IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 16, 2020, accepted November 26, 2020, date of publication December 1, 2020,
date of current version December 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041672

Analyzing Multifunctionality of Head Movements in Face-to-Face Conversations Using Deep Convolutional Neural Networks

KAZUHIRO OTSUKA[1], (Member, IEEE), AND MASAHIRO TSUMORI[2]
[1]Faculty of Engineering, Yokohama National University, Yokohama 240-8501, Japan
[2]Graduate School of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology, Nagaoka 940-2188, Japan

Corresponding author: Kazuhiro Otsuka (otsuka@ieee.org)

This work was supported by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, under a joint research agreement between NTT and Yokohama National University.

ABSTRACT A functional head-movement corpus and convolutional neural networks (CNNs) for detecting head-movement functions are presented for analyzing the multiple communicative functions of head movements in multiparty face-to-face conversations. First, focusing on the multifunctionality of head movements, i.e., that a single head movement can simultaneously perform multiple functions, this paper defines 32 non-mutually-exclusive function categories, whose genres are speech production, eliciting and giving feedback, turn management, and cognitive and affect display. To represent and capture arbitrary multifunctional structures, our corpus employs multiple binary codes and logical-sum-based aggregations of multiple coders' judgments. A corpus analysis targeting four-party Japanese conversations revealed multifunctional patterns in which the speaker modulates multiple functions, such as emphasis and eliciting listeners' responses, through rhythmic head movements, and listeners express various attitudes and responses through continuous back-channel head movements. This paper proposes CNN-based binary classifiers for detecting each of the functions from the angular velocity of the head pose and the presence or absence of utterances. The experimental results showed that the recognition performance varies greatly, from approximately 30% to 90% in terms of the F-score, depending on the function category, and the performance was positively correlated with the amount of data and inter-coder agreement. In addition, we noted a tendency toward overdetection that added more functions to those originally in the corpus. The analyses and experiments confirm that our approach is promising for studying the multifunctionality of head movements.

INDEX TERMS Deep neural networks, gesture recognition, meeting analysis, multimodal sensor, nonverbal behaviors, social signal processing.

I. INTRODUCTION

Computational analysis of human social behavior has emerged as a frontier in computer science, referred to as *social signal processing* [1]. Among the various social behaviors, an attractive target of study in social signal processing is behaviors appearing in face-to-face conversations [2], [3]. Here, the knowledge and means of automatically recognizing conversational behaviors are useful for developing embodied conversational agents (ECAs) [4], [5] and other applications. In particular, *nonverbal* behaviors, such as head movements, gaze, facial expressions, bodily gestures, and vocal prosody, play important roles in face-to-face settings. As a type of visual nonverbal behavior, this paper focuses on *head movements*, which, in conversations, serve multiple essential functions related to speech production, eliciting and giving feedback, managing speaking turns, and displaying one's attitude, cognitive states, and emotions. In addition, together with other nonverbal behaviors, head movements appearing in conversations can provide useful clues to assessing various aspects of individuals and groups, such as communication skill [6], [7], personality traits [8], [9], leadership [10], [11], and team performance [12].

In particular, head movements have been recognized as a useful nonverbal communication modality in the context of embodied conversational agents, and the technologies for head-movement recognition and synthesis have been considered keys to improving both the intelligence and interactivity of an agent. This is because the user's head movements reveal a variety of internal states, including interest, motivations,

The associate editor coordinating the review of this manuscript and approving it for publication was Sandra Baldassarri.

and level of understanding of what the agent is trying to say. Therefore, if the agent can recognize the user's head movements and 'mind-read' the user's attitudes and cognitive state from them, it can promptly regulate its own behaviors, including its head movements, toward the mutual goal of the users and agent; e.g., it can build rapport through mimicry of the user's nods [13]–[15]. In addition, synthesis of the ECA's head movements is also an important task because it enables users to better understand the intentions and internal states of the agent, e.g., the level of attentiveness, in an intuitive way [13], [16], [17].

Another application area involving head movements is computer-mediated communications (CMCs), especially robotic telepresence [18], which exploits robotic heads as surrogates of remote users. The robotic head, which typically represents or displays the remote user's face, is controlled automatically by tracking the user's head [18], [19] and/or controlled manually through user interfaces [19]. Intelligent control of robotic heads is an important research topic that seeks to compensate for communication deficits caused by transmission delays and limited audio/visual channels and to improve the interactivity among spatially separated people so that they can interact as if they were all together in the same place; here, the aim is to adaptively regulate the robot's movements depending on the user's intentions, as recognized from his/her head movements and other behaviors, and on the communication environment. The computational techniques for analyzing and understanding spontaneous head movements appearing in natural human-human conversations are thus considered to be worth exploring as the basis for developing these applications.

One distinctive aspect of head movements is their diversity of function, called *multiple functionality*, wherein a single head movement can manifest more than one function. For example, a *nod* is generally interpreted as a sign of 'yes', and a *shake* is interpreted as a sign of 'no' [20]; these are called *emblem* gestures [21]. However, a single *nod* in a conversation can also be interpreted as indicating *listening*, *understanding*, or *agreement*. Nods with different meanings might have distinct kinetics, but the boundaries of these kinetics seem rather vague and may change depending on the context, individual, and culture or language. Moreover, a single nod sometimes manifests multiple functions at the same time, e.g., by simultaneously displaying signs of listening and understanding. In addition, intentional behaviors have communicative functions; unintentional behaviors can influence others' thoughts and actions. Considering the above, this study views multifunctionality as a mixture of potential meanings encoded in the sender's behaviors and a diversity of interpretations by different receivers in multiparty conversations. Furthermore, we assume that this multifunctionality could serve as a driving force of conversations, taking the perspective that conversation is a process of disambiguating each other's expressions and meanings to create mutual understanding and shared consensus among the participants.

The multiple functionality of head movements has also been referred to as being *multiple level* [22], [23], of a *multitude* [23], and an example of *polysemy* [24], [25]. Note the term *polysemy* originally referred to a phenomenon in linguistics whereby a single word is associated with multiple meanings [26]. Poggi *et al.* borrowed this term to shed light on the multifunctionality of head nods [24], [25] under the assumption that the true meaning of a nod can be singled out among other candidates when the context is fixed. Hereafter, we will abbreviate the communicative functions of head movements as head-movement functions or just functions.

Compared with facial expressions [27] and hand gestures [28], [29], head movements are difficult to describe and analyze. Facial expressions and hand gestures can be decomposed into a set of elements, of which the spatial configuration/articulation clearly represents a specific form of meaningful gesture. For example, in Ekman's facial action coding system (FACS) [30], a combination of action units such as AU6 (cheek raise) + AU12 (lip corner pull) represents the expression 'smile'. On the other hand, head movements cannot simply be segmented into elements in space and time. Assuming the existence of such prototypical elements, which Birdwhistell's *Kinesics* [31] called *kinemes*, a number of studies have tried to describe head movements using kinetic categories such as nod, jerk, and shake [32]–[34], and/or *kinetic features* such as amplitude, frequency, and cyclicality [32], [35]–[37]. As reviewed in the next section, many researchers have tried to establish a link between kinetic categories/features and specific functions; however, a solid one-to-one correspondence between kinetics and functions has yet to be established. In fact, the study of the multifunctionality of head movements has so far been hampered by the complexity of the kinetics and diversity of the functions.

This study aims to provide a novel approach to exploring the structure of multiple communicative functions manifested through head movements in conversations. To this end, this paper first presents a *functional head-movement corpus* and reveals some aspects of multifunctionality by analyzing how and identifying which functions co-occur in conversations. Second, we examine the possibility of automatic recognition of multiple head-movement functions using the corpus. Reflecting on the past finding that there is no fixed one-to-one relation between kinetic categories/features and functions, we hypothesized that raw head-movement data, i.e., a time series of head-pose angles, still contain rich information for distinguishing functions. To verify this hypothesis, we employed modern machine-learning models called deep neural networks (DNNs) [38], in particular, convolutional neural networks (CNNs) [39], and we examined their capability of automatically detecting head-movement functions from raw head-pose data.

First, this paper presents our *functional head-movement corpus*. To represent the multifunctionality of head movements, this corpus employs non-mutually-exclusive binary codes, each of which represents the presence or absence of a function in a time frame (e.g., 1/30 sec.) in a conversation.

The set of multiple binary codes can represent an arbitrary multitude of functions. Specifically, we define 32 function categories—13 related to speaking, 12 for reacting, and seven others—that cover a wide range of functions, including speech rhythm and emphasis, feedback interactions, turn management, and cognitive and affect display. To capture the diversity of functional interpretation as well as possible, the *logical sum* of multiple coders' interpretations is used to create the final code. These strategies make our approach distinct from previous ones that assume mutually exclusive classes and use single-valued attributes as the ground truth agreed upon by the coders. Targeting conversations among four Japanese females, we built a corpus with the cooperation of three coders. Our co-occurrence analysis of the corpus revealed simultaneous multifunctional structures in which speakers modulate rhythmic head movements to perform various functions, such as emphasis and eliciting responses, and listeners modulate continuous back-channel head movements to express various attitudes and responses. We find that various cognitive and affect display functions are densely interrelated with each other and form a rich functional spectrum of meaning expressed through head movements.

Next, to automatically detect head-movement functions from head-movement data, this paper presents classification models based on CNNs. To date, most automated methods of recognizing head movements have targeted classifications of kinetic categories such as nods and shakes, and only a few studies have addressed the recognition of head movement functions [40]–[42]. The existing approach employs manually annotated kinetic categories as the input data and aims to find a link between the kinetic categories of the head movements and functions. However, this approach has trouble dealing with subtle differences in movements and in the manual categorization of kinetics without ambiguity [41]. In contrast, our approach aims to find a direct link between raw head-movement data and functions without using intermediate symbolic representations. To this end, we used CNN models with the expectation that they could model subtle but crucial differences in kinetics to distinguish different functions. To highlight head movements as the main information source, we employed a minimum bimodal approach, which uses only head pose and speaking activity, i.e., the presence or absence of an utterance, as the input data.

CNNs are widely used forms of DNNs that show excellent performance in various tasks [38], [39]. According to a survey [43], [44], CNNs have been the most frequently used DNN models for human behavior recognition based on the time series data of bodily movements [45]–[50], and its performance has been well verified for the purpose. Therefore, we decided to employ the CNN as the reasonable baseline that could reveal the potentials for the automatic recognition of head-movement functions. We built separate binary classifiers based on CNNs for the ten most frequent functions found in our corpus. The experimental results were insightful and promising. The recognition performance varied greatly, from approximately 30% to 90% in terms of the F-score, depending on the function category. Positive correlations were found between the performance measures and amount of data and between the performance measures and inter-coder agreement. These results support our hypothesis in the case in which a function category contains a large amount of data and there is substantial inter-coder agreement, while they do not support it in the case of a small amount of data and large variations in the coders' interpretations. The qualitative and quantitative analyses of the recognition results showed a tendency toward overdetection that added more functions to the original ones in the corpus, which implies denser interactions between the speaker and listeners through multifunctional head movements. In addition, we found that our CNN models outperformed support vector machine models and that the CNN models that use image-based head-pose measurements provided a comparable level of performance to that of sensor-based measurements.

The contributions of this paper are summarized as follows:

- This paper proposes a novel *functional head-movement corpus* that represents the multiple functionality of head movements in face-to-face conversations. Analyses reveal the structures of multiple head-movement functions, such as the speaker's rhythmic movements, the listener's back-channel movements, and cognitive/affective expressions.
- This paper explores the potential of using head-movement data as a rich source of information for automatically detecting multiple functions, and it shows the possibility and limitations of the automatic recognition of head-movement functions by using CNN models.

In these two respects, we believe this study will contribute to a deeper understanding of human nonverbal behaviors and the automatic recognition thereof in future applications.

This paper is organized as follows: Section 2 reviews the communicative functions of head movements, coding schemes, and the automatic recognition of head movements. Section 3 presents our functional head-movement corpus together with its definitions, the coding scheme, and corpus analyses. Section 4 describes our CNN models for the automatic recognition of head-movement functions and presents the experimental results. Section 5 discusses future work. Section 6 draws conclusions.

