MSP-AVATAR Corpus: Motion Capture Recordings to Study the Role of Discourse Functions in the Design of Intelligent Virtual Agents

Najmeh Sadoughi[1], Yang Liu[2], Carlos Busso[1]
[1] Multimodal Signal Processing Lab
[2] Human Language Technology Research Institute
University of Texas at Dallas, USA
Emails : nxs137130@utdallas.edu, yang.liu@utdallas.edu, busso@utdallas.edu

***Abstract*— Nonverbal behaviors and their co-occurring speech interplay in a nontrivial way to communicate a message. These complex relationships have to be carefully considered in designing *intelligent virtual agents* (IVAs) displaying believable behaviors. An important aspect that regulates the relationship between gesture and speech is the underlying discourse function of the message. This paper introduces the MSP-AVATAR data, a new multimedia corpus designed to explore the relationship between discourse functions, speech and nonverbal behaviors. This corpus comprises motion capture data (upper-body skeleton and facial motion), frontal-view videos, and high quality audio from four actors engaged in dyadic interactions. Actors performed improvisation scenarios, where each recording is carefully designed to dominate the elicitation of characteristics gestures associated with a specific discourse function. Since detailed information from the face and the body is available, this corpus is suitable for rule-based and speech-based generation of body, hand and facial behaviors for IVAs. This study describes the design, recording, and annotation of this valuable corpus. It also provides analysis of the gestures observed in the recordings.**

I. INTRODUCTION

We use multiple nonverbal behaviors when we communicate with others. In addition to speech, we use head motion, facial expressions, hand gestures, and body postures, which are intricately coordinated, and synchronized with speech to convey the message [4], [6], [16], [24]. Valbonesi et al. [24] showed that more than 90% of acoustic events (e.g., maximum and minimum of the pitch and RMS energy) occur during hand gesture strokes. While backchannel gestures are definitely important, most of these gestures are generated when we are speaking [16].

Synthesizing human like behaviors can be helpful for animation, entertainment, virtual reality applications and hearing impaired aid devices. They do not need to look like human (e.g. cartoons), but their gestures should display human-like behaviors (i.e, natural movements, synchronized with speech and gestures). To accomplish this, the complex relationship between gestures and speech has to be carefully considered in the design of *intelligent virtual agents* (IVAs) [3], [11].

Previous studies on creating *conversational agents* (CAs) can be categorized into two main approaches; rule based systems [6], [10], and data driven systems [3], [12], [14]. Rule based systems define certain rules for behaviors based on contextual information. These frameworks are prone to create repetitive behaviors that are not properly synchronized with speech. Data driven systems, in particular methods that use speech prosody features as input to generate animations can model the complex relationship between speech and gesture. Previously studies have shown that prosodic features such as the *fundamental frequency* (F0) and energy are effective modalities to synthesize nonverbal human-like behaviors [3], [8], [11], [12], [14]. They can provide information about the gestures, and their kinematic (i.e., timing, dynamic range, speed, acceleration). The higher the energy or the F0 range, the stronger the underlying motion of the gesture. A challenge in speech-driven animation is to generate behaviors that respond to the underlying discourse context (e.g., questions, statements, affirmation, negation) [21]. We expect to observe specific facial expressions, hand and head motions, given a specific discourse function in the message. For example, statements with affirmations are usually accompanied with head nods, while statements with negation are accompanied with head shakes and probably frowned eyebrows.

This paper introduces the MSP-AVATAR database, a multimedia corpus comprising motion capture data, audio and frontal view videos of actors engaged in dyadic conversation. A key feature of the database is the use of body and facial markers, which provide detailed information for modeling nonverbal behaviors for IVAs. Each scenario is designed to elicit characteristic gestures for a specific type of discourse function. We consider contrast, confirmation/negation, question, uncertainty, suggest, giving orders, warn, inform, large/small, and pronouns. After describing the design, collection, and annotation of the corpus, we provide preliminary analysis of the behaviors conveyed during each discourse function. The analysis relies on mid-level gesture representation automatically derived from the data.

