NIH Public Access
Author Manuscript
Multimed Tools Appl. Author manuscript; available in PMC 2013 November 01.

Published in final edited form as:
Multimed Tools Appl. 2012 November 1; 61(1): 7–20. doi:10.1007/s11042-010-0701-1.

Gestural cue analysis in automated semantic miscommunication annotation

Masashi Inoue
Collaborative Research Unit, National Institute of Informatics, Tokyo, Japan

Mitsunori Ogihara
Department of Computer Science/Center for Computational Science, The University of Miami, Miami, FL, USA ogihara@cs.miami.edu

Ryoko Hanada
Graduate School of Clinical Psychology/Center for Clinical Psychology and Education, Kyoto University of Education, Kyoto, Japan hanada@kyokyo-u.ac.jp

Nobuhiro Furuyama
Information and Society Research Division, National Institute of Informatics, Tokyo, Japan furuyama@nii.ac.jp

Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Tokyo, Japan

Abstract

The automated annotation of conversational video by semantic miscommunication labels is a challenging topic. Although miscommunications are often obvious to the speakers as well as the observers, it is difficult for machines to detect them from the low-level features. We investigate the utility of gestural cues in this paper among various non-verbal features. Compared with gesture recognition tasks in human-computer interaction, this process is difficult due to the lack of understanding on which cues contribute to miscommunications and the implicitness of gestures. Nine simple gestural 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 in our setting.

Keywords

Semantic indexing; Gesture; Psychotherapy; Face-to-face

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. Nowadays, 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 manual inspection. Video data usually lacks an essential piece of information, the semantic annotations. As the amount of recorded conversations increases, there is a growing

© Springer Science+Business Media, LLC 2011

M. Inoue Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan mi@yz.yamagata-u.ac.jp.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

need 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 low-level gestural signals in annotating semantic information, miscommunication. Miscommunication is an important obstacle in solving psychological problems such as those in psychotherapeutic interviews. We investigated the possibility of identifying a miscommunication in actual psychotherapy conversation data where miscommunication must be avoided, but often occurs.

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 arise. Therapists need to engage the clients' problem. However, miscommunication often occurs during the interview sessions because the participants' versions of subjective events [4] must be interpreted in social context. 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 the above examples suggests; sometimes they even develop into conflicts between therapists and clients, making finding a solution to a 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 clients' statements frequently occurs for 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 their problems.

Researches on finding problematic events in conversations have involved human observation and subjective interpretation. For example, miscommunication patterns in survey interviews have been studied based on observation [17]. The main concern in this study is the awkwardness resulting from the adherence to strictly pre-defined formats for questions. An example is that when a listener addresses that he or she understood what was asked before the completion of the question, the interviewer has to continue the sentence because it is a rule. Also, emotional conflicts in face-to-face conversations have been qualitatively categorized [14].

Finding problematic events in conversations using computers is a relatively new research topic. The difficulty is that problematic events are semantic and not easily detectable using computers that cannot mimic human subjective judgments. However, if the conversation is somewhat structured, a computational approach could be usable. An example is the analysis of telephone conversations at the contact center of a rent-a-car business [18]. The goal was to computationally identify the successful conversational state transitions for booking as many car rentals as possible.

Another aspect that has not been computationally examined is the role of gestural cues. In the field of human-computer interaction, the use of gestures as the means of natural input has been investigated [15]. In addition to hand movments, hand shapes are now

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

automatically classified without using specialized sensors [6]. Many of the researches were on understanding explicit gestures performed by humans to be used as alternative (i.e., non-textual) inputs to computers. A typical example is the use of gestures as commands in an intelligent environment [16]. In these researches, however, gestures are assumed clearly expressed. In real-world conversations, gestures are subtle and their meanings are implicit. We aim at bridging the gap between researches on multimedia methodologies and a complex conversational data domain.

Gestures, or hand gestures in particular, have been qualitatively studied in professional conversations such as in medical counseling. For example, it is observed that when a patient coordinates his head gestures with the doctor's hand gestures when the effect of treatment is explained [9]. Gestures are suggested to have certain relationships with moods in conversations [2]. However, there have been only a 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 [3].

We are taking the initial step towards a data-driven understanding of high-level semantic events and gestures in psychotherapeutic interviews in this paper. In the following sections, 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 the features used, and our experimental results.

