gestures, we label the entire gesture sequence as the most significant gesture type. This resulted in generating longer gesture durations as data than isolated gestures.

The second label is the occurrence of a miscommunication. Identification of miscommunications is not based on reports from participants but from video observations. First, transcripts are created from videos. Next, points at which any word or phrase

[1] http://mcneilllab.uchicago.edu/topics/annotation.html

Table 1 Gestural features calculated on each segment

Feature Index	Gestural Feature	Interpretation
x_1	Frequency in current segment	Degree of gestural activity
x_2	Frequency in previous segment	”
x_3	Frequency in next segment	”
x_4	Frequency difference from previous segment	Degree of change in gestural activity
x_5	Frequency difference in next segment	”
x_6	Duration (Mean)	Degree of complexity of gesture
x_7	Duration (Maximum)	”
x_8	Duration (Minimum)	”
x_9	Mean interval	Degree of gestural continuity

that may indicate the existence of a miscommunication are listed. Then, these points of suspicion are checked against the original video taking into account speech sound and other modalities. If a check confirms that an interaction contains a point of miscommunication, the starting time of the interaction is considered to be the time point of miscommunication.

2.2 Feature set

For both communicative and non-communicative gestures produced by clients, we derive the following features from the basic gesture code data defined in 2.1,

1. The gesture frequencies on, before, and after the target time slot are respectively denoted as x_1, x_2, and x_3. Gesture frequencies were calculated at gesture starting points: how many times gesture were initiated in a given window of size W. We computed the gesture frequencies at the current segment, at W seconds in the past and at W seconds in the future.
2. The differences in gesture frequencies between the sth and $(s-1)$th or $(s+1)$th segments were also calculated, respectively denoted as x_4 and x_5.
3. The mean, maximum, and minimum duration of gestures in each segment, respectively denoted as x_6, x_7, and x_8.
4. The mean interval of a speaker's gesturing, x_9.

The resulting representation of gestural cues is $X = \langle \mathbf{x}_1, ..., \mathbf{x}_S \rangle$ where $\mathbf{x} = (x_1, ..., x_9)$. Among these feature values, x_1, x_2, x_3 are nonnegative integers, x_4, x_5 are integers, and x_6, x_7, x_8, x_9 are nonnegative real numbers. For the window size, W, we used 5 and 50 seconds. The two segmentation bin sizes correspond to the short-term and long-term dependencies between gestural signals and semantic miscommunication. The 5-second window can be used to find how gestures are used while miscommunication is occurring. In contrast, the 50-second window captures the overall gestural trend that induces miscommunication. All the above-mentioned features are summarized in Table 1. The left column shows the features and the right column states what these features are expected to measure. Note that the goal of our study is not building accurate classifiers but finding useful cues. Therefore, we limited the features to those that seem noticeable by humans; we did not examine long-term dependencies or complicated interactions.

After we divide a dialogue into segments using bin size, we assign a binary label $y \in \{0, 1\}$ to each segment based on whether it contains miscommunication. That is,

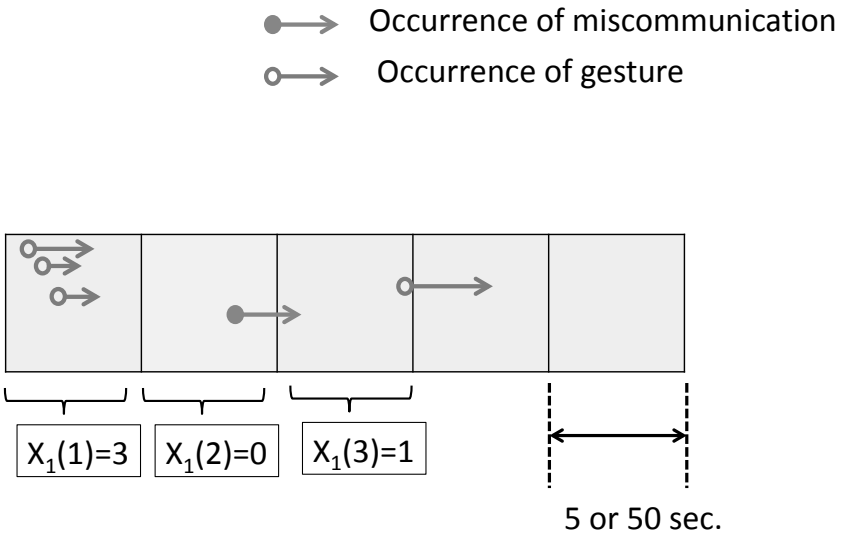

Fig. 1 Schematic illustration of segmentation and coding of segments for different window sizes. Root circles indicate starting point of gestures and miscommunications. Arrows represent duration of gestures and miscommunications.

as shown in Figure 1, occurrences of multiple miscommunications in a segment are ignored. Because the number of segments is dependent on the bin size, and the degree of class bias changes, different segmentation results in different task difficulties. Feature values are derived by checking gesture starting points in each segment. The lower part of Figure 1 shows the extraction process of x_1 feature values for each segment in this example.