The corpus provides an ideal resource to investigate the role of discourse functions during nonverbal human interactions, and its application to IVAs. After learning the underlying structures in the data, we expect that we will be able to synthesize talking avatars that convey more naturalistic behaviors and respond to the underlying discourse function.

II. REVIEW AND MOTIVATION

Previous studies have presented various motion capture corpora to study human behaviors. This section describes

This work was not supported by This work was funded by NSF (IIS 1352950)

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

some of these corpora. It also provides the motivation behind the collection of the MSP-AVATAR corpus.

A. Current Corpora with Motion Capture Data

The motion capture database HDM05 [18] was recorded at the University of Bonn, Germany, to study several human motions (e.g., walking, dancing, throwing). The target research problems for this corpus include the analysis, synthesis and classification of human motions. This database includes 100 motion categories, where five subjects performed multiple realizations, providing a broad range of variations for each motion. The corpus comprises three hours of data, where reflective markers were placed on the subjects' bodies.

The Carnegie Mellon University motion capture (MoCap) database (`http://mocap.cs.cmu.edu`) is another corpus to analyze human motion. It consists of two parts: kitchen and motions. The former comprises the recording of 55 subjects cooking five meals in a kitchen, where each recording took 15 minutes. The kitchen corpus includes, motion capture data, plus audio and video of the subjects. The latter has the motion capture data of various types of movements performed by 144 subjects in a number of trials. This corpus is rich in terms of variations of the behaviors and can be useful for research in animation.

The HumanEva corpus [23] was recorded at Brown University, USA. This corpus provides video (7 cameras) and motion capture data, where the goal is to establish a systematic evaluation of pose and motion estimation methods. It includes recordings from four subjects performing six predefined motions (walking, jogging) a number of times.

The Biological Motion Library [13] is a motion capture database recorded at the University of Glasgow, Scotland. It aims to study the role of personality, gender and emotion on people's behavior. The corpus consists of 30 nonprofessional actors performing actions (knocking, walking, lifting, and throwing) under four affective states (neutral, anger, sadness and happiness). The subjects provided 10 repetitions over knocking, lifting and throwing actions.

The *Korea University Gesture* (KUG) [9] is a corpus comprising high resolution video and motion capture data. This database was recorded from 20 subjects who performed 14 normal actions (e.g., "Raising a right hand", "Sitting on a chair"), and 10 abnormal actions that may occur during emergency scenarios (e.g., "Falling forward"). It also includes 30 command-based gestures for human-computer interactions (e.g., "Pointing at top-left"). The corpus is suitable for the study of human action recognition.

Another corpus is the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database [2], recorded at the *University of Southern California* (USC), USA. The corpus was recorded to study the role of emotion in spontaneous human interaction. The corpus comprises over 12 hours of motion capture data including 53 facial, two head and six hand markers. Ten trained actors participated in the recordings during dyadic sessions. In each recording, only one of the actors had the markers at a time. After collecting the scenarios, the markers were placed on the other actor and the sessions were collected again. The CreativeIT corpus [17] is another database recorded at USC. The database was recorded to study creativity in theatrical improvisation, providing resources to study expressive behaviors in human interactions. This database is recorded in a dyadic setting from 19 actors. It contains video, audio and motion capture recordings of the actors (only the body markers).

B. Motivation to Record the MSP-AVATAR database

Synthesizing a talking avatar that displays believable human-like behaviors is a challenging task. There are studies that derive gestures using data-driven approaches [3], [8], [11], [12], [14], [21], or rule-based approaches [5], [7], [19]. In both cases, considering the underlying discourse function is important to design behaviors that not only are timely aligned with speech, but also convey the right "meaning" of the message. For data-drive approaches, it is important to rely on a corpus that comprises characteristic behaviors associated with various discourse functions.