II. RELATED WORK

This section reviews the communicative functions of head movements, as described in the literature. Then, it reviews the annotation schemes for head movements and their functions. Finally, it surveys the trends in the automatic recognition of head movements.

A. COMMUNICATIVE FUNCTIONS OF HEAD MOVEMENTS

There have been a number of review papers surveying head movements and their functions in conversation [23]–[25], [51]–[54]. On the basis of these papers and other literature, we review the head-movement functions related to speech production, speaker cognitive expression, speaker feedback request, listener feedback and cognitive expressions, affect display, and turn management, as described below.

1) SPEECH PRODUCTION

The speaker's body movements are rhythmically coordinated with the articulated segmentation of his/her speech [36], [55], [56]. Hadar *et al.* found that the speaker's head is constantly in motion during his/her utterances, while the listener's head tends to remain stationary when listening [36]. This kind of movement is called *batonic* or *rhythmic* movement. The speaker also uses head movements to put stress on his/her speech [35], [52], [53], [57]–[59]. Rapid head movements tend to coincide with the peak loudness of speech [35]. Repeated movements can emphasize words and phrases [59]. Head shakes co-occurring with adverbs or exclamations such as "very" and "really" emphasize verbal content [52], [58]. Furthermore, head movements are used for gesticulation [21], [60], [61]. McClave *et al.* pointed out that head pose and head movement serve narrative functions, including expressing mental-image characters, making references to physical or mental space, listing alternatives, and marking direct quotes [52].

2) SPEAKER's COGNITIVE EXPRESSIONS

The speaker's head movements also indicate his/her internal cognitive state with regard to speech production. When a speaker thinks while speaking, e.g., performs a word search in his/her mind, the speaker turns his/her head away from the other participants to reduce the visual cognitive load [62]. Hesitations, which indicate the speaker is having difficulty producing speech, can be identified from irregular head movements that are different from rhythmic movements [56], [63]. Hadar *et al.* discovered that a high-frequency and low-amplitude head movement appears just after such a dysfluency period [63]. A small lateral shake of the head can indicate a speaker's uncertainty when he/she feels a piece of information is missing [52]. Repeated head movements appear when the speaker performs self-affirmation and self-reflection [59]. Lateral shaking of the head typically appears before lexical self-repair [52], [58].

3) SPEAKER's FEEDBACK REQUESTS

Feedback from listeners to the speaker, which does not signify the intent to claim a speaking turn, is known as a "back-channel" [64], [65]. Typical back-channel signals include short utterances, such as 'uh huh' and 'mm hmm', and nods. Traditionally, the back channel was considered to consist of minimal, non-interruptive, and spontaneous responses of the listener to the speaker [64], [66]. Later, it was reframed as a joint activity between the speaker and listener [52], [67]–[71]. This change in perspective was related to the advent of the concept that conversations are organized through *mutual orientation*, in which both the speaker and listener need to orient and secure each other's attention to retain their positions as speaker and listener [72]. The speaker's nods and jerks (i.e., nods starting with upward movements) of the head are used to elicit listener responses to monitor their attitudes and level of understanding of what the speaker is trying to express, and the speaker adaptively changes the course of his/her own subsequent speech accordingly [52], [70], [73]. Maynard found that nearly 40% of listener back-channels in Japanese conversations occurred in the context of the speaker's head movement [74]. In addition, to elicit listener responses, the speaker turns to or gazes at the listener [75], who looks back at the speaker, at which point *mutual gaze* occurs [76]. Such mutual gazing, called the *gaze window*, is a strong cue for eliciting responses from listeners [77], [78].

4) LISTENER's FEEDBACK AND COGNITIVE EXPRESSIONS

Different types of listener feedback involving head movements have been identified. Note that here we exclude aspects of vocal feedback such as newsmarkers (e.g., 'really?'), change-of-activity tokens (e.g., 'alright'), and assessment tokens (e.g., 'great') [79].

First, a *continuer*, as termed by Schegloff [80], is a typical feedback signal often consisting of a listener's short utterances and/or nods [64], [68]. It can be interpreted as a sign of listening with attention to the speaker or a supportive stance toward the current speaker's turn and their listener-hood. Typical *continuer* head movements can be characterized as small single-shot nods [54], [81]. Second, when the speaker elicits listener feedback as mentioned above, the listeners respond with behaviors including gazes and nods [75], [76].

Third, listeners intentionally or unintentionally exhibit their internal cognitive state in feedback signals, which are layered on one another, as suggested by Clark's four-stage model (attending, identifying, understanding, and considering) [67], [71] and Allwood's four-dimensional model (contact, perception, understanding, and attitudinal reaction) [82]. In these models, *understanding* is one of the core cognitive processes, and listeners' jerks (up-down-up nods) are often used as a feedback signal to display their understanding to the speaker [32], [34].

Fourth, head movements can indicate other cognitive states, such as thinking and uncertainty [83]. Fifth, the listener's head movements show their attitude toward what is being said by the speaker. Usually, nods are considered to indicate agreement, whereas a lateral shake of the head is considered to indicate disagreement [20], [84]. Both movements are cyclic (approximately 3 cycles) [37]. High-amplitude nods indicate affirmative responses [81], [85], [86]. Stivers *et al.* suggested that listeners' nods in 'mid-telling position' indicate their support for the speaker's stance [87]. However, the degree of kinetic arbitrariness is uncertain [20], [37], and the correspondence between movement and meaning, e.g., a nod means yes and a shake means no, is not universal and is culturally dependent [20]. Furthermore, listeners' head movements are also used for performing gesticulation [21], [60], [61].

As seen from the above, the listeners' back-channel head movements are highly multifunctional, and their timing and meaning have large variations depending on the individual and context. For example, Iwan de Kok and Heylen gave an example in which multiple listeners, who were listening to the same speaker on live video, responded in different ways depending on their individual state and motivation [88]. In addition, back-channel nods appear differently depending on language. For example, Maynard found that Japanese listeners nod nearly two times more frequently than American listeners [74].

5) AFFECT DISPLAY

In recent years, head movements have been revealed as an effective means of *affect display*. Livingstone *et al.* discovered that observers can accurately perceive the emotions (happy, sad, and neutral) of speakers and singers from only their head movements, i.e., without being able to see or hear their facial expressions or voice [89]. Adams *et al.* found that head movements carry emotional information complementary to the facial expression [90] and can be useful input for the automatic recognition of emotional categories [91]. Otsuka suggested that head movements expressed with a robotic telepresence avatar increase the observer's perception of emotion compared with a static display showing the same face image without physical movements [18].

6) TURN MANAGEMENT

Switching from the current to the next speaker is called turn-transition or turn-taking. Head movements play important roles in turn-taking. According to the turn-taking system of Sacks [92], conversation participants generally follow three rules with priority levels: i) the current speaker selects the next speaker, ii) the next speaker self-selects, and iii) the current speaker continues to speak. The turn-taking rules are implemented by exchanging verbal and nonverbal signals or cues between the speaker and listeners [53], [60], [65], [93]. A speaker's nod is considered to be a turn-yielding cue [53]. Maynard found that nods are used to mark clause boundaries and the ends of turns [53]. Hadar pointed out that a postural shift of the speaker's head indicates the completion of a sentence or clause [94]. On the other hand, a shift away from the listener's head is considered to be a turn-taking cue, and listeners' back-channel nods are interpreted as turn-avoiding cues [60], [93]. Kendon suggested [95] that a speaker gazing at the listener at the end of his/her turn is a turn-yielding cue, and a listener gazing back at the speaker creates a short mutual gaze; the speaker subsequently averting his/her gaze is a turn-taking cue. Because gaze shifts are often accompanied by head-pose shifts, head movements are also considered to be turn-yielding/taking cues.

B. CODING SCHEMES FOR HEAD MOVEMENTS

Several coding schemes and corpuses have been developed for analyzing and understanding the functions of head movements in face-to-face conversations.

Allwood *et al.* developed a multimodal coding framework called MUMIN for feedback, turn management, and sequencing functions [96]. MUMIN targets multimodal behaviors, including head movements, facial expressions, gaze direction and movements, hand gestures, and body postures. Based on Allwood's four-dimensional model mentioned above [82], the scheme codes listener feedback in terms of contact-perception-understanding (CPU) status, acceptance/non-acceptance, and attitudinal response, where the CPU status is either C-P (contact-perception) or C-P-U (contact-perception-understanding) and the response conforms to Ekman's basic emotion categories [27]. In addition, the direction of feedback, i.e., eliciting or giving feedback, and turn management (gain, hold, and end), are also annotated. The sequencing functions consist of opening, continuing, and closing speech act sequences. MUMIN has strengths in annotating multimodal expressions and in its ability to code multiple functionalities, which can include one or more of the feedback, turn management, and sequencing functions.

Upon building a MUMIN-based corpus called NOMCO, Paggio and Navarretta addressed the multifunctionality of head movements and employed an annotation procedure for resolving the simultaneous multifunctionality of head movements and other gestures. In this procedure, each coder chooses one *primary* function per gesture out of the plausible functions, and inter-coder disagreements are resolved through post-discussions, followed by a third coder's final judgment [42]. Our standpoint differs from theirs in that we do not regard the simultaneous multifunctionality and inter-coder disagreement as an *issue* to be resolved but instead regard the intrinsic nature of conversational head movements as an issue to be explored.

Poggi *et al.* targeted nods appearing in a TV broadcast of a political debate and proposed a typology of nods [24], [25]. The typology consists of six types of speaker nods, 14 types of addressee nods, and 5 types of side-participant nods. The strength of this typology is in detailing the functions of nods, such as listener agreement, approval, permission, and submission, each of which represents a different dialogue act. However, the typology does not include shakes of the head or other movements, and it lacks certain functions such as turn management. In addition, due to the nature of the TV program, the pre-edited video data contained only parts of the participants' behaviors, which made it hard to analyze the conversational interactions in the debate itself.

Head movements have been categorized on the basis of their kinetics as well as their function. Włodarczak *et al.* proposed the head gesture unit (HGU), which consists of a kinetic category ('nod,' 'jerk,' 'tilt,' 'turn,' 'protrusion,' and 'retraction'), number of cycles, duration, frequency per unit time, and complexity (number of gesture types in a single gesture phase) [32]. Kousidis *et al.* added five other categories, consisting of 'shake,' 'bobble,' 'slide,' 'shift,' and 'waggle' [33]. Buschmeier *et al.* built an Active Listening Corpus (ALICO), which employs HGUs [32] for listener expressions, Kousidis's extension [33] for speaker expressions, and vocal feedback expressions [34].

Looking more broadly than head movements, there has been several attempts to create corpuses that include diverse annotations from multiple coders as follows. Cowie *et al.* annotated the emotion categories using multiple coders, who perceived audio and/or visual cues from the interactions in TV shows [97]. They confirmed the low inter-coder agreement, which indicates the nature of real-life communication, including blended emotions and contradictory multimodal cues. Kumano *et al.* attempted to capture the diverse interpretations by multiple coders on the pairwise empathy in group meetings, which was represented as the distribution over *empathy*, *antipathy*, and *neither* categories [98]. They built a Bayes model that could predict the distribution of the interpretations from the facial expressions of meeting participants.

In the field of language processing, Passonneau *et al.* conducted a word sense annotation via cloud sourcing, which requested multiple workers to annotate multiple sense labels per word [99]. They concluded that the annotations of the cloud-sourced workers were preferable to those from a single trained annotator for building their probabilistic annotation models. To study the semantic relationship in discourse, Rohde *et al.* conducted cloud-source experiments, which asked multiple workers to answer an appropriate conjunction that could be inserted immediately before an adverbial in the given sentences [100]. They observed the diverse patterns of conjunction-adverbial combinations, which indicated multiple possible connections between the sentence and its context, and suggested the importance of different judgements of multiple annotators to fully characterize the discourse relations.