The gestural features calculated for both communicative $\mathbf{x}^c$ and non-communicative $\mathbf{x}^{nc}$ categories and miscommunication labels are combined. We obtain the final representation of a conversation in the following form: $\langle\{y_1, \mathbf{x}^c(1), \mathbf{x}^{nc}(1)\}, ..., \{y_S, \mathbf{x}^c(S), \mathbf{x}^{nc}(S)\}\rangle$.

2.3 Classifier

We train binary classifiers to assess if there is a cue that can predict whether a time segment contains miscommunications. The first classification method we use is the following linear discriminant analysis (LDA) [5], which is often used as a good baseline

Table 2 Overview of datasets

Dataset	Duration (min.sec)	Therapist (Experience)	Client
1-(1)	24.17		
1-(2)	25.43	Female (Expert)	Female
1-(3)	9.07		
2-(1)	12.48		
2-(2)	21.41	Female (Intermediate)	Male
2-(3)	40.58		
3-(1)	17.02		
3-(2)	26.43	Female (Beginner)	Male
3-(3)	22.41		

classifier:

$$\delta_k(x) = x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad (1)$$
$$\text{with } y = \arg\max_k \delta_k(x)$$

where k represents a class (miscommunication or smooth communication in our case) index, Σ the covariance matrix of observations, μ the mean of observations, and π the prior probability of the class. For π, we can use either uniform prior (i.e., $\pi_{+1} = \pi_{-1} = 0.5$) or empirically estimated prior (i.e., $\hat{\pi}_k = N_k/N$ where N_k is the number of class-k observations). The second classifier is support vector machine (SVM), which can generate more complex classification boundaries. We use the LibSVM implementation [3] with radial-basis functions (RBFs) as its kernel.

3 Experiment

3.1 Basic data statistics

We prepared three conversational datasets each consisting of video files with gesture and miscommunication codings. The therapists, clients, and topics varied. Each dataset were recorded on different dates and consisted of three interview sessions between a psychotherapist and client. The number of sessions in a day and the length of each session could be controlled by the participants. The problems they discussed were actual problems and not role-plays. The properties of the datasets are listed in Table 2. All participants used the Japanese as the language for communication. The participants consented to the use of the video files for research purposes.

3.2 Experimental settings

We conducted a leave-one-out cross-validation to assess the best achievable classification accuracy for each feature.

First, we segmented the entire conversation into either 5- or 50-second segments as we extracted feature values, i.e., $\mathbf{s} = s(1), ..., s(t), ..., s(S)$. Among these S segments, we took $s(t)$ out and trained a classifier using a gesture feature that belonged to the remaining $S-1$ time slots. Then, we classified the $s(t)$ segment into miscommunication or smooth communication classes. Some features require earlier or later time segments

to be calculated. For the boundary conditions where there is no earlier or later time segment, we simply skipped the classification.

When S is large, most segments do not contain miscommunications. That is, two classes are extremely biased. In that case, a reasonable baseline classification rule judges all segments as smooth conversation segments. However, even if the machine is successful in terms of accuracy, that baseline classifier does not offer any new information to practitioners or system designers. Therefore, we evaluated the experimental results using precision and recall measures. Precision represents how many were actual miscommunications out of all the events the machine determined to be miscommunications, and recall presents the fraction of the miscommunications that are successfully identified. There are sometimes trade-offs in achieving high scores between these two measures, so we needed a score that reflects both aspects. We used the F-measure, which is calculated using precision value p and recall value r as follows: $2\frac{pr}{p+r}$. We generally do not know which of the two are more important; therefore, we did not assign any weights to either precision or recall.