To the best of our knowledge, there are no corpora available to explore the role of discourse functions on human interactions. Therefore, the goal of the MSP-AVATAR corpus is to provide us with rich audiovisual recordings to design data-driven algorithms for IVAs, capturing the relationship between speech and gestures constrained by the underlying discourse function in the message.

III. Design of Scenarios

In this corpus, actors performed improvisation scenarios during dyadic conversations. We chose this framework, since a dyadic setting, where two actors improvise, creates more naturalistic behaviors than a setting where each actor is separately recorded in a monologue. It is not feasible to collect the data with natural interactions in less constrained recordings. However, this corpus provides suitable recordings for studies on synthesis. We carefully created the scenarios, which were designed to elicit characteristic behaviors for various discourse functions. The scenarios consist of common situations, which are prone to elicit speaking turns conveying the target discourse function.

The selection of the discourse functions considered in this corpus was motivated by the work of Poggi et al. [20] and Marsella et al. [15]. Poggi et al. [20] presented Greta, an *embodied conversational agent* (ECA). This toolkit has several discourse-related labels which create specific configurations for the ECA's movements. Marsella et al. [15] considered a set of semantic functions, where typical gestures were defined for each of them [15]. Following these studies, we selected ten categories, which we referred to as discourse functions: contrast, confirmation/negation, question, uncertainty, suggest, giving orders, warn, inform, large/small, and pronouns. These discourse functions usually elicit specific behaviors, which we aim to capture in the corpus. Table I lists the discourse functions and their definitions.

We created 2-5 scenarios for each of these discourse functions. In each recording, we presented a description of the scenario with the roles of the actors. We also describe

TABLE I

DEFINITION OF THE CONSIDERED DISCOURSE FUNCTIONS.

Discourse Function	Definition
Contrast	Contrasting two ideas, usually accompanied with contrast conjunctions such as *but*, *nevertheless*, *as*, and *as opposed to*.
Confirmation/Negation	Showing agreement and disagreement, usually accompanied with phrases such as *Yes*, *No*, and *I don't think so*.
Question	Asking a question of any type: *Yes-No* and *Wh-questions*, .
Uncertainty	Showing uncertainty in making a decision, might be accompanied by sentences such as *I really don't know what to do!*
Suggest	Suggesting ideas to the listener, e.g., *How about the new Japanese restaurant?*
Giving Orders	Ordering any type of service, e.g. *ordering food in a restaurant*.
Warn	Warning the listener of a danger, e.g. *Be careful about* ....
Inform	Inform something to the listener.
Large/Small	The act of referring to something as small or large during speaking. These scenarios target iconic gesture usually accompany these two words or any of their synonyms.
Pronouns	The act of referring to any pronoun (I/You/She/He/They). These scenarios target deictic gestures.

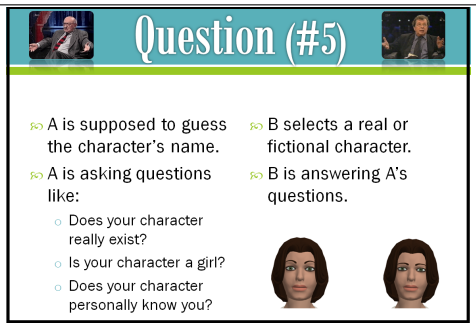


Fig. 1. A slide describing one scenario for the actors. It shows the description of the scenario and some prototypical behaviors associated with the target discourse function.

the context of the improvisation. We told the participants to incorporate as many gestures as they feel is natural in their performance (notice that the main purpose of the corpus is synthesis of behaviors). To clarify their understanding, we present illustrations of one or two prototype gestures accompanying the target discourse function. These gestures are created with Greta. Figure 1 shows a slide describing one of the scenarios for the discourse function *question*.