Upon the importance of different interpretations of multiple coders as we reviewed above, this paper proposes a novel functional head-movement corpus that can represent the diverse interpretations by multiple coders on multiple functionalities achieved through head movements in multiparty conversations. Because nonverbal behaviors such as head movements are an integral part of a language, our corpus and its analysis on the multifunctionalities of head-movements are expected to contribute to a deeper understanding of the real-life nature of human communication.

C. AUTOMATIC RECOGNITION OF HEAD MOVEMENTS

1) RECOGNIZING KINETIC CATEGORIES

Previous studies mainly targeted automatic recognition of nods and shakes of the head. A typical framework consists of *feature extraction* from a video sequence, followed by *classification*, which determines the category of the head movements depicted in the input video. A number of techniques have been used for feature extraction, such as facial landmark detection [101]–[103], optical flow calculation [104]–[106], appearance modeling [107], and head pose tracking [108]–[111]. Facial landmarks include the eyes and/or the midpoint between the eyes [101], [102], [112], the tip of the nose [113], and facial points detected by, e.g., scale-invariant feature transform (SIFT) [103]. Although the use of image-based tracking to measure the head pose has been considered challenging [114], it has seen significant progress thanks to the advent of depth image sensors [110], [111], [115], accurate 3-D head models [103], [115], and deep neural networks [116]. The head pose sequences obtained from the tracker are often used in further feature extraction processes such as the fast Fourier transform (FFT) [108] and discrete wavelet transform (DWT) [109]. In addition, various classifiers have been used, including those that incorporate rule-based decision making [101], hidden Markov models (HMMs) [102]–[104], [112], finite state machines [105], Bayesian networks [107], variants of conditional random field (CRF) models [117], [118], support vector machines (SVMs) [108], [109], independent component analyses (ICAs) [119], non-negative matrix factorization (NMF) [113], and deep neural networks (DNNs) [120], [121]. Relatedly, Akakın *et al.* found that fusion models, which integrate the results of multiple classifiers, can significantly improve classification performance [119].

Many recent advances have been made by DNNs [38]. In particular, the authors' group has used CNNs to detect nods from head pose sequences in multiparty conversations [120] and has found that the CNN model significantly outperformed the Wavelet+SVM model [109]. We employed this CNN model as a reference for examining the possibility of detecting head-movement functions. Sharma *et al.* recently employed a composite DNN called ConvLSTM, consisting of CNNs followed by a long short-term memory (LSTM), for classifying ten head-movement classes (nod, jerk, up, down, tick, tilt, shake, turn, forward, and backward) from head positions/poses and facial landmarks [121]. They found that a multiscale version of the ConvLSTM model with bimodal head and face inputs outperforms a single-scale model and/or single-modal input.

Another approach is to use contextual information and/or multimodal features. Morency *et al.* targeted interactions between humans and artificial agents and developed a context-based gesture recognizer that takes context cues as additional input features, such as timing (e.g., end of utterance or not), utterance type (e.g., question or not), and lexical features (e.g., ''Do you'') [122]. An example of the multimodal approach is using the speaking status, i.e., speaking or silent, of the target person together with visual motion features [123].

2) RECOGNIZING ATTITUDINAL EXPRESSIONS

Pioneering studies on the automatic recognition of attitudinal expressions from head movements include that of Bousmalis *et al.*, who attempted to automatically recognize 'agreement' and 'disagreement' from manually annotated behaviors, including nodding and shaking of the head, hand and body gestures (e.g., forefinger raise, hand wag, shoulder shrug), and audio features (e.g., pitch and energy) [85], [124]. They concluded that nods and shakes of the head are the most discriminative features and that multimodal data are more useful than single-mode data. Kaliouby and Robinson demonstrated automatic recognition of six attitude classes, i.e., 'agreeing,' 'disagreeing,' 'concentrating,' 'interested,' 'thinking,' and 'unsure', from head poses, facial landmark points, and mouth actions detected in frontal face images [83]. Although they did not target spontaneous conversations, their findings indicate that head movements are a rich source of information for recognizing human attitudes. Using similar features to those in [83], Sheerman-Chase *et al.* targeted four categories, 'thinking,' 'understanding,' 'agreeing,' and 'questioning', in spontaneous conversations and claimed that the resulting classifications were comparable to manual classifications [125]. Xiao *et al.* proposed a method for classifying 'acceptance,' 'blame,' 'positive behavior,' and 'negative behavior' in dyad conversations by modeling optical flow patterns over a person's head region in videos by using line spectral frequencies (LSFs) and a Gaussian mixture model.

3) RECOGNIZING COMMUNICATIVE FUNCTIONS

There have only been a few studies targeting automatic classification of the communicative functions of head movements in conversations. These were based on the manual annotation of kinetic categories of head movements and other gestures [40]–[42]. Jokinen *et al.* tried to classify communicative functions by performing MUMIN-based annotations on Danish and Estonian data, including head movements, facial expressions, eye movements, gaze directions, mouth openness, and lip positions [40]. The target functions were feedback (CP, CPU, accept, non-accept, elicit, emphasis), turn management (take, accept, yield, end, hold), and semiotic type (index-deictic, index-non-deictic, iconic, symbolic). They achieved approximately 60% to 80% classification accuracy. Navarretta and Paggio tried to recognize the dialogue acts of feedback expressions from head movements and facial expressions in a MUMIN-based corpus and from prosody features [41]. The dialogue acts were 'accept,' 'decline,' 'repeat rephrase,' and 'answer'. They found that a hidden Naïve Bayes classifier of head movements and facial expressions performed better than using prosodic features and achieved an f-measure of approximately 50%. Later, Paggio and Navarretta summarized their series of machine learning experiments on their NOMCO corpus for predicting communicative functions from manually annotated gestures [42]. They concluded that the best classifier for the feedback functions of head movements and facial expressions was the SVM, which achieved an F-score of 0.76 and outperformed a naïve Bayes classifier.

In contrast to the above studies, our approach does not use manual annotations of kinetic categories as its input data but rather aims to establish a link between head-movement functions and the raw data composing a head-pose sequence. Our method also employs a minimum bimodal approach using only the head pose and speaking activity, which represents the presence or absence of one's utterance without any other modality or interactional context.

III. FUNCTIONAL HEAD-MOVEMENT CORPUS

This section reviews the functional head-movement corpus that we have developed for analyzing the multiple functionality of the head movements in conversations.[1] First, we define a set of communicative functions related to head movements, followed by the coding scheme. Then, we present an analysis of multifunctionality and the kinetics of head movements.

A. DEFINITION OF FUNCTION CATEGORIES

The set of communicative functions related to head movements is shown in Table 1.[2] This set was chosen to cover as many head-movement functions as possible among those found in the past studies discussed in Section II while keeping them as simple as possible for practical coding. The table lists 32 functions, including 13 speech-related functions (`s1`∼`s13`), 12 reaction functions (`r1`∼`r12`), and seven other functions (`c1`∼`c7`), where the symbol (`s`, `r`, or `c`) + number denotes a single function category. To represent multifunctionality, the function categories are not mutually exclusive and are described by multiple binary codes, each of which indicates the presence or absence of each functionality. In theory, $2^{32} \sim 4.3$ billion different patterns can be described in our corpus. The temporal resolution of the codes is assumed to be a single video frame, e.g., 1/30 sec, which enables the analysis of fine-grained dynamics. We target head movements including vertical movements (including nods and jerks), lateral (horizontal) shakes, and tilts. Note that we do not assume any prerequisite relationship between the kinetic classes and the functions. For simplicity, we exclude lateral head turns accompanied by gaze shifts and their related functions.

The speech-related functions (`s1`∼`s13`) are head-movement functions related to speech and vocal production, which include rhythmic articulation, emphasis, hesitation, word-search, repair, and turn yielding. These are mainly assigned to the speaker, i.e., the floor holder, but the listeners' short back-channel utterances are also within the scope of these functions. The reaction functions (`r1`∼`r12`) are related to the reactions and responses to others' utterances and behaviors. These are assigned not only to the listener but also to the speaker, who reacts to the listener's feedback. We distinguish speech-related functions and reaction functions because speech production involves voluntary cognitive processing and motor activities that are different from perceiving and reacting to others' behaviors and/or utterances. Since our coding scheme is fully non-mutually exclusive, both speech-related function(s) and reaction function(s) can be simultaneously assigned to a head movement, such as the listener's nodding with a short utterance. Such flexibility is an advantage of the proposed coding scheme. The other

[1] Consult with the authors about the availability and licensing of the corpus.

[2] The supplemental movie shows example scenes for each functional category. Note that the faces in the movie are partially hidden to draw attention to the head movements rather than the facial expressions and to anonymize the people in the movie.

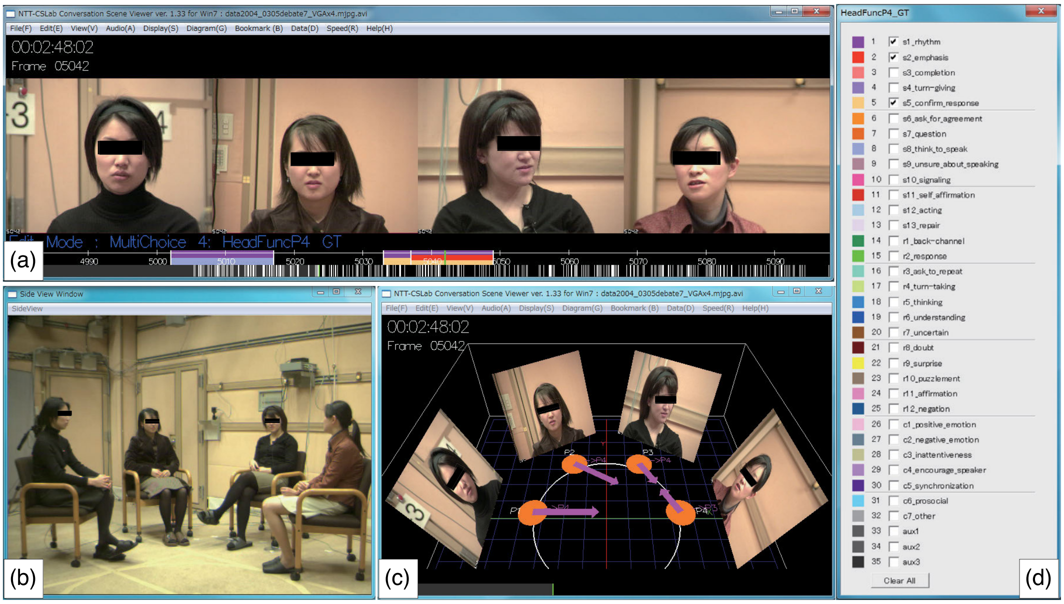
TABLE 1. Definitions of the function categories of head movements. s1∼s13 are speech-related functions, r1∼r12 are reaction functions, and c1∼c7 are other functions.

#	category	description
s1	rhythm	rhythmic movements occurring with speech production
s2	emphasis	putting stress on one's speech and/or action
s3	completion	indicating the completion of one's speech with a nod
s4	turn-yielding	attempting to give a listener the next turn
s5	confirm response	eliciting listeners' response to monitor other's attitudes/responses
s6	ask for agreement	asking for agreement in the form of a tag-question
s7	question	nodding at the end of a sentence to indicate it is a question
s8	think to speak	indicating thinking while speaking, e.g., during a word search
s9	unsure about speaking	indicating hesitation or lack of confidence
s10	signaling	attempting to attract another's attention
s11	self-affirmation	reflecting on one's own prior speech with satisfaction
s12	acting	indicating acting or gesticulation
s13	repair	repairing one's prior utterance while shaking one's head
r1	back-channel	listening with attention; a continuer, *aizuchi* in Japanese
r2	response	responding to the speaker's confirmation, question, etc.
r3	ask to repeat	asking the speaker to repeat previous speech
r4	turn-taking	accepting or attempting to take the next turn
r5	thinking	indicating thinking while listening
r6	understanding	indicating understanding of what the speaker meant to say
r7	uncertain	indicating that one could not understand well what the speaker meant to say
r8	doubt	showing doubt about what the speaker said
r9	surprise	being surprised
r10	puzzlement	being confused
r11	affirmation	replying as with 'yes', acceptance, agreement, approval, etc.
r12	negation	replying as with 'no', rejection, disagreement, disapproval, etc.
c1	positive emotion	displaying positive emotions, e.g., empathy or satisfaction
c2	negative emotion	displaying negative emotions e.g., antipathy or dissatisfaction
c3	inattentiveness	indicating boredom or lack of interest
c4	encourage speaker	encouraging the speaker or eliciting another to speak
c5	synchronization	synchronizing with the actions of another person
c6	prosocial	nodding or bowing to greet, thank, apologize, etc.
c7	other	function that does not fall into any of the above categories

functions (c1∼c7) include head-movement functions that do not fall into either the speech-related or reaction function categories.