Before going into the details of our experiment, we compared different configurations of classifiers. Because the number of smooth segments is far larger than miscommunication segments, we expect calibration is needed to adjust to the data imbalance between the two classes. For LDAs, the empirical weighting by means of class priors as explained in 2.3 can be used. For SVMs, one-class SVMs that identify a single class rather than discriminate two classes can be introduced. Through preliminary experiments, we found that empirically weighted LDAs performed better than the uniform prior models, and one-class SVMs performed better than standard SVMs. Based on these results, we show the results for empirically weighted LDAs and one-class SVMs only.

3.3 Experimental results

We summarized the classification results in terms of F-measures in two tables. Table 3 corresponds to the smaller window size and Table 4 corresponds to the larger window size. When classifiers did not identify any segments as miscommunication, when classifiers regarded all segments are smooth and did not assign any positive labels, we left the entry blank (-). The best F-scores and corresponding classifiers in each session are shown in bold face. The gesture types, either communicative (c) or non-communicative (nc), are associated with classifiers.

By comparing the F-scores in the two tables, we can see that the scores are quite low in Table 3 and adequate in Table 4. This implies that there may not be any relationship between the occurrence of miscommunications and immediate gesturing at that time. In contrast, there might be some relationships between long-term gesture use and the emergence of miscommunications. However, the relationships between gestural cues and miscommunications are not easy to understand. When both tables are viewed column-wise, we can see that there is no single gestural feature consistently marked with the highest F-scores; rather, the useful feature is data session dependent. The gesture frequency of current segment x_2 could be somewhat more useful than others; however, what we can assume is only that some features, including gesture frequencies in previous and subsequent segments (x_1 and x_3), mean and minimum values of gesture duration (x_6 and x_8), are not good candidate cues for annotating miscommunications. When the tables are viewed row-wise, there are no consistently strong classifier and

Table 3 F-scores for clients' gestures (short-term, 5 sec.-window)

Dataset	Classifier	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1-(1)	**LDA(c)**	0.04	**0.05**	0.02	0.03	-	0.03	0.03	0.03	0.04
	LDA(nc)	0.04	0.04	0.04	-	0.04	0.04	0.03	0.04	0.04
	SVM(c)	-	-	-	-	-	-	-	-	-
	SVM(nc)	-	-	-	-	-	-	-	-	-
1-(2)	LDA(c)	0.03	0.03	0.04	-	0.04	0.03	0.02	0.03	0.03
	LDA(nc)	0.03	0.03	-	**0.07**	**0.07**	0.03	0.03	0.03	0.03
	SVM(c)	-	-	-	-	-	-	-	-	-
	SVM(nc)	-	-	-	-	-	-	-	-	-
1-(3)	LDA(c)	-	-	0.14	0.07	-	0.17	0.14	0.17	-
	LDA(nc)	0.15	0.15	0.09	-	0.09	0.15	**0.18**	0.15	0.17
	SVM(c)	-	-	-	-	-	-	-	-	-
	SVM(nc)	-	-	-	-	-	0.13	0.13	0.15	-
2-(1)	**LDA(c)**	0.08	**0.16**	0.09	-	0.13	-	0.13	-	0.05
	LDA(nc)	0.13	0.07	0.09	-	-	0.08	0.11	0.07	0.08
	SVM(c)	-	-	-	-	-	0.06	0.04	0.06	-
	SVM(nc)	-	-	-	-	-	-	-	-	-
2-(2)	**LDA(c)**	0.04	-	0.03	-	**0.05**	0.04	0.04	0.02	0.02
	LDA(nc)	0.03	-	0.03	-	0.03	0.03	0.03	0.04	0.04
	SVM(c)	-	-	-	-	-	0.02	-	0.03	-
	SVM(nc)	-	-	-	-	-	-	-	-	-
2-(3)	**LDA(c)**	-	0.06	0.03	-	**0.07**	0.05	0.06	0.06	-
	LDA(nc)	0.06	0.06	0.02	0.10	0.02	0.06	0.06	0.05	**0.07**
	SVM(c)	-	-	-	-	-	0.06	0.06	0.06	-
	SVM(nc)	-	-	-	-	-	-	-	-	-
3-(1)	LDA(c)	0.08	0.03	0.07	0.04	-	0.09	0.09	0.06	0.08
	LDA(nc)	0.09	0.08	0.05	**0.11**	0.04	0.06	0.07	0.06	0.09
	SVM(c)	-	-	-	0.08	0.07	-	-	-	-
	SVM(nc)	-	-	-	0.07	0.08	-	-	-	-
3-(2)	**LDA(c)**	0.07	**0.11**	0.05	0.05	0.05	0.08	-	0.09	0.05
	LDA(nc)	0.06	0.03	0.06	0.02	0.06	0.06	0.06	0.03	0.06
	SVM(c)	-	-	-	-	-	-	-	-	-
	SVM(nc)	-	-	-	-	-	0.06	0.05	0.05	-
3-(3)	LDA(c)	0.07	0.07	0.04	0.09	0.03	0.07	0.07	0.06	0.06
	LDA(nc)	0.03	**0.11**	-	0.08	0.04	0.08	0.07	0.07	0.08
	SVM(c)	-	-	-	-	-	-	-	-	-
	SVM(nc)	-	-	-	-	-	0.07	0.04	0.07	-