2 Methodology

2.1 Data representation

The conversation data used in our analysis was captured as video files depicting two speakers sitting facing each other. Each video has a length of T. All the videos were segmented into S segments of size W ($S = T/W$) 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. That is, there was no automatic audio-visual feature extraction involved.

Gestures are specific movements of the 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. The second class consists of *beat* gestures and *adapters*. McNeill [11] defined four of these gesture types. *Iconic* gestures are those that are pictorial and bear a close formal relationship to the semantic and concrete 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 [7] suggests that adapters represent the mental status such as the 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 [12].

In psychotherapy, nonverbal cues were tested to see whether they can be used to predict a counselor's expertise, trustworthiness and attractiveness. Among nonverbal cues, the presence or absence of gestures and their expressiveness were used as the gestural features. The result was negative; nonverbal cues were not effective in predicting the counselors' qualifications. Considering this outcome, we speculate that the analysis should be more detailed in capturing semantic content. Therefore, we analyze the data as a time series rather

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

than summarized counts as in the previous study. We divide the entire dialogue into segments. In each segment, regardless of which hand the speaker uses, we identify every gesture and measure its duration. Each gesture falls into one of the two classes. 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 the reports from participants but from the video observations. First, transcripts are created from videos. Next, the points at which any word or phrase that may indicate the existence of a miscommunication are listed. Then, these points of suspicion are checked against the original video taking into account the speech sound and other modalities such as facial expressions and eye gaze. 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 the miscommunication.

2.2 Feature set

For both the communicative and non-communicative gestures produced by clients, in each segment s from all total S segments in a dialogue, we derive the following features from the basic gesture code data defined in Section 2.1: First, the gesture frequencies on, before, and after the target time slot are respectively denoted as $x_1(s)$, $x_2(s)$,and $x_3(s)$. The gesture frequencies were calculated at the gesture starting points: how many times a gesture was initiated in a given window of size W. For simplicity in calculating the gesture frequencies, we took into consideration only the gestures starting in the segment and not the continuing or ending ones. We computed the gesture frequencies of the current segment, at W seconds in the past and at W seconds in the future. Second, the differences in the gesture frequencies between the sth and $(s-1)$th or $(s+1)$th segments were also calculated, respectively denoted as $x_4(s)$ and $x_5(s)$. Third, the mean, maximum, and minimum duration of the gestures in each segment, respectively denoted as $x_6(s)$, $x_7(s)$, and $x_8(s)$. Finally, the mean interval of a speaker's gesturing, $x_9(s)$. The resulting representation of gestural cues in a segment s is represented as $\mathbf{x}(s) = (x_1(s), \ldots, x_9(s))$ that is a 9 dimensional vector. Among these feature values, $x_1(s)$, $x_2(s)$, and $x_3(s)$ are the nonnegative integers; $x_4(s)$ and $x_5(s)$ are the integers; $x_6(s)$, $x_7(s)$, $x_8(s)$, and $x_9(s)$ are the nonnegative real numbers. For the window size, W, we used 5 and 50 sec. The two segmentation window sizes correspond to the short-term and long-term dependencies between the gestural signals and semantic miscommunication. The 5-sec window can be used to determine how the gestures are used during the miscommunication. The number is determined based on the intuition that a single exchange of messages may be completed in about 5 sec. In contrast, the 50-sec window captures the overall gestural trend that induces miscommunication. This number is 10 times the smaller window size that is considered long enough to contrast with the short-term phenomena. 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 the window size, we assign a binary label $y \in \{0, 1\}$ to each segment based on whether it contains miscommunication, as shown in Fig. 1. Occurrences of multiple miscommunications in a segment are ignored. Since the number of segments is dependent on the window size, and the degree of class bias or the ratio of positive data changes, different segmentation results in different task difficulties. The feature

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

values are derived by checking the gesture starting points in each segment. The lower part of Fig. 1 shows the extraction process of $x_1(s)$ 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, where each set of braces represents a set of class label (either miscommunication or not) and corresponding feature values at each segment: $\langle\{y_1, \mathbf{x}^c(1), \mathbf{x}^{nc}(1)\}, \ldots, \{y_s, \mathbf{x}^c(s), \mathbf{x}^{nc}(s)\}, \ldots, \{y_S, \mathbf{x}^c(S), \mathbf{x}^{nc}(S)\}\rangle$.