As listed in Table 1, the speech-related functions are as follows: Rhythmic head movements naturally accompanying an utterance are labeled as *rhythm* (s1). *Emphasis* (s2) indicates head movements used to put stress on one's utterance. As a turn management function, *completion* (s3) indicates a nod used to mark the end of one's speaking turn. When the current speaker explicitly assigns the next speaker by using a head movement, e.g., a nod, the movement is labeled *turn-yielding* (s4). When a person elicits another's response by using head movements, the movement is labeled as *confirm response* (s5). When a person explicitly asks for a favorable reaction from the other(s), which typically appears at the end of a tag question, the associated head movements are labeled as *ask for agreement* (s6). In contrast, head movements appearing at the end of a question are labeled with *question* (s7), where the receiver's attitude is not assumed to be positive. As a sign of the cognitive state, *think to speak* (s8) indicates thinking while speaking, such as when the person searches his/her mind for a particular word, while *unsure about speaking* (s9) indicates a lack of confidence and hesitation in speaking. For example, s9 without s8 indicates that the person already has specific words or phrases in mind, but he/she is hesitating to vocalize them because of possible negative reactions from the receivers. As with other speech-related functions, *signaling* (s10) is used to gain another's attention, *self-affirmation* (s11) indicates reflection on one's own prior speech, *acting* (s12) indicates gesticulation involving head movements, typically appearing in a parenthetical remark, and *repair* (s13) indicates lexical self-repair with shakes of the head for denying one's previous utterance.

As listed in Table 1, the reaction functions are as follows: *Back-channel* (r1) indicates a sign of listening with attention, a continuer, which is *aizuchi* in Japanese. It includes pretending to listen. Note that the term back-channel has a narrower definition than in the literature reviewed in II-A4. *Response* (r2) indicates an explicit response to the person who elicited or asked. *Ask to repeat* (r3) is a nod used when a person fails to catch what the other has just said or a nod used as a request for clarification. *Turn-taking* (r4) is a nod or nods when a person accepts or attempts to take the next turn. Some functions are defined for displaying one's cognitive state, as follows: *Thinking* (r5) indicates that the person is thinking while listening. *Understanding* (r6) indicates that the person successfully understood what the other meant to say or intended to convey. In contrast, *uncertain* (r7) indicates that the person still does not understand what the other meant to say, and *doubt* (r8) clearly shows suspicion


NTT-CSLab Conversation Scene Viewer ver. 1.33 for Win7 : data2004_0305debate7_VGAx4.mjpg.avi
File(F) Edit(E) View(V) Audio(A) Display(S) Diagram(G) Bookmark (B) Data(D) Speed(R) Help(H)
00:02:48:02
Frame 05042
Mode : MultiChoice 4: HeadFuncP4_GT
4990
5000
5010
5020
5030
5040
5050
5060
5070
5080
5090
(a)
Side View Window
SideView
(b)
00:02:48:02
Frame 05042
P1
P2
P3
P4
(c)
HeadFuncP4_GT
Clear All
(d)

- 1 ☑ s1_rhythm
- 2 ☑ s2_emphasis
- 3 ☐ s3_completion
- 4 ☐ s4_turn-giving
- 5 ☑ s5_confirm_response
- 6 ☐ s6_ask_for_agreement
- 7 ☐ s7_question
- 8 ☐ s8_think_to_speak
- 9 ☐ s9_unsure_about_speaking
- 10 ☐ s10_signaling
- 11 ☐ s11_self_affirmation
- 12 ☐ s12_acting
- 13 ☐ s13_repair
- 14 ☐ r1_back-channel
- 15 ☐ r2_response
- 16 ☐ r3_ask_to_repeat
- 17 ☐ r4_turn-taking
- 18 ☐ r5_thinking
- 19 ☐ r6_understanding
- 20 ☐ r7_uncertain
- 21 ☐ r8_doubt
- 22 ☐ r9_surprise
- 23 ☐ r10_puzzlement
- 24 ☐ r11_affirmation
- 25 ☐ r12_negation
- 26 ☐ c1_positive_emotion
- 27 ☐ c2_negative_emotion
- 28 ☐ c3_inattentiveness
- 29 ☐ c4_encourage_speaker
- 30 ☐ c5_synchronization
- 31 ☐ c6_prosocial
- 32 ☐ c7_other
- 33 ☐ aux1
- 34 ☐ aux2
- 35 ☐ aux3

FIGURE 1. Annotation software, called the NTT-CSL Multimodal Conversation Viewer: (a) bust shot, (b) whole shot, (c) 3-D gaze view, and (d) checkbox for coding functions (the function names in English are translations from the original Japanese).

of the other's opinion. For example, r7 without r8 indicates that the person has not yet judged the other's opinion as being good or bad, while r8 without r7 indicates that the person is clearly expressing his/her negative attitude after he/she has understood the meaning of the other's utterance. *Surprise* (r9) indicates that the person is surprised, and *puzzlement* (r10) indicates confusion. In addition, *affirmation* (r11) is a positive reply such as 'yes' or one indicating acceptance, agreement, approval, and submission. *Negation* (r12) is a negative reply such as 'no' or one indicating rejection, disagreement, disapproval, and disobedience. Note that *affirmation* (r11) and *negation* (r12) do not necessarily involve emotional responses and can be simple business-like replies.

The other functions are as follows: *Positive emotion* (c1) indicates positive emotions such as empathy, satisfaction, respect, and happiness. *Negative emotion* (c2) indicates negative emotions such as antipathy, dissatisfaction, ridicule, sadness, anger, and fear. Here, we decided to put the affect display functions into the *other* category because the conversation participants express their emotions not only as reactions toward the other participants but as voluntary or involuntary actions either in speaking or silence. *Inattentiveness* (c3) indicates boredom and lack of interest. *Encourage speaker* (c4) includes head movements used to elicit another to speak, e.g., a speaker in the *unsure about speaking* (s9) state. We categorized c4 into the *other* category so that it can cover a situation in which no one is willing to take the next turn, and the participants nonverbally encourage someone else to start speaking. *Synchronization* (c5) indicates head movements for regulating joint actions with others. *Prosocial* (c6) denotes a nod or bow such as for greeting, thanking, or apologizing. Finally, *other* (c7) indicates a head-movement interval with a function that does not fit any of the above categories.

Note that our function definition does not employ Ekman and Friesen's basic emotions [27] as the affect display functions, which were employed in the MUMIN coding scheme [96] reviewed in II-B. This is because the conversation participants in our data almost never or never showed *anger*, *fear*, *sadness*, or *disgust* among the fixed basic emotions. Instead, we employ the concept of *emotional valence*, in which an emotion is perceived as either positively or negatively valenced in general [126]. For simplicity, our corpus defines two function categories, positive emotions (c1) and negative emotions (c2), each of which indicates a positive or negative domain of emotional valence, which lies in a pleasure-displeasure dimension used in such systems as the PAD (pleasure-arousal-dominance) emotional model [127] and the Circumplex model [128]. In addition, because we believe that the detailed differences in participants' reactions are important for analyzing how people express their acceptance/rejection and understanding to their partner to build mutual understanding and group consensus, we define some reaction functions, including doubt (r8), surprise (r9) and puzzlement (r10).

B. DATA AND CODING SCHEME

This study employs the NTT-CSL multimodal conversation corpus 2004 (NTT-CSL-MMCC'04) and annotation software called the NTT-CSL Multimodal Conversation Viewer, which the authors' group developed, as shown in Fig. 1. Our corpus targets four-party conversations in face-to-face settings, as shown in Fig. 1(b), and consists of conversations conducted

TABLE 2. Some statistics on our corpus. Rate indicates the percentage duration of the head movements for each function relative to the duration of all head movements in our dataset. Frequency indicates the number of head-movement intervals with consecutive functions per minute. Duration is the mean duration in seconds of a head-movement interval with consecutive functions. ICC indicates the inter-coder correlation ICC(2, *k*) among the three coders.

Functions		Rate	Frequency per minute			Duration	ICC
#	category	%	All	Speaker	Listener	sec.	$(2, k)$
s1	rhythm	24.3	6.39	20.6	1.8	.672	.614
s2	emphasis	12.5	3.84	12.9	.849	.573	.576
s3	completion	.639	.160	.564	.123	.706	.836
s4	turn-yielding	.655	.160	.413	.135	.724	.467
s5	confirm response	15.6	3.80	12.2	1.05	.724	.605
s6	ask for agreement	3.81	.996	2.25	.591	.675	.633
s7	question	1.87	.507	1.16	.308	.649	.582
s8	think to speak	12.2	2.66	7.74	1.12	.812	.680
s9	unsure about speaking	2.94	.573	1.84	.197	.907	.453
s10	signaling	.140	.047	.150	.012	.527	$< 10^{-3}$
s11	self-affirmation	2.93	.667	1.65	.406	.774	.228
s12	acting	.429	.085	.263	.025	.896	.826
s13	repair	.050	.019	.075	0.00	.467	N/A
r1	back-channel	63.3	12.1	2.33	15.2	.923	.745
r2	response	14.7	2.69	2.18	2.88	.967	.365
r3	ask to repeat	.025	.009	0.00	.012	.467	N/A
r4	turn-taking	1.06	.310	.752	.197	.606	.632
r5	thinking	17.5	2.63	.376	3.34	1.17	.361
r6	understanding	8.01	.949	.338	1.14	1.49	.489
r7	uncertain	1.85	.301	.075	.369	1.08	.228
r8	doubt	1.77	.310	.038	.394	1.01	.656
r9	surprise	.731	.188	0.00	.246	.687	.798
r10	puzzlement	2.13	.366	.150	.443	1.03	.182
r11	affirmation	9.33	1.48	.564	1.78	1.11	.462
r12	negation	.392	.075	.075	.074	.921	.610
c1	positive emotion	17.4	2.56	1.24	2.98	1.2	.586
c2	negative emotion	.867	.169	.038	.209	.906	.296
c3	inattentiveness	.122	.028	.038	.037	.767	0.00
c4	encourage speaker	4.77	.827	.827	.849	1.02	.124
c5	synchronization	.369	.056	.075	.049	1.16	.874
c6	prosocial	.662	.122	0.00	.160	.956	.317
c7	other	.298	.103	.038	.123	.509	$< 10^{-3}$

by two groups of four females. The data include two sessions for each group, four sessions in total: G1S1, G1S2, G2S1, and G2S2, where G*i* denotes group *i* and S*j* denotes session *j*. The participants were female Japanese native speakers in their 20s $\sim$ 30s. They met each other for the first time on the day of recording. In each session, they had a discussion on a controversial topic given by the experimenter and were instructed to try to reach a conclusion as a group within 5 minutes. The topics were related to love and matrimony, legislation on euthanasia, and favorable taxation for childcare. The conversations were captured with cameras, microphones, and motion capture sensors. The frontal face images of each participant were captured by separate cameras, as shown in Fig. 1(a). Lapel microphones were used to record each person's voice. A head-worn sensor (Fastrak [129]) was attached to the back of the head with a hairband. The data were all synchronized at 30 fps (frames per second). The sensor measured the 3-D coordinates and 3-degrees-of-freedom rotation angles of each person. The data lengths were 17200, 9900, 10100, and 10700 frames ($\sim$ 9.6, 5.5, 5.6, and 5.9 minutes) for G1S1, G1S2, G2S1, and G2S2, respectively.