gesture type combinations, although in Table 3, LDAs always outperformed SVMs that did not produce positive outputs. This result indicates that the even complex decision boundaries produced by SVMs cannot explain the relationships between gestures and semantic miscommunications. Also, it should be noted that the gesture type, either communicative or non-communicative, did not directly relate to miscommunication.

4 Discussion

We have compared many features in terms of the predictability of miscommunication. There is another important criterion to be considered: usability. Different features appeal differently to therapists. The first distinction is between clients' gestures and those of therapists themselves. We considered only clients' gestures because we are often not fully aware of our own behavior, and clients' gestures are considered more useful for therapists. The next difference is the temporal range that a therapist has

Table 4 F-scores for clients' gestures (long-term, 50 sec.-window)

Dataset	Classifier	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1-(1)	LDA(c)	0.24	0.24	0.12	0.32	0.40	0.29	0.11	0.24	0.33
	LDA(nc)	0.30	0.36	0.40	0.13	0.27	0.25	0.26	0.35	0.40
	SVM(c)	0.35	0.40	0.11	0.25	0.29	0.21	**0.62**	0.20	-
	SVM(nc)	-	-	-	0.25	0.37	0.12	-	-	0.21
1-(2)	LDA(c)	0.27	0.25	-	0.10	0.13	0.43	0.17	0.21	0.32
	LDA(nc)	0.24	0.35	0.12	0.32	0.11	0.17	0.33	0.08	0.19
	SVM(c)	0.32	0.11	0.21	0.11	0.43	0.12	0.33	0.25	**0.50**
	SVM(nc)	-	-	-	0.08	0.16	0.32	-	0.29	0.35
1-(3)	LDA(c)	0.60	0.33	0.40	0.50	0.25	0.60	0.60	0.60	0.50
	LDA(nc)	0.29	**0.67**	0.57	0.40	0.29	0.36	0.40	0.40	0.25
	SVM(c)	-	-	-	0.50	0.57	0.25	-	0.50	-
	SVM(nc)	-	0.33	-	0.33	-	0.57	-	0.33	-
2-(1)	LDA(c)	-	0.40	0.60	0.18	0.40	0.18	0.15	0.17	0.20
	LDA(nc)	0.55	0.18	0.55	0.36	0.40	-	-	0.31	0.20
	SVM(c)	0.25	**0.67**	0.29	0.20	0.44	0.22	0.29	0.44	-
	SVM(nc)	0.29	0.22	0.25	-	0.33	-	-	-	-
2-(2)	**LDA(c)**	0.23	0.37	0.31	**0.47**	0.27	0.17	0.26	0.31	0.11
	LDA(nc)	0.35	0.37	0.27	0.29	0.13	0.35	0.17	0.33	0.30
	SVM(c)	0.15	0.37	0.14	0.32	0.22	-	0.20	0.35	-
	SVM(nc)	0.21	0.40	0.30	-	-	-	0.20	0.36	-
2-(3)	LDA(c)	0.07	0.24	0.37	0.36	0.19	0.36	0.33	0.43	0.28
	LDA(nc)	0.29	0.34	0.13	0.29	0.23	0.07	0.29	0.33	0.21
	SVM(c)	0.44	**0.47**	0.24	0.22	0.38	0.33	0.24	0.23	0.39
	SVM(nc)	0.42	0.31	0.26	0.33	0.43	0.13	0.34	0.35	0.07
3-(1)	**LDA(c)**	0.32	0.33	0.22	0.33	0.27	0.50	**0.63**	0.42	0.50
	LDA(nc)	0.40	0.15	0.18	0.17	0.50	0.31	0.31	0.43	0.53
	SVM(c)	0.31	0.59	0.31	0.22	0.40	0.36	0.31	0.13	-
	SVM(nc)	0.32	0.57	0.42	0.50	0.17	0.20	0.40	0.14	-
3-(2)	LDA(c)	0.29	0.44	0.36	0.32	0.08	0.30	0.33	0.27	0.28
	LDA(nc)	0.38	0.40	**0.55**	-	0.19	0.32	-	0.40	0.20
	SVM(c)	0.41	0.42	0.32	0.30	0.29	0.29	0.48	0.37	-
	SVM(nc)	0.17	0.24	0.42	0.30	0.35	0.10	0.32	0.24	0.26
3-(3)	LDA(c)	0.50	0.43	0.20	0.52	0.35	0.55	0.45	0.11	0.25
	LDA(nc)	0.48	0.33	0.54	0.42	0.32	0.48	0.45	0.54	0.40
	SVM(c)	0.30	0.44	0.38	0.13	0.40	0.44	0.33	0.26	0.25
	SVM(nc)	0.52	**0.63**	0.52	0.41	-	0.35	0.17	0.29	0.44