2.3 Classifier

We train binary classifiers to assess if there is a cue that can predict whether a time segment contains miscommunications. That is, the classes of a segment that are denoted by y are assumed to be binary (positive or negative). The first classification method we use is the following linear discriminant analysis (LDA) [8], which is often used as a good baseline classifier:

$$\begin{aligned} \delta_k(x) &= x^T \textstyle\sum_k^{-1} \mu k - \frac{1}{2} \mu_k^T \sum^{-1} \mu k + \log \pi_k \\ \text{with } y &= \arg\max_k \delta_k(x) \end{aligned} \quad (1)$$

where k represents a class (miscommunication or smooth communication in our case) index, Σ the covariance matrix of the observations, μ the mean of the observations, and π the prior probability of the class. The second classifier is a support vector machine (SVM), which can generate non-linear classification boundaries:

$$y = \operatorname{sgn}((w \cdot \Phi(x) - b) \quad (2)$$

where w is the weight, b is a bias, Φ is a feature map $\chi \rightarrow F$ where χ is a set of observations and F is a dot product space. SVMs are reported to discriminate classes well in various domains and worked fine in a deception detection task [13]. We use the LibSVM implementation [5] with radial-basis functions (RBFs) as its kernel function (Φ).

In the experiment, we compared different configurations of classifiers. Since the number of smooth segments is far larger than the miscommunication segments, we believe calibration is needed to adjust to the data imbalance between the two classes. For LDAs, we can use either a uniform prior (i.e., $\pi_{+1} = \pi_{-1} = 0.5$) or empirically estimated prior as the ratio of the observations (i.e., $\widehat{\pi}_k = N_k/N$ where N_k is the number of class-k observations) that accounts for the data imbalance. For SVMs, one-class SVMs that identify a boundary of a single class distribution rather than discriminate two classes are used. Through preliminary experiments, we found that empirically weighted LDAs performed better than the uniform prior models, and one-class SVMs performed better than the standard SVMs. Based on these results, we show the results for empirically weighted LDAs and one-class SVMs only.

In contrast to the segmental classification we conducted, there is the possibility of modeling dialogues as a time series. For example, in [1], hidden Markov models are used to detect dialogue scenes in TV programs. Since we do not have any knowledge that supports that there is a stationary hidden state corresponding to the miscommunication, and the amount of data is too small to reliably estimate the probabilistic models with many parameters, we used the above simplified segment-based classification approach.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 was 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 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 extracted from clients' behavior. First, we segmented the entire conversation into either 5- or 50-sec segments as we extracted the feature values. Among these S segments, we took sth out and trained a classifier using a gesture feature that belonged to the remaining $S - 1$ time slots. Then, we classified the sth 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.

3.3 Experimental results

We summarized the classification results in terms of the F-measures in the 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 miscommunications, when the 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. The gesture types, either communicative (c) or non-communicative (nc), are associated with the 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 easily understood. When both tables are viewed column-wise, we can see that there is no single gestural feature consistently marked with the highest F-

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

scores; rather, the useful feature is data session dependent. The gesture frequency of previous segment $x_2(s)$ could be somewhat more useful than others; however, what we can assume is only that some features, including the gesture frequencies in current and subsequent segments ($x_1(s)$ and $x_3(s)$), which are the mean and minimum values of gesture duration ($x_6(s)$ and $x_8(s)$), are not good candidate cues for annotating miscommunications. When the tables are viewed row-wise, there is no consistently strong classifier or 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 the gestures and semantic miscommunications. Also, it should be noted that the gesture type, either communicative or non-communicative, did not directly relate to a 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 the clients' gestures and those of the therapists themselves. We took into consideration only the clients' gestures because we are often not fully aware of our own behavior, and clients' gestures are considered more useful to therapists. The next difference is the temporal range that a therapist has to observe to detect the saliency of the gestural features. If therapists have to find the changes that occur in a long time range, say 50 sec in our experiment, it would force a more cognitive load on the therapists than detecting changes within 5 sec. 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 the problematic point is over the long term. Therefore, even though machines can easily annotate miscommunication over the long term 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.

We used two classifiers: LDAs and SVMs. As shown in the experimental results, the classification results differ among the classifiers. If other classifiers are used, we might have obtained different results. However, we consider the two classifiers representative in that they are extremes in terms of the number of parameters to be estimated from the data.