The duration of the recorded scenarios varies across the recordings (MEAN = 143.1 sec, STD = 74.7 sec.). We recorded motion capture data from four actors who improvised 21, 15, 22, and 16 scenarios, respectively. In total, we have 74 sessions. The number of recordings in each session is given by the pace of the actors during the data collection.

IV. DATA COLLECTION

A. Motion Capture Data

We recorded six nonprofessional actors in three dyadic sessions, one female, and five males. We asked the actors to glue 43 reflective markers to their faces following the layout illustrated in Figure2(a). The location of the facial markers include most of the *feature points* (FPs) defined in the MPEG-4 standard, facilitating the animation of talking avatars. We follow the Vicon Skeleton Template to capture the position of the joints of the actors' upper body. The actors wore a headband with 4 markers, and a suite with 28 markers. The body markers include three fingers (thumb, index, and pinky) as well. Figure 3 shows the positions of the skeleton markers.

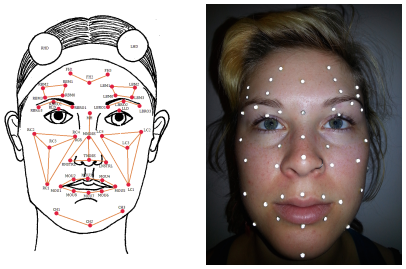
Collecting facial (small size, reduced volume) and body (big size, large volume) markers at the same time is a challenging task, since they require different resolutions. The cameras need to be close to the person to capture facial motion, which reduces the spatial area required to track the movements from the body markets. We create a setting where we have enough resolution to track facial and upper body markers, by placing a VICON system with ten Bonita Optical Cameras directed to one of the actor at different distances – see Figure 4(a). This setting allowed us to capture both the upper-body skeleton and facial motions. The only drawback of this approach is that when the actors were in rest position, their hands sometime were hidden from the cameras.

For the first session, we placed the markers on the second actors after collecting the scenarios for the first actors. Then, we recollected the scenarios. For the last two sessions, we only motion captured one actor. In total, we have motion capture data from four actors.

B. Audiovisual Recordings

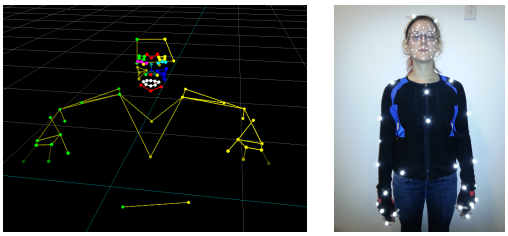
While the motion capture system provides data from one actor, we placed microphones and cameras for both actors. For the audio, we used a microphone connected to a digital recorder (TASCAM DR-100MKII). For the first session, we used a head-worn microphone (SHURE BETA 53), which attenuated the speech coming from the other subject. We noticed that using a head-worn microphone occluded some of the facial markers, making the post-processing steps to clean the data more difficult. Therefore, we decided to use a lavalier microphone (SHURE MX150) for the next sessions. We told the actor to avoid talking at the same time as much as possible. The digital recorder was set at 16 bit resolution and at a sampling rate of 44.1 kHz. Note that our instructions about overlapped speech served as high-level guidelines. In practice, we do observe overlapped speech. Since we use lapel microphones, cross-talking speech is very low.

We also recorded frontal view videos of both of the actors.



(a) Facial Markers (b) Actress

Fig. 2. The position of 43 facial markers.



(a) Marker Skeleton (b) Actress

Fig. 3. Placement of the body and facial markers for the data collection.

We used two Sony handycams HDR-XR100, which record with 1920 $\times$ 1080 resolution in Full HD. Figure 4(b) shows a snapshot of one of the cameras. We use these videos to annotate the data. They can also be useful in extracting facial features from the actors. Notice that we use a clapboard with two reflective markers on the corners, to synchronize audio, video and motion capture data. We collected 74 sessions.

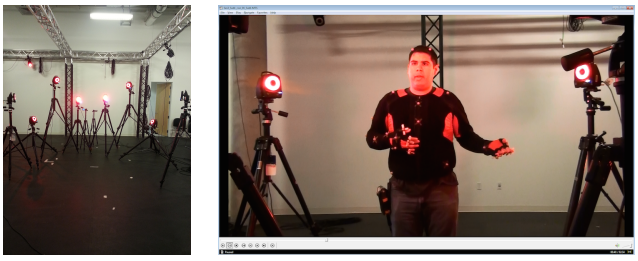
V. Post-Processing and Data Annotation

After collecting the data, we started the post processing steps to clean the motion capture data. We use the software Blade to track the markers. First, we created a skeleton personalized to each of the actors with all the markers. Then, we use the auto-labeling option in Blade. Capturing face and upper-body markers introduce noise in this process. Therefore, we are manually correcting the labels of the markers. For example, we have to fill the gaps for the hands while the subjects are standing, having their hands outside the volume captured by the cameras. At the moment of this submission, we have corrected the upper body markers of 37 scenarios, and the facial markers of 3 scenario.

Another important post processing step is to segment the corpus into speaking turns. We manually segmented the dialogs focusing on the recordings from the subjects being motion captured. We consider speaking turns conveying the discourse functions considered in this study, which were annotated by one subject. We are planning to have a second annotator to increase the reliability of the labels. The annotation process included 74 scenarios recorded from four actors. At the moment of this submission, the corpus has 1751 segments annotated with discourse function labels.

VI. Analysis of the Database

This section analyzes the content of the MSP-AVATAR corpus. The analysis considers the annotations of the discourse functions, and the different head and hand gestures observed under different discourse functions. For the analysis, we split the class *large/small* into *Large-iconic* and *Small-iconic*, and the class *pronouns* into *I-deictic*, *You-deictic* and *Other-deictic*. We also consider *Statement*.



(a) VICON Cameras (b) Snapshot from the Video

Fig. 4. Setting for the data collection. The second actors is behind this setting facing the target actor.

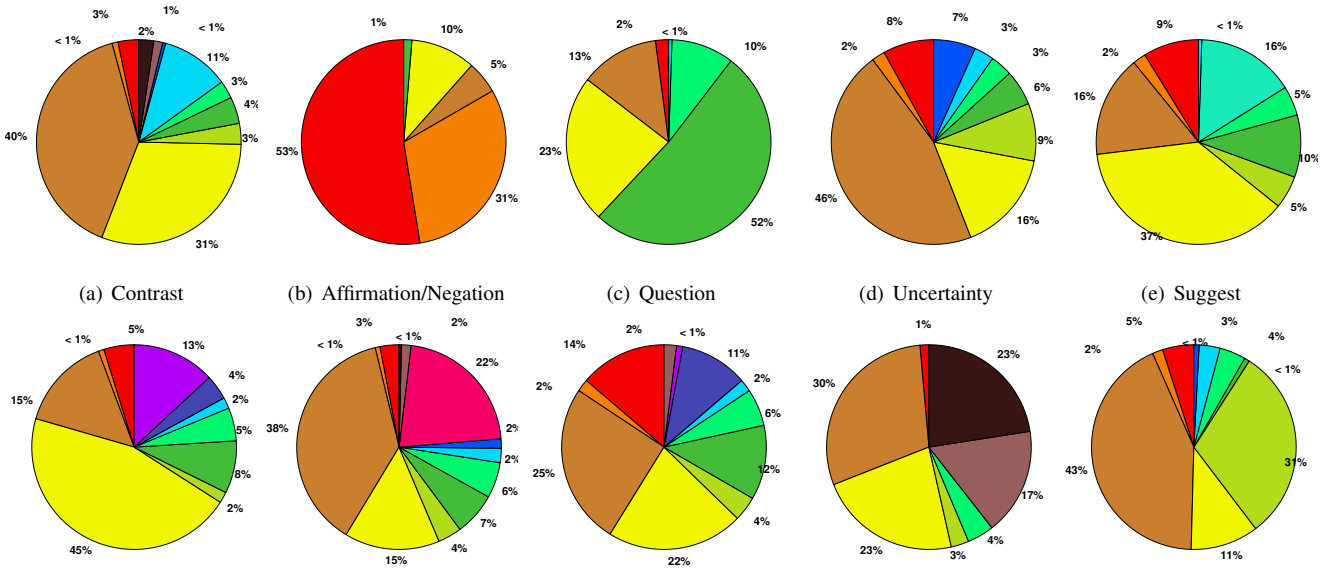
A. Analysis of the Discourse Functions Annotations

Figure 6 displays the distribution of the discourse functions annotated in the data. Figure 5 provides the pie charts per scenario, where we use the same color convention used in Figure 6 to identify the discourse functions. For example, Figure 5(f) shows that most of the discourse functions found on the scenarios for class *warn* are *warn* (43%) and *Question* (23%). Figure 6 shows that the distribution of the segments is not uniform across the discourse function. This is due to two reasons. First, the distribution of the performed scenarios is not uniform across the target discourse functions. We created two to five scenarios per class, but we only collected a subset of them. Second, some discourse functions, such as *Question*, are more common than others, such as *Contrast*, which is reflected in the corpus.

Figure 5 shows that the target discourse function dominates the classes observed in the recordings. The scenarios were successful in eliciting their target discourse function. Overall, the database provides several examples of the selected discourse functions. This corpus will facilitate the design of speech-driven animations constrained by the underlying discourse function.

B. Analysis of Head Motion and Hand Gestures

This section analyzes hand and head gestures generated by each discourse function. Our goal is to determine whether the underlying discourse function of the message affects the behaviors of the subjects. As mentioned in Section V, we have manually corrected the motion capture data for 37 scenarios (body motion), and this analysis relies on these recordings. We separately examine head movements and hand gestures, since they may not co-occur. Instead of manually annotating specific behaviors in the data, we decide to rely on a data-driven framework based on *Parallel hidden Markov Models* (PHMMs). The model clusters hand and head behaviors into automatically determined data-driven mid-level classes. PHMM is an appropriate clustering method for gestures, since it allows for dynamic segmentation and clustering of the data. This model consists of a number of parallel branches (i.e., individual left-to-right HMMs) with the same number of states. All the states within the branches have self transitions which give the flexibility to model dynamic behaviors having different temporal durations. We use the implementation provided by Sargin et al. [22]. Previous studies have used PHMM to segment head motion [22], hand gestures [25], and upper-body [1] into mid-level representations. We use a similar approach in this analysis. After segmenting the behaviors into different clusters, we study the distribution of these clusters associated with each of the discourse functions. Below we describe the details.



(a) Contrast (b) Affirmation/Negation (c) Question (d) Uncertainty (e) Suggest

(f) Warn (g) Giving Orders (h) Inform (i) Large/Small (j) Pronouns

Fig. 5. Distribution of the labels assign to each of the target discourse functions.

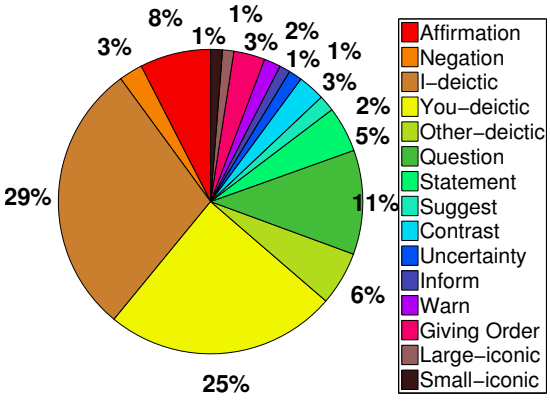


Fig. 6. Distribution of the discourse function labels in the corpus.

The number of states per cluster determines the minimum duration of the clusters which we set equal to 15 (approximately 125ms). We consider 15 branches for the head and hand models, which provide a reasonable number of classes for the behaviors. Related studies have used 5 (only head [22]) and 50 (hand gestures [25]) branches. Note that several segments describe gestures while listening, which should be considered in PHMM as well. Increasing the number of clusters provides finer representation of behaviors, and results in higher likelihood of the data given the model (it decreases the within class variance). However, it increases the number of parameters. We chose 15 clusters, where the total number of parameters for head and hands are 2743 and 9043. The average numbers of frames for training each parameter are 320.9 and 97.3, respectively.

From the motion capture data, we extract the rotation of the bones around the joints in the Skeleton, which serves as our features. For head motion, we consider the three angular rotations, and their first derivatives (6 D). For hand gestures, we consider the arm (3 DOF per arm) and forearm (2 DOF per forearm) angular rotations and their first derivatives (20 D). We expect that the clusters and their durations will be different for hand and head gestures. Therefore, we separately run the PHMM to cluster their behaviors. We process the entire recordings, without considering the discourse function segmentation. The data was Z-normalized across all the recordings.

After segmenting the recordings into data-driven clusters, we analyze their distributions associated with each discourse function, using the manual annotations (see Sec. V). Figure 7(a) and 7(b) give the results for head motion and hand gestures, respectively. The y axis in these figures is the normalized count of the frames shared between each cluster and discourse function, and the x axis is the cluster number. The results show that there are marked differences in head motions and hand gestures across different discourse functions. For hand gesture, *Affirmation* and *Negation* have a distribution with a peak in cluster 14. By inspecting the hand gestures assigned to this cluster, we noticed that the movements are less active than in other clusters, suggesting a reduction of hand motion for these discourse functions. We also found out that cluster 15 (for hand), which appears mostly in *Uncertainty* incorporates the gesture of moving both hands from waits to chest. Moreover, we investigated head motions and found the peak in *Affirmation* to be related to the cluster of head nodes, and cluster 4 which mostly occurs in *Negation* to include head shakes.

VII. Conclusions

This paper introduces the MSP-AVATAR database, a multimodal corpus comprising motion capture data (body and face), high quality audio, and high definition video of actors engaged in dyadic conversations. Six actors performed improvisation scenarios designed to elicit target discourse functions. The corpus includes over one thousand speaking turns annotated with different discourse functions. This corpus is ideal to study the relationship between speech and

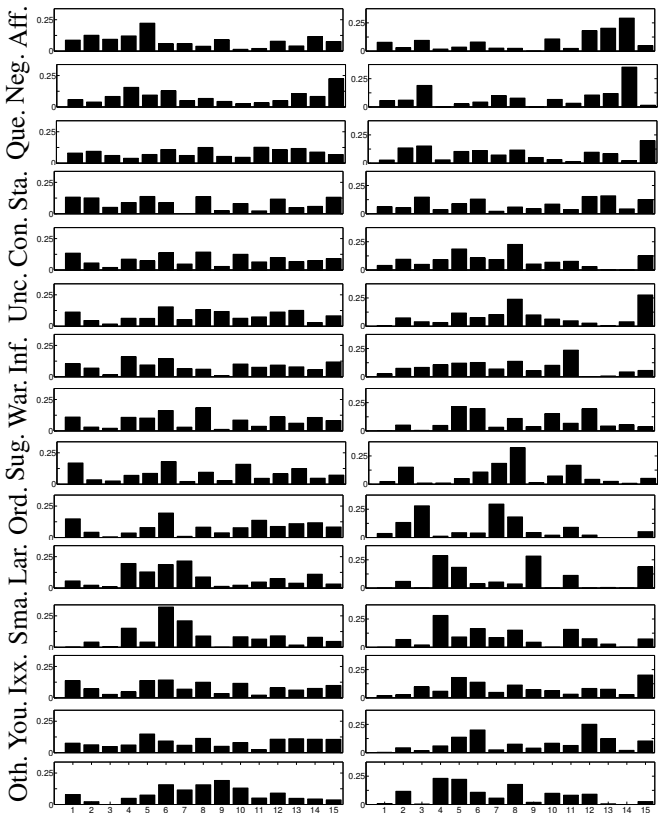


Fig. 7. The distribution of the PHMM-based head (a) and hand (b) clusters for each discourse function (Aff: Affirmation, Neg: Negation, Que: Question, Sta: Statement, Con: Contrast, Unc: Uncertainty, Inf: Inform, War: Warn, Sug: Suggest, Ord: Order, Lar: Large-iconic, Sma: Small-iconic, Ixx: I-deictic, You: You-deictic, Oth: Other-deictic).

gestures as dictated by the underlying discourse function.

By simultaneously collecting facial expression and body motions, the corpus provides unique opportunities to explore interesting research questions. One drawback of the of this database is the small number of actors that were motion captured. While we have the facilities to collect more sessions, the main bottleneck is cleaning the motion capture data, which has been slower than expected. After finishing this process, we expect to release this corpus to the research community. We hope that this corpus will be a valuable resource to explore the role of discourse functions in the synthesis of human-like behaviors for IVAs.

REFERENCES

[1] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and E. Erzin. Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3652–3656. IEEE, 2013.

[2] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.

[3] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.

[4] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, November 2007.

[5] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14:55–64, March 2001.

[6] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, pages 413–420, Orlando, FL,USA, 1994.

[7] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer, 2004.

[8] Z. Deng, C. Busso, S. Narayanan, and U. Neumann. Audio-based head motion synthesis for avatar-based telepresence systems. In *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP 2004)*, pages 24–30, New York, NY, October 2004. ACM Press.

[9] B. W. Hwang, S. Kim, and S. W. Lee. A full-body gesture database for automatic gesture recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 243–248. IEEE, 2006.

[10] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents*, pages 205–217. Springer, 2006.

[11] B. H. Le, X. Ma, and Z. Deng. Live speech driven head-and-eye motion generators. *IEEE transactions on visualization and computer graphics*, 18(11):1902–1914, 2012.

[12] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics (TOG)*, 29(4):124, 2010.

[13] Y. Ma, H. M. Paterson, and F. E. Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, 38(1):134–141, 2006.

[14] S. Mariooryad and C. Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2329–2340, October 2012.

[15] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35. ACM, 2013.

[16] D. McNeill. *Hand and Mind: What gestures reveal about thought*. The University of Chicago Press, Chicago, IL, USA, 1992.

[17] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan. The USC CreativeIT database: A multimodal database of theatrical improvisation. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, Valletta, Malta, May 2010.

[18] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Computer Graphics Technical Reports CG-2007-2, University of Bonn, Bonn, Germany, June 2007.

[19] I. Poggi and C. Pelachaud. Performative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, page 154188. MIT Press, Cambridge, MA, USA, 2000.

[20] I. Poggi, C. Pelachaud, F. Rosis, V. C. B., and Carolis. Greta. a believable embodied conversational agent. *Multimodal Intelligent Information Presentation*, pages 3–25, 2005.

[21] N. Sadoughi, Y. Liu, and C. Busso. Speech-driven animation constrained by appropriate discourse functions. In *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014.

[22] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(8):1330–1345, 2008.

[23] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120, 2006.

[24] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. McCullough, and R. Bryll. Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures. In *European Signal Processing Conference (EUSIPCO 02)*, pages 75–78, Tolouse, France, September 2002.

[25] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan. Analysis of interaction attitudes using data-driven hand gesture phrases. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 699–703. IEEE, 2014.