to observe to detect the saliency of gestural features. If therapists have to find the changes that occur in a long time range, say 50 seconds in our experiment, it would force a more cognitive load on the therapists than detecting changes in 5 seconds. Furthermore, if the segment is detected as potentially including miscommunication, the therapist may not fully utilize that information because it is not quite clear where in the long time range the problematic point is. Therefore, even though machines can easily annotate miscommunication in a long time range, the information might not be semantically practical. In addition, we expect that humans can detect relative saliency such as the increase and decrease in frequency better than absolute saliency such as the average frequency of gestures. However, the above-mentioned categorization of usability is hypothetical and requires verification.

5 Conclusions

We tested the automatic classification of dialogue segments into smooth communications and miscommunications based on gestural cues taken from psychotherapeutic interview sessions. The classifiers were trained on actual dialogue data. This process clarified which gesture cues were useful in predicting miscommunications. Experimental results suggest that we could not find any distinct gestural feature that serves as a useful cue for automatically annotating the occurrences of miscommunications consistently among different data sessions. The distinction between two types of gestures, communicative and non-communicative, was not helpful in identifying miscommunications.

Although we could not find strong gestural cues that are tightly connected with semantic miscommunication, we will study detailed gesture types, gesture sub-units, handedness, or gesture strength, which could not be investigated with the current experimental setting. Also, the size of the dataset should be enlarged by adding more dialogue sessions for generalizability. In addition, since many of miscommunications are triggered by verbal contents, we can study the relationships between gestures and speech types. Categorizing miscommunications into two types, those that could be and those that could not be identified from gestures, might be an interesting next step.

Acknowledgments

We thank for the discussion with m-project members especially Kunio Tanabe and Tomoko Matsui for commenting on an earlier version of this paper. This research was partially supported by the Grant-in-Aid for Scientific Research 19530620, 21500266, the National Science Foundations under Grant CCF-0958490 and the National Institute of Health under Grant 1-RC2-HG005668-01, and the Function and Induction Research Project, Transdisciplinary Research Integration Center of the Research Organization of Information and Systems.

References

1. Burgoon, J., Adkins, M., Kruse, J., Jensen, M.L., Meservy, T., Twitchell, D.P., Deokar, A., Nunamaker, J.F., Lu, S., Tsechpenakis, G., Metaxas, D.N., Younger, R.E.: An approach for intent identification by building on deception detection. Hawaii International Conference on System Sciences **1**, 21a (2005)
2. Buttny, R.: Talking Problems. State University of New York Press (2004)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
4. Freedman, N.: Hands, word and mind: On the structuralization of body movements during discourse and the capacity for verbal representation, pp. 219–235. Plenum Press, New York (1977)
5. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
6. Heath, C.: Body Movement and Speech in Medical Interaction. Cambridge University Press, Cambridge (1986)
7. McNeill, D.: Hand and mind. The University of Chicago Press, Chicago (1992)
8. Mortensen, C.D.: Human conflict. Rowman & Littlefield Publishers (2006)
9. Suchman, L., Jordan, B.: Ineractional troubles in face-to-face survey interviews. Journal of the American Statistical Association **85**(409), 232–241 (1990)

10. Takeuchi, H., Subramaniam, L.V., Nasukawa, T., Roy, S.: Getting insights from the voices of customers: Conversation mining at a contact center. Information Sciences **179**(11), 1584 – 1591 (2009)