5 Conclusion

We tested the automatic classification of conversational segments into smooth communications and miscommunications based on the gestural cues taken from psychotherapeutic interview sessions. The classifiers were trained on actual conversational data. This process clarified which gesture cues were useful in predicting miscommunications. The 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 further study detailed gesture types, gesture sub-units, handedness, or gesture strength, which could not be investigated under the current experimental setting. Also, the size of the dataset should be enlarged by adding more dialogue sessions for generalizability. In addition, since many miscommunications are

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

triggered by verbal contents, we can study the relationships between gestures and speech types. The integration of gesture signals with other modalities in the framework of human-computer interaction might be interesting as well [10]. 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 would like to thank the 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.

Biographies

Masashi Inoue is an assistant professor at Yamagata University. He received the B.A. degree from International Christian University (ICU), Japan in 1999, the M.S. and D.S. degree from Nara Institute of Science and Technology (NAIST), Japan in 2001 and 2004, respectively. He was a research associate at the National Institute of Informatics (NII), Japan from 2004 to 2009. His research interests are in cognitive science and computational knowledge discovery.

Mitsunori Ogihara is Professor of Computer Science at the University of Miami and Director for Data Mining in the University's Center for Computational Sciences. He received his Ph.D. degree in Information Sciences from Tokyo Institute of Technology in 1993. From 1994 until 2007 he was a faculty member of the Department of Computer Science at the University of Rochester, where he served as chair between 1999 and 2007. He is a recipient of a prestigious NSF CAREER Award and an ACM Distinguished Scientist Member. He has published two books (one co-authored), more than 60 journal articles, and more than 80 conference articles. He is currently on the editorial board of International Journal of Foundations of Computer Science (World Scientific Press), Theory of Computing Systems (Springer), and Centeral European Journal of Computer Science (Versita).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Ryoko Hanada is Associate Professor of the Department of Clinical Psychology in the Center for Clinical Psychology and Education at Kyoto University of Education. She received her M.A. (2002) and Ph.D. (2007) in clinical psychology from Tohoku University. She completed her clinical internship and JSPS Research Fellow (2004) and joined as Lecturer the faculty of the Department of Clinical Psychology at Kyoto University of Education the same year. Her therapeutic approach is Family therapy, particularly Brief Therapy. Her research interests are counselor education, family education, law related education and school improvement plan.

Nobuhiro Furuyama is an associate professor at National Institute of Informatics, Research Organization of Information and Systems in Japan, The Graduate University for Advanced Studies, and Tokyo Institute of Technology. He received his Ph.D. in psychology from the University of Chicago in 2001. His research interests include inter-personal and human-machine interaction from the viewpoint of ecological psychology. He is a member of International Society of Ecological Psychology, Japanese Cognitive Science Society.

References

1. Alatan AA, Akansu AN, Wolf W. Multi-modal dialog scene detection using hidden Markov models for content-based multimedia indexing. Multimed Tools Appl. 2001; 14:137–151.
2. Brown D, Parks JC. Interpreting nonverbal behavior, a key to more effective counseling: review of literature. Rehabil Couns Bull. 1972; 15(3):176–184.
3. Burgoon, J.; Adkins, M.; Kruse, J.; Jensen, ML.; Meservy, T.; Twitchell, DP.; Deokar, A.; Nunamaker, JF.; Lu, S.; Tsechpenakis, G.; Metaxas, DN.; Younger, RE. An approach for intent identification by building on deception detection. Hawaii international conference on system sciences 1, 21a; 2005.
4. Buttny, R. Talking problems. State University of New York Press; New York: 2004.
5. Chang, CC.; Lin, CJ. [Accessed 24 May 2010] LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
6. Erol A, Bebis G, Nicolescu M, Boyle RD, Twombly X. Vision-based hand pose estimation: a review. Comput Vis Image Underst. 2007; 108(1–2):52–73.
7. Freedman, N. Hands, word and mind: on the structuralization of body movements during discourse and the capacity for verbal representation. Plenum Press; New York: 1977. p. 219-235.
8. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer; New York: 2001.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