Our functional head-movement corpus focuses only on the time intervals showing the head movements. The target movements are nod (including jerk), shake, and tilt, which were manually detected in a previous study [109]. Subtle movements at a barely noticeable level are included. In all four sessions, head movements existed for 29.4% of the total time length per person, on average. The percentages of nod, shake, and tilt are 95.8%, 0.37%, and 3.82%, respectively. The majority of the movements are categorized as nods. The frequency of the head-movement intervals, which are separated by nonmovement intervals, is 20.3 times per minute on average. The mean duration is 0.87 seconds. The total number of head movement intervals in all four sessions is 2166.

To create the functional head-movement corpus, we employed three coders who did not participate in the conversations and were Japanese females in the same age bracket as the conversation participants. They performed the coding process without direct communication. The initial and follow-up instruction was provided by the first author. The coders used the annotation software shown in Fig. 1, which displays (a) the bust shot, (b) the whole shot, and (c) the interpersonal gaze directions and timeline, as shown in Fig. 2. In Fig. 2, the timeline window shows the utterance intervals in gray and the head-movement intervals with color codes

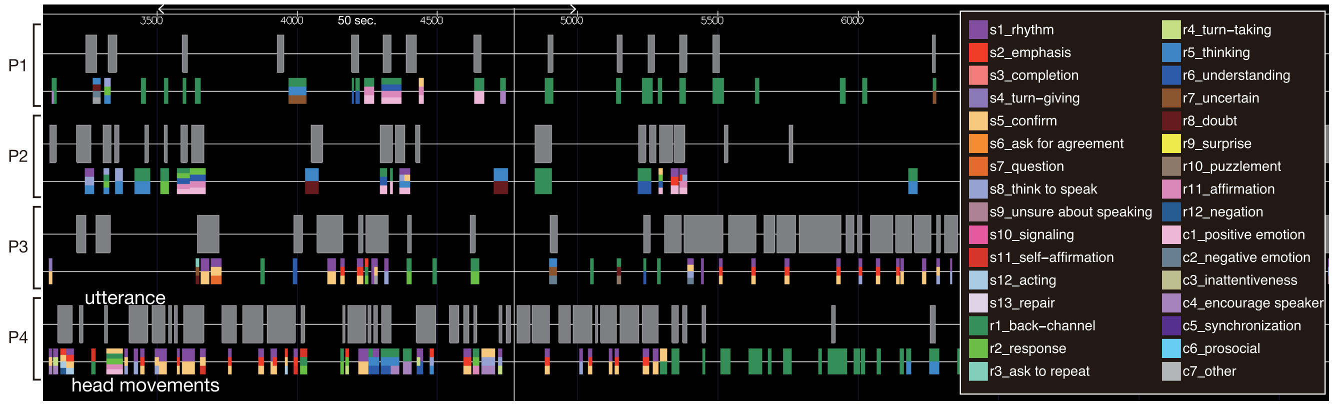


FIGURE 2. Example timeline of head-movement functionality codes. The gray bars indicate utterance intervals, and the colored bars show the functionality codes. Single-color bars indicate a single functionality, whereas multiple-color bars indicate the multiple functions during the head-movement interval.

representing multiple functions. The coders were allowed to freely play back the video, move in framewise increments/decrements at arbitrary speed, and jump to another frame by operating the keyboard, mouse, jog-shuttle, and 3-D mouse. They judged the set of functions by considering not only the head movements seen in the video but also all available information, including the voice, verbal content, gaze direction, facial expressions, body movements, and prior and posterior interactional contexts, including others' reactions. All head-movement intervals were pre-coded as being in the *other* category. Each coder judged the function set for each movement interval by clicking the checkboxes in Fig. 1(d) for as many functions as the coder recognized. If necessary, the coder could split an interval into two or more subintervals to assign different sets of functions. The annotation software itself allows frame-by-frame coding, but we employed the coarse-to-fine approach to minimize the time and effort of the manual coding process.

To minimize the chance of coders making operational errors, such as clicking the wrong checkbox, we asked them to double-check their codes. During the coding process, the supervisor, i.e., first author, was always available to answer coders' questions about the function definitions and the annotation process on demand. After that, the supervisor jointly reviewed the whole code sequence with each coder so that there were no obvious mistakes or misinterpretations of the functions left in the codes.

Last, the final functional codes were obtained by aggregating the codes from the three coders. To capture the multifunctionality as much as possible, we computed the *logical sum* of the three coders' outputs. This means that if at least one coder recognized a function, the final code includes a positive determination, even if the other two coders did not recognize it. Fig. 2 illustrates an example timeline of head functions; the color scheme shows different functions as different colors, so a bar consisting of different colors indicates multifunctionality.

C. BASIC STATISTICS OF THE FUNCTIONAL HEAD-MOVEMENT CORPUS

Table 2 shows the statistics of the corpus. The *rate* column is the time during which each function occurred as a proportion of all head-movement intervals. The frequency indicates how many times each function appeared per minute on average, depending on the participant's role. To calculate the frequency, we counted the number of time intervals in which a function continued without interruption. The *duration* is the mean duration of the interval for each function. The inter-coder correlation, ICC(2, k), shows the degree of agreement among the three coders [130]. The ICC is typically used to assess the quality of the annotations made by more than two coders. We chose to use ICC(2, k), two-way random effects, absolute agreement, and multiple raters/measurements because they indicate the mean characteristics of multiple coders who judged the same set of data in terms of absolute agreement [130]. We calculated the ICC by using the psych library in R ver. 3.4.1.

Back-channel (`r1`) was found to be the most frequent function, followed by *rhythm* (`s1`). Emphasis (`s2`), confirm response (`s5`), and think to speak (`s8`) were frequent speech-related functions, while response (`r2`) and thinking (`r5`) were frequent reaction functions. Positive emotion (`c1`) was a relatively frequent other function. The role-based frequency, which was calculated on the basis of the participant's role (here, the speaker means the floor holder and the listener means a non-speaker), indicated that the speech-related functions were mainly given to, but were not limited to, the speaker. Likewise, the reaction functions were mainly given to, but were not limited to, the listener.

Depending on the function category, the ICC scores varied widely, from zero to greater than 0.8, i.e., ranging from *poor* to *good* agreement; according to Koo's guidelines, *poor* $< 0.5 \leq$ *moderate* $< 0.75 \leq$ *good* $< 0.9 \leq$ *excellent* [130]. Higher ICC scores were obtained for completion (`s3`) = 0.836, acting (`s12`) = 0.826, and back-channel (`r1`) = 0.745. Lower ICC scores were obtained for speaker's

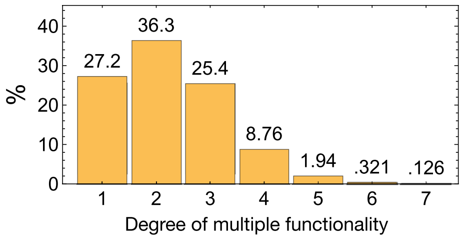


FIGURE 3. Frequency of simultaneous functions. The horizontal axis denotes the degree of multiple functionality, i.e., the number of functions that appeared simultaneously, and the vertical axis indicates the percentage of time that each degree of multifunctionality appeared relative to all head-movement intervals on a time-frame basis.

self-affirmation (`s11`) = 0.228, listener's thinking (`r5`) = 0.361, uncertain (`r7`) = 0.228, and negative emotions (`c2`) = 0.296. The functions with lower ICC scores were related to inner cognitive states, which could be difficult to infer with confidence and could be interpreted differently by different observers. Interestingly, in contrast to the *good* agreement on negation (`r12`), the lower ICC score for negative emotions (`c2`) could imply that conversation participants tried to maintain a cooperative mood and suppress expressions of negative emotion. Note that we do not consider a low inter-coder agreement to indicate low data reliability or procedural flaws in the coding process. We rather consider these ICC scores as evidence of multiple functionality. Another possible interpretation is that the inter-coder disagreement stems from the ambiguity of the words used to define the functional categories. Therefore, our functional corpus is considered to be one that jointly encodes both the multiple functionality embedded in head movements and the ambiguity of human linguistic cognition.

Other possible factors of the diverging codes include coders' operational errors and individual differences in cognitive styles. The cognitive style here refers to how the coders attend to the details of behaviors. Some coders tended to focus more on the details and detect subtle changes in the functions over time. Such a difference caused the temporal granularity of the code sequences to differ among the coders. The cognitive style of a coder may also be influenced by the operational skill with the annotation software. With an equal time limit given to all coders, lower-skilled coders might have more chances to make a mistake and might not have enough time to pay attention to the functions' fine-scale temporal changes.

D. SOME ASPECTS OF MULTIPLE FUNCTIONALITY

Here, we briefly analyze the multifunctionality found in our corpus. First, the number of different combinations of functions was 421 among 2^{32} possible combinations. The degree of multifunctionality, which is defined as the number of functions appearing at the same time, is summarized in terms of the frequency distribution shown in Fig. 3. The figure shows that two functions most often appeared at the same time, followed by a single function and three functions. No more than seven functions appeared at the same time.

On the basis of the above analysis, we examined multifunctional patterns involving two functions. Fig. 4 illustrates the conditional probability of a function (in a *column*) given another function (in a *row*), $P(column|row)$, i.e., the percentage of time a *column* function co-occurred with a *row* function when the *row* function appeared. The blue boxes *A* to *D* in Fig. 4 show a pattern of function co-occurrence. Part A shows the conditional probability of the speech rhythm (`s1`), i.e., $P(\texttt{s1}|\cdot)$. It indicates that rhythm (`s1`) frequently co-occurred with emphasis (`s2`), confirm response (`s5`), and ask for agreement (`s6`), i.e., $P(\texttt{s1}|\texttt{s2}) = 0.88$, $P(\texttt{s1}|\texttt{s5}) = 0.86$, and $P(\texttt{s1}|\texttt{s6}) = 0.64$. Part B shows that back-channel (`r1`) co-occurred with response (`r2`), thinking (`r5`), understanding (`r6`), and positive emotion (`c1`), i.e., $P(\texttt{r1}|\texttt{r2}) = 0.63$, $P(\texttt{r1}|\texttt{r5}) = 0.87$, $P(\texttt{r1}|\texttt{r6}) = 0.72$, and $P(\texttt{r1}|\texttt{c1}) = 0.60$. These high conditional probabilities indicate that rhythm (`s1`) and back-channel (`r1`) movements constitute the fundamental layers of multiple functions related to speech production and reaction.

Parts C and D in Fig. 4 show some multifunctional structures in cognitive and affective display functions. Part C reveals the relationships among thinking (`r5`), understanding (`r6`), uncertain (`r7`), and doubt (`r8`). It shows that thinking (`r5`) and understanding (`r6`) did not often co-occur, as evidenced by the low probabilities, $P(\texttt{r5}|\texttt{r6}) = 0.14$ and $P(\texttt{r6}|\texttt{r5}) = 0.07$. This indicates that thinking and understanding have distinct meanings and expressions. Instead, thinking (`r5`) often co-occurred with uncertain (`r7`) and doubt (`r8`); $P(\texttt{r5}|\texttt{r7}) = 0.52$ and $P(\texttt{r5}|\texttt{r8}) = 0.52$. Part D indicates that positive emotion (`c1`) co-occurred with affirmation (`r11`) and understanding (`r6`); $P(\texttt{c1}|\texttt{r11}) = 0.71$ and $P(\texttt{c1}|\texttt{r6}) = 0.42$. This indicates that when a person gave an affirmative response, it was often accompanied by a positive emotion. In addition, displaying one's understanding was sometimes accompanied by a positive emotion. This analysis reveals some aspects of the multifunctional structure of cognitive and affective display functions.

In addition, multifunctionality can be confirmed from the functional timeline. Fig. 2 shows that speakers (P4 in the first half and P3 in the second half) repeatedly expressed combinations of rhythm (in purple), emphasis (in red), and confirm response (in light orange) and sometimes expressed various attitudes, such as thinking (in light blue) and positive emotions (in light pink), in the response to feedback from the listeners. Listeners regularly expressed back-channels (in green) and sometimes showed understanding (in blue) and positive emotions in response to the speaker.

The above analysis revealed some aspects of the multifunctionality of head movements. Speakers can modulate rhythmic head movements to manifest various functions, such as emphasis and eliciting responses, and listeners can modulate continuous back-channel head movements to express various attitudes and responses. Various cognitive and affective display functions are densely interrelated with each other and

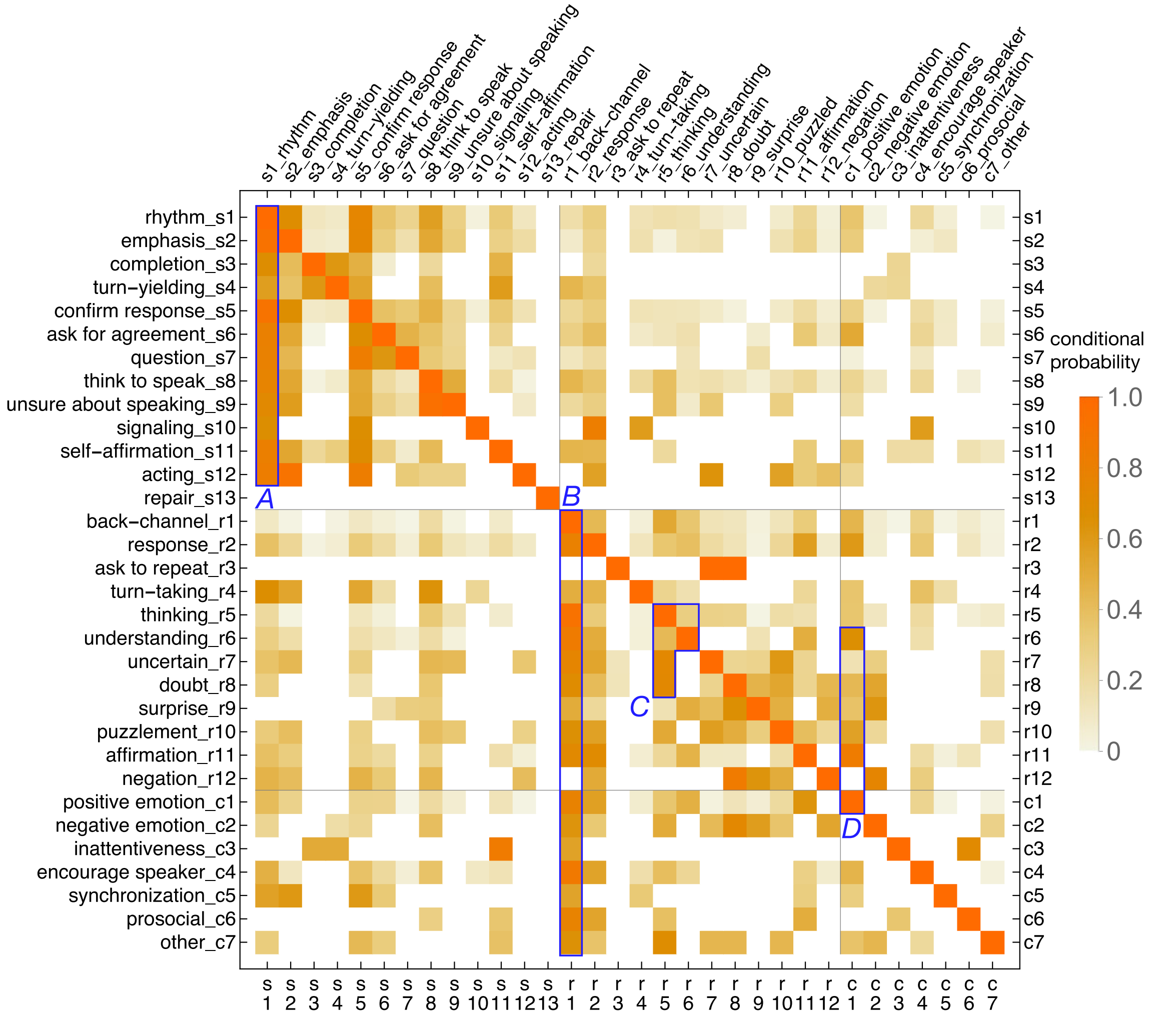


FIGURE 4. Co-occurrence matrix of two functions. The rate indicates the conditional probability of a *column* function for a given *row* function, *P*(*column*|*row*), i.e., the percentage of time a *column* function co-occurred with a *row* function. The blue boxes A, B, C, and D respectively indicate co-occurrent patterns related to the rhythm (`s1`), the back-channel (`r1`), thinking (`r5`), and positive emotions (`c1`).

form a rich functional spectrum. Although this analysis is preliminary, it illustrates that our corpus and visualization schemes, such as the conditional probability matrix (Fig. 4) and function timeline (Fig. 2), are useful resources for analyzing the deeper aspects of the multifunctionality of head movements.

E. KINETICS OF HEAD MOVEMENTS

To determine the kinetic features of head movements when communication functions emerge, the velocity profiles of head movements, measured with a head-worn sensor called Fastrak, were drawn for some function categories. Fig. 5 shows the time series of the angular velocity of the head pose angle in the nodding direction, which we call *elevation*, for up to 100 samples chosen randomly for each category. Each subplot, (a), (b), and (c), shows separate waveforms for each of two categories (blue and red lines) that do not overlap and common waveforms for both categories (black line). The intra-category variations in the velocity profiles are great, while the inter-category distinctions are ambiguous. Looking more closely, the emphasis movements (s2) can be characterized as having sharp peaks in the velocity profile, i.e., high acceleration, compared with the relatively slow velocity changes (low acceleration) in the rhythm movements (s1). As seen in Fig. 5(b), back-channel (r1) and thinking (r5) movements can be characterized as relatively low-amplitude periodic movements compared with affirmation (r11) and positive emotion (c1) in Fig. 5(c), which show highly accelerated fast downward strokes (i.e., nods).

To characterize the kinetic features of the head movements depending on the function categories, we calculated two features, *amplitude* and *frequency*, from the data. In Fig. 6, the amplitude is defined as the difference between the maximum and minimum angles in the head-pose elevation direction within a head-movement interval. The frequency is the

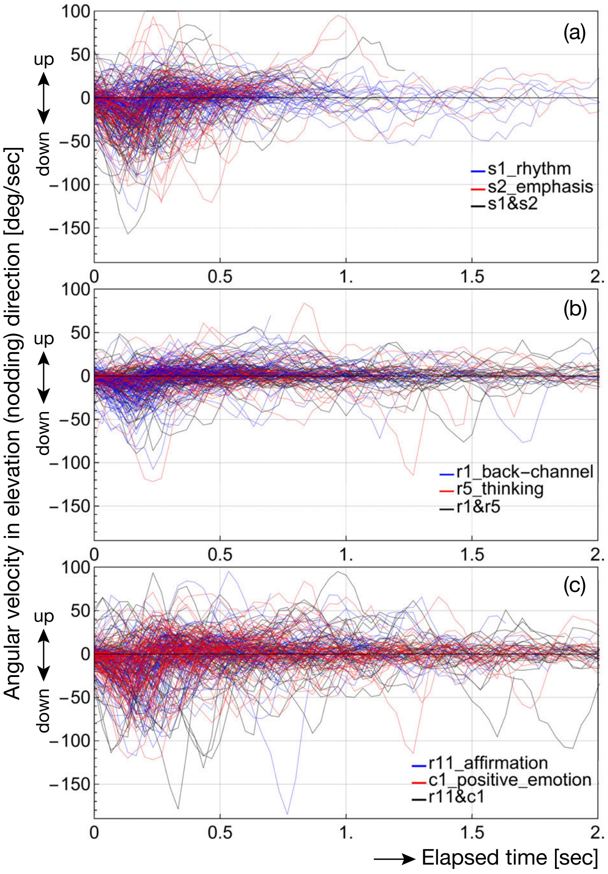


FIGURE 5. **Velocity profiles of head movements in the elevation (nodding) direction within the first 2-second interval labeled as (a) rhythm & emphasis, (b) back-channel & thinking, and (c) affirmation & positive emotion. Up to 100 samples were randomly chosen for each category (blue and red lines) and for both categories (black lines).**

number of upward and downward peaks in the head pose elevation per second, which is calculated by counting the zero-crossing points on the velocity profiles. Fig. 6 shows that each function has its own distribution range, but these substantially overlap and seem to be difficult to separate from one another, at least using these two kinetic features.

From the above observations, it seems to be difficult to manually define explicit rules or kinetic features that can distinguish between these functional categories because of the diversity and complexity of the head movements.

IV. AUTOMATIC DETECTION OF HEAD-MOVEMENT FUNCTIONS

This section examines the possibility of the automatic classification of head-movement functions in light of our hypothesis that head-movement data are a rich source of information for providing and distinguishing multiple communicative functions. Among the 32 functions defined in Table 1, we mainly targeted the ten most frequent functions, i.e., rhythm (s1), emphasis (s2), confirm response (s5), think to speak (s8), back-channel (r1), response (r2), thinking (r5), understanding (r6), affirmation (r11), and positive emotion (c1), in IV-D1, and two examples, back-channel (r1) and positive emotion (c1), are discussed in the following subsections. This section first describes our classification model structures based on convolutional neural networks. Then, it describes the parameter settings and experimental conditions, followed by the validation schemes. After that, we show the experimental results and discuss them from several performance perspectives by making comparisons with different sensing methods, classifiers, and CNN structures.

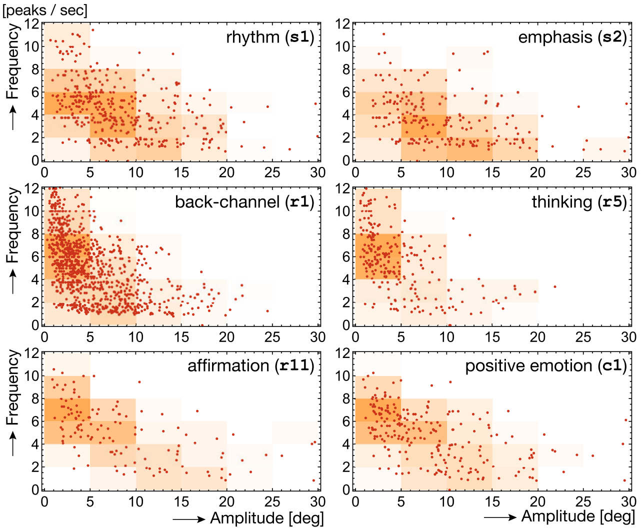


FIGURE 6. **Two kinetic features of head movements, *amplitude* vs. *frequency*, are shown for six function categories in the form of scatter plots. The amplitude is the (maximum − minimum) angle in the elevation (nodding) direction. The frequency is the number of peaks in the head-pose elevation per second. Each red point corresponds to a movement interval that lasted more than 0.5 seconds. A density histogram of the distribution is shown in the background with gradation colors.**

Our approach differs from those of past studies on the automatic recognition of attitudinal expressions II-C2 and on communicative function recognition II-C3. First, we focus on the multifunctional aspects of head movements and formulate the recognition problem as *multiple binary* classification, not the *multiclass* classifications of the previous studies [40], [41], [83], [124], [125]. Second, to investigate the potential uses of head-movement data, we take a *minimum bimodal* approach, not the fully multimodal approach on which the past studies relied. Note that the multimodal approach has been employed on facial expressions and action units (AUs) [40], [41], [125], facial landmarks [83], gaze directions [40], [124], [125], hand gestures [124], body postures [124], and vocal prosody [41] in addition to head movements.

A. MODEL STRUCTURES OF CNNs

We built a binary classifier for each head-movement function. Each classifier was built independently by using supervised learning in which the functional head-movement corpus presented in III was used as the training data. The input data consisted of windowed sequences of head-pose angular velocity and utterance features. The head pose was represented by

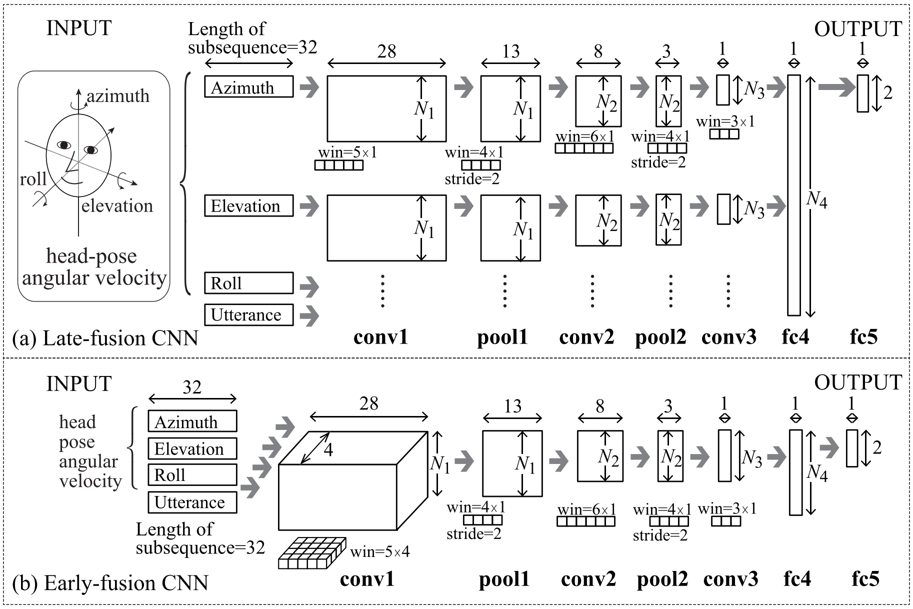


FIGURE 7. Convolutional neural network (CNN) structures for detecting communicative head-movement functions: (a) late-fusion model; (b) early-fusion model. This figure is a modification of the author's previous models in [131].

three rotational angles, called *azimuth*, *elevation*, and *roll*, as illustrated in Fig. 7, where *azimuth* represents the lateral shake direction, *elevation* indicates the nodding direction, and *roll* corresponds to the tilting direction. The angular velocity of each rotational angle was calculated by differentiating values between two adjacent time frames. The angular velocity was normalized so that the CNN could input bounded values within the range $[-1, 1]$ by using a symmetric saturating linear transfer function that limits the velocity to $[-v_{th}, v_{th}]$ and normalizing by dividing by v_{th}, where v_{th} is the threshold velocity. The utterance feature was a binary-valued feature that represents the presence or absence of an utterance in each time frame. Temporal smoothing using a moving average filter was applied to the binary sequence of utterances so that the CNNs could take continuous-valued inputs. The output of the CNNs was a binary decision, i.e., the presence or absence of a specific function, at the time frame located at the center of the temporal input window.

Fig. 7 shows the structures of the proposed CNN models: one is called the late-fusion model (Fig. 7(a)), and the other is called the early-fusion model (Fig. 7(b)). Both models consist of a convolution layer (**conv1**) and pooling layer (**pool1**), a second convolution layer (**conv2**) and pooling layer (**pool2**), a third convolution layer (**conv3**), a fully connected layer (**fc4**), and a final fully connected layer (**fc5**). The activation function in the **conv***l* layers is ReLU, and maxpooling is used in the pooling layers **pool***l*. The softmax function is used in **fc5** to determine the presence or absence of each function. In the late-fusion model in Fig. 7(a), **conv1** to **conv3** process each feature component separately. The **fc4** layer integrates multiple streams of features, which are passed to the final decision in **fc5**. In contrast, the early-fusion model in Fig. 7(b) concatenates multiple features in the first **conv1** layer.

CNNs of this type were first developed for HAR, which aimed at recognizing daily indoor activities by using accelerometers worn on the body [45]. With their success in HAR tasks, the authors' group applied them to nod detection and found that they outperformed the traditional SVM classifier [120]. The authors' group then used them for gaze-direction estimation in multiparty meetings [131] and found that the late-fusion models slightly outperformed the early-fusion models [131]. On the basis of these findings, we decided to use CNNs as a baseline model with sufficient performance for investigating the multiple functionalities of head movement. Note that this paper is intended to offer a comprehensive description and evaluation of CNN models for head-movement detection and is a large expansion of our previous brief report [120]. Hereafter, we employ the late-fusion model as the baseline model; we compare it with the early-fusion model later in IV-D5.

B. EXPERIMENTAL SETTINGS

1) INPUT DATA

We examined two different methodologies to measure the head pose: a head-worn sensor (Polhemus *Fastrak* [129]) and an image-based tracker (*OpenFace* [116], [132]). As mentioned briefly in III-B, *Fastrak* is a magnetic-based motion capture sensor that can measure the 3-D position and 3-degrees-of-freedom rotational angles relative to reference world coordinates. The time series of each person's head pose components, azimuth, elevation, and roll, which were recorded at 30 Hz, were converted into angular velocity components by temporal differencing, which makes our classifiers invariant to the absolute position and direction. The strengths of *Fastrak* are its robustness and accuracy (it does not accumulate errors) as well as freedom from the optical occlusion problem of vision-based sensing. The experiments in the following sections were based on *Fastrak* data unless otherwise noted. In addition, *OpenFace* (ver. 0.4.0) was used for comparison. The velocity threshold v_{th} used for the normalization was set to 120 degrees per second for all participants.

From the audio signal of the lapel microphone that each participant wore, the utterance intervals were manually determined by a single labeler at a 30 fps resolution, synchronized with the head pose and video. Here, we employed an interpausal unit (IPU); each utterance interval was separated by at least 200 msec silence. Note that the utterance intervals can be detected automatically with the voice activity detection (VAD) techniques used in the authors' previous study [131]. In [131], we confirmed that CNN models with VAD-based utterance input performed comparably to CNN models with manually detected utterances in the context of gaze direction estimation. Thus, we believe that the results presented in this paper are valid when VAD data are used instead of manually detected data. The length of the moving average filter, which was applied to the binary utterance sequence, was 101 frames.

2) IMPLEMENTATION OF CNN MODELS

The implementation of our CNN models was based on MatConvNet ver. 1.0-beta23 [133], [134]. From a preliminary analysis, we determined the common parameters for all CNN models and all function categories as follows: The width of the convolution filter kernels was set to 5, 6, and 3 for **conv1**, **conv2**, and **conv3**, while the width of pooling was set to 4 for both **pool1** and **pool2** with stride = 2. Regarding the parameters used for training, a mini-batch of size 12 was used, and the number of epochs was set to 4. A log loss function was used. As seen in Table 2, some function categories exhibited imbalances in the data; e.g., the positive data (presence of a function) were far fewer than the negative data (absence of a function). To address the imbalanced data problem, we employed class-aware sampling, which randomly samples positive and negative data to equalize the sample sizes of the classes, in each mini-batch [135]. It is known that imbalanced data degrades the performance of CNNs [136]. The number of mini-batches per epoch was chosen such that the total number of samples was equal to half the number of training data.

3) OPTIMIZING THE CNN PARAMETERS

The optimum number of units in each CNN layer, $\hat{N}_l, l = 1, \cdots, 4$, were searched for each function category. Here, we employed a *grid greedy* search, which can alleviate the computational burden entailed by a large search space [131]. The first stage searched for the best $\hat{N}_1$ and $\hat{N}_4$ in terms of the F-score (the details are given later) for fixed initial values, $N_2 = 200$ and $N_3 = 100$, which were found in a preliminary study. The default search space was set to $N_1 \in \{150, 200, \cdots, 400\}$ and $N_4 \in \{100, 150, \cdots, 300\}$. If the optimal $(\hat{N}_1, \hat{N}_4)$ was found on the border of the search space, the search space was repeatedly expanded by 50 until the optimum fell inside the search space. The second stage searched for the optimal $(\hat{N}_2, \hat{N}_3)$ for $(\hat{N}_1, \hat{N}_4)$. The default search space was set to $N_2 \in \{50, 100, \cdots, 250\}$ and $N_3 \in \{50, 100, \cdots, 250\}$. The search space was expanded in the same manner as in the first stage. The reason for prioritizing $(\hat{N}_1, \hat{N}_4)$ over $(\hat{N}_2, \hat{N}_3)$ was that the first convolution layer, **conv1**, and the fully connected layer, **fc4**, may play crucial roles in processing the raw data and in integrating multichannel and multimodal features, respectively.

C. TRAINING AND VALIDATION

1) CROSS-VALIDATION METHOD

This study targeted the head-movement intervals in our functional head-movement corpus, which occupied approximately 30% of the conversation periods, as described in III-B, and it evaluated the performance of our classifiers separately for each function category in each time frame within the head-movement intervals. To train and evaluate the classifiers, we employed a cross-validation scheme, *leave one person in one session out*. Separate classifiers were built for each person in each conversation session from the training data, which consisted of the data of the other three people in the same session and all the data from the other three sessions. For example, a model for G1S1-P1 used training data consisting of G1S1-P2∼P4, G1S2-P1∼P4, G2S1-P1∼P4, and G2S2-P1∼P4, where Pi denotes the i-th person in a session. Therefore, 16 models in total were created for each function. The advantage of this cross-validation scheme is that it can make the most use of the available data and achieve a reasonable balance between generalization and individualization; i.e., the majority of the data (approx. 14 out of 15) was from other people, but individual characteristics could still be captured from the remaining data (1 out of 15) to some extent.

As a performance measure, to alleviate data imbalances in individuals and groups, we employed *micro* measures, including micro-*accuracy*, micro-*precision*, micro-*recall*, and micro-F-score. Here, *micro* means that framewise results from all 16 models were aggregated to calculate these statistics, in contrast to *macro* measures, e.g., averages of the individual scores. *Accuracy* is the ratio of correct answers to all data, where correct answers include both positive and negative classes. *Precision* is the ratio of true-positive cases to all positive predictions. *Recall* is the ratio of true-positive cases to all positive cases in the ground truth. The F-score is the harmonic mean of *precision* and *recall*.

2) SVM MODEL

For the performance comparison, we employed a classifier using a DWT and SVM, which were first used for head

TABLE 3. Summary of the classification performance on the ten most frequent function categories. All performance measures are *micro* measures. Acc., Prec., and Rec. denote accuracy, precision, and recall, respectively.

Functions	Baseline	CNN models (late-fusion)					Human performance		
	Acc.	Acc.	Prec.	Rec.	F-score	Cohen's κ	Prec.	Rec.	F-score
s1_rhythm	.679	.868	.673	.888	.766	.676	.627	.395	.485
s2_emphasis	.750	.842	.433	.867	.577	.493	.561	.340	.423
s5_confirm_response	.702	.814	.450	.859	.590	.485	.554	.342	.423
s8_think_to_speak	.701	.750	.286	.702	.407	.282	.614	.395	.481
r1_back-channel	.610	.857	.902	.869	.885	.697	.863	.669	.754
r2_response	.631	.668	.249	.625	.356	.185	.328	.182	.234
r5_thinking	.597	.679	.319	.735	.445	.266	.347	.193	.248
r6_understanding	.755	.768	.191	.581	.287	.189	.440	.253	.322
r11_affirmation	.708	.728	.193	.606	.293	.177	.404	.231	.294
c1_positive_emotion	.622	.700	.333	.721	.456	.285	.579	.357	.442

gesture recognition in [109]; here, we simply call it the SVM model. First, to obtain the feature vectors, wavelet decompositions using the Daubechies wavelet of order 10 (db10) were conducted up to level 4 for a temporal windowed subsequence (32 frames long) of the head-pose angular velocity components, azimuth, elevation, and roll directions. Then, the maximum, minimum, mean, and standard deviation of the DWT coefficients, called *details*, D2, D3, and D4, were calculated at multiple resolution bands. The resulting feature vector was 36 dimensional, i.e., (3 rotations) × (4 statistics) × (3 wavelet bands). In addition, a temporally smoothed utterance subsequence, similar to the one used in the CNN model, was added to the feature vector. To calculate the wavelet coefficients, we used the MATLAB Wavelet Toolbox R2017b. Our SVM employed a radial basis function (RBF) kernel and a soft margin criterion. To implement the SVM, we employed LIBSVM [137]. The optimum parameters, C and γ, were found by performing a grid search over $C \in \{2^{-11}, 2^{-9}, \cdots, 2^{5}\}$ and $\gamma \in \{2^{-11}, 2^{-9}, \cdots, 2^{3}\}$. To balance the positive and negative samples, we used cost-sensitive learning; i.e., we gave different weights to the cost functions depending on the size of the samples [138].

The CNN model and SVM model are similar in their use of *convolution*. In the SVM model, convolutions with wavelet kernels can detect head movements with various kinetics at multiple temporal scales, e.g., from slow nods to fast nods, due to cascading convolutions. In addition, calculating statistics over the wavelet coefficients is similar in effect to maxpooling in the CNN. However, the CNN model has overwhelmingly better representation power due to its large number of multilayered adaptable convolution kernels.

D. EXPERIMENTAL RESULTS

Next, we describe the results of the automatic detection of head-movement functions and discuss the potential of our classifier and head-movement data.

1) PERFORMANCE ON THE TEN MOST FREQUENT FUNCTIONS

Table 3 summarizes the classification performance of our CNN models (late fusion) for the ten most frequent function categories, which were obtained with the best parameters $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$ shown in Table 4. In Table 3, the baseline accuracy is that obtained from a virtual classifier, which always outputs the majority class in the ground-truth data. The results in the table indicate that the accuracy of the CNN models is the highest for all functions. The F-scores vary greatly, from less than 0.3 to close to 0.9 depending on the function category. Higher F-scores are given to the rhythm (s1) and back-channel (r1). Lower F-scores are mainly given to the functions related to cognitive processing, e.g., think to speak (s8) and understanding (r6), and to some reactions, e.g., response (r2) and affirmation (r11). The precision scores were much lower than the recall scores, except with back-channel (r1). The lower precision means that the CNN classifiers tended to detect more head-movement functions than appeared in the corpus data.

TABLE 4. Best CNN parameters $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$ found by a greedy grid search for each function category.

Functions	$(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$
s1_rhythm	(200, 200, 100, 200)
s2_emphasis	(250, 200, 100, 250)
s5_confirm_response	(250, 200, 100, 300)
s8_think_to_speak	(300, 200, 150, 250)
r1_back-channel	(200, 200, 100, 100)
r2_response	(350, 200, 50, 250)
r5_thinking	(150, 100, 200, 100)
r6_understanding	(200, 50, 50, 200)
r11_affirmation	(200, 200, 50, 100)
c1_positive_emotion	(250, 200, 100, 200)

As another evaluation measure, Table 3 also shows Cohen's kappa coefficient, κ, which measures how well the prediction of the CNN matched the corpus data. Cohen's kappa is commonly used to measure the level of agreement between two coders while taking chance agreement into account. According to Landis's guidelines [139], $0.0 \leq$ *slight* $\leq 0.2 <$ *fair* $\leq 0.4 <$ *moderate* $\leq 0.6 <$ *substantial* $\leq 0.8 <$ *almost perfect* ≤ 1.0; the CNN model's performance varies greatly, from a *slight* matching in response (r2), understanding (r6), and affirmation (r11) to a *substantial* matching in rhythm (s1) and back-channel (r1). The trend of κ roughly matches that of the F-score.

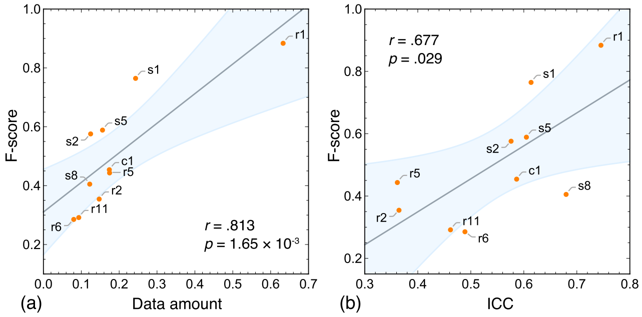


FIGURE 8. Performance characteristics: (a) correlation between the amount of data and F-score; (b) correlation between the ICC and F-score. The orange dots represent the relations between the amount of data and F-score in (a) and the relations between the ICC and F-score in (b), where the F-scores are from Table 3 and the amount of data and ICC are from Table 2. Pearson's correlation coefficient *r* and the p-value *p* are also indicated. The light-blue regions indicate the mean prediction bands at the 95% confidence level, which show that 95% of new data in the future will fall into these regions.

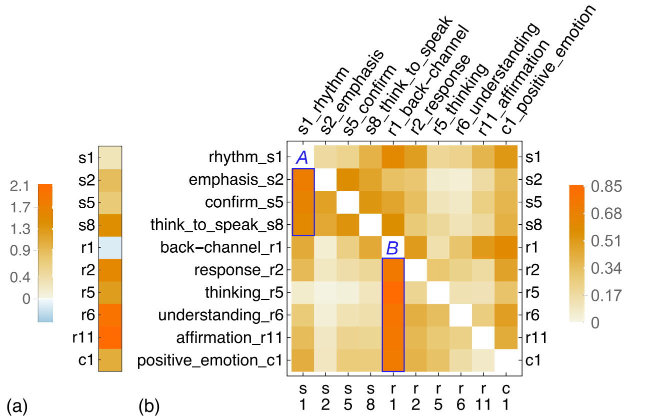


FIGURE 9. Trend of the overdetection of functions: (a) ratio of overdetection; (b) conditional probability of the *column* function in the corpus for a given overdetected *row* function that was not in the corpus. Part *A* indicates that the overdetected speech-related functions (s1∼s8) overlap rhythm (s1) in the corpus. Part *B* indicates that the overdetected reaction functions (r2∼r11 and c1) overlap back-channel (r1) in the corpus.

For reference, the table also shows virtual human performance values when a coder tries to predict the logical sum of the other two coders' results. Although this does not constitute a rigorous comparison, the human performance trends roughly match those of the CNN models. The low scores of the human coders indicate significant diversity in interpretation due to the multifunctional nature of the head movements. Looking more closely, we can see that the manual recall rates are much lower than those of the CNN models. This indicates that the CNN models were more sensitive than a single human coder to subtle expressions because they were learned with logical-sum-based aggregated data from the three coders.

Let us take a closer look at the variation in the detection performance. In particular, let us investigate the relationships between the amount of data and the F-score and between the ICC and the F-score, as shown in Fig. 8 (a) and (b). Note that the amount of data is relative to all head-movement intervals. Fig. 8 (a) shows that there is a positive correlation between the amount of data and the F-score. This indicates that the prediction performance improves when more data are available. Fig. 8 (b) shows that there is a positive correlation between the ICC and the F-score. This indicates that the more the human coders agreed on the interpretation, the better the performance was. Note that there is no significant correlation between the amount of data and the ICC value; here, Pearson's correlation coefficient is 0.56, which does not show statistical significance at the 5% level in a Spearman rank test.

These results partially support our hypothesis when a function category has a large amount of data and the inter-coder agreement is high. However, the current CNN models did not fare well in other cases, i.e., when there was a small amount of data and large variation in the coders' interpretations.

2) MULTIFUNCTIONAL ASPECTS OF THE DETECTION RESULTS

To examine the trend in overdetection suggested by the low precision rate, we tried to locate where the overdetections occurred. Fig. 9(a) shows the overdetection ratio of each function category. It shows that all categories except back-channel (r1) were overdetected compared with their lengths in the corpus. Understanding (r6) and affirmation (r11) increased by more than 200%. Fig. 9(b) shows the conditional probability of functions (in a *column*) for a given overdetected function (in a *row*) that does *not* exist in the corpus. This plot indicates which functions in the corpus co-occur with overdetected functions. Part *A* indicates that overdetected speech-related functions such as emphasis (s2), confirm response (s5), and think to speak (s8) overlap rhythm (s1) in the corpus. Part *B* indicates that overdetected functions such as response (r2), thinking (r5), understanding (r6), affirmation (r11), and positive emotion (c1) overlap back-channel (r1). These results suggest that the CNN models tended to add more functions to the rhythm and back-channel movements. Moreover, the figure indicates that the speech-related functions widely overlap the reaction functions. In addition, the CNN output added more reaction functions along with speech-related functions.

3) QUALITATIVE ANALYSIS

To assess the trend of the CNN model output, Fig. 10 illustrates the timeline of the detected functions compared with the functions in the corpus, which were used as the ground truth for the evaluation. Note that this scene was the last part of a discussion where all persons were about to reach an agreement, and it was very interactive. Fig. 10 shows that the CNN model detected more functions simultaneously than were in the corpus, and its temporal granularity was much finer than that of the manual code, which typically remained constant over each single head-movement interval. More specifically, Fig. 10 indicates that the CNN output replicated roughly the same functionalities, e.g., rhythm (s1), back-channel (r1), and response (r2), as in the corpus but also added more functions, including think to speak (s8),

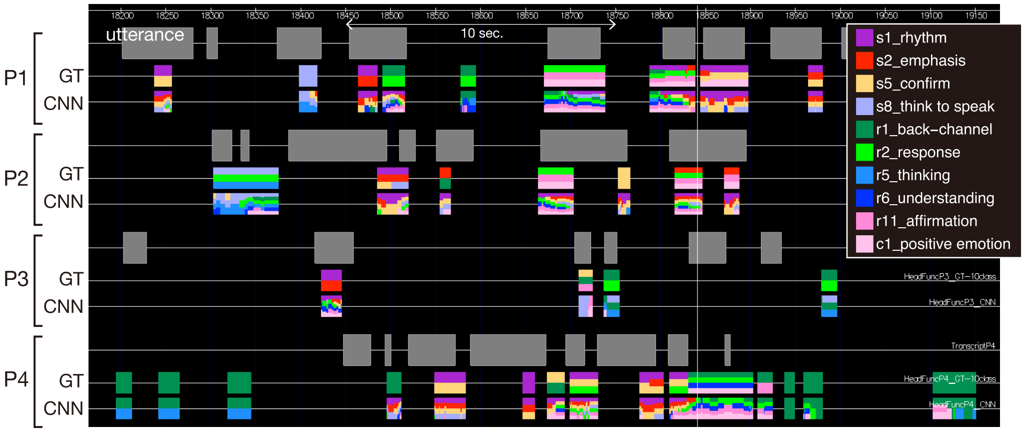


FIGURE 10. Example timeline of the detected functions compared with corpus data. CNN denotes the detection results, and GT denotes the corpus data. Multiple functions are illustrated using the color code. The gray bars indicate the utterance intervals.

TABLE 5. Performance comparison between the *OpenFace* image-based tracker and *Fastrak* head-worn sensor. *OF* denotes *OpenFace*. Acc., prec., rec., and F denote the accuracy, precision, recall, and F-score, respectively. The results of *Fastrak* are excerpted from Table 3. ∧ and ∨ indicate the best of the compared methods.

Sensors	r1(back-channel)				c1(positive emotion)			
	acc.	prec.	rec.	F	acc.	prec.	rec.	F
OF	.851	.895	.866	.880	.727	.350	.662	.458
	∧	∧	∧	∧	∨	∨	∧	∨
Fastrak	.857	.902	.869	.885	.700	.333	.721	.456

thinking while listening (r5), understanding (r6), and positive emotions (c1).

4) IMAGE-BASED TRACKER VS. HEAD-WORN SENSOR

We also compared the CNN models that input head-movement data obtained by a *Fastrak* sensor with the CNNs that input head-movement data obtained by an image-based tracker, *OpenFace*. Table 5 focuses only on back-channel (r1) and positive emotion (c1) among the ten most frequent functions; these can be considered representative functions that exhibit high and middle-level F-scores, respectively. Table 5 indicates that the CNN models with these sensing methods have comparable performances. This result replicates those of our previous study using CNNs