9. Heath, C. Body movement and speech in medical interaction. Cambridge University Press; Cambridge: 1986.
10. Jaimes A, Sebe N. Multimodal human-computer interaction: a survey. Comput Vis Image Underst. 2007; 108(1–2):116–134.
11. McNeill, D. Hand and mind. The University of Chicago Press; Chicago: 1992.
12. McNeill, D. Annotative practice. 2008. Available at: http://mcneilllab.uchicago.edu/pdfs/susan_duncan/Annotative_practice_REV-08.pdf
13. Meservy TO, Jensen ML, Kruse J, Twitchell DP, Tsechpenakis G, Burgoon JK, Metaxas DN, Nunamaker JF Jr. Deception detection through automatic, unobtrusive analysis of non-verbal behavior. IEEE Intell Syst. 2005; 20:36–43.
14. Mortensen, CD. Human conflict. Rowman & Littlefield; Cambridge: 2006.
15. Pavlovic VI, Sharma R, Huang TS. Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans Pattern Anal Mach Intell. 1997; 19:677–695.
16. Rahman, AM.; Hossain, MA.; Parra, J.; Saddik, AE. Motion-path based gesture interaction with smart home services. Proceedings of the seventeen ACM international conference on multimedia, MM '09; 2009. p. 761-764.
17. Suchman L, Jordan B. Interactional troubles in face-to-face survey interviews. J Am Stat Assoc. 1990; 85(409):232–241.
18. Takeuchi H, Subramaniam LV, Nasukawa T, Roy S. Getting insights from the voices of customers: conversation mining at a contact center. Inf Sci. 2009; 179(11):1584–1591.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

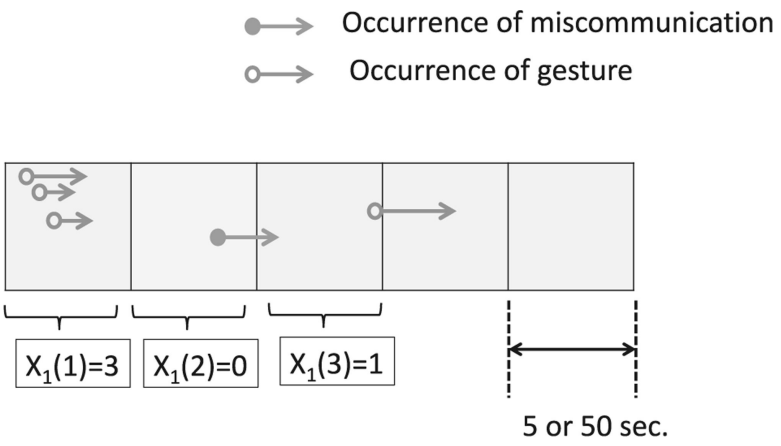

Fig. 1.
Schematic illustration of segmentation and coding of segments for different window sizes. The *root circles* indicate the starting points of gestures and miscommunications. The *arrows* represent the duration of the gestures and miscommunications

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Gestural features calculated on segment s

	Gestural feature	Interpretation
$x_1(s)$	Frequency in current segment	Degree of gestural activity
$x_2(s)$	Frequency in previous segment	Degree of gestural activity
$x_3(s)$	Frequency in next segment	Degree of gestural activity
$x_4(s)$	Frequency difference from previous segment	Degree of change in gestural activity
$x_5(s)$	Frequency difference in next segment	Degree of change in gestural activity
$x_6(s)$	Duration (mean)	Degree of complexity of gesture
$x_7(s)$	Duration (maximum)	Degree of complexity of gesture
$x_8(s)$	Duration (minimum)	Degree of complexity of gesture
$x_9(s)$	Mean interval	Degree of gestural continuity

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Overview of datasets

Dataset	Duration (min.sec)	Therapist (experience)	Client
1-(1)	24.17		
1-(2)	25.43		
1-(3)	9.07	Female (expert)	Female
2-(1)	12.48		
2-(2)	21.41		
2-(3)	40.58	Female (intermediate)	Male
3-(1)	17.02		
3-(2)	26.43		
3-(3)	22.41	Female (beginner)	Male

Each dialogue consists of three videos

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

F-scores short-term, 5-sec window

Dataset	Classifier	Features								
		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	0.010	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	–	–	–	–

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Dataset	Classifier	Features								
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
	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	–

Multimed Tools Appl. Author manuscript; available in PMC 2013 November 01.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 4

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

Dataset	Classifier	Features								
		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	–

Dataset	Classifier	Features								
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
3-(2)	SVM(nc)	0.32	0.57	0.42	0.50	0.17	0.20	0.40	0.14	–
	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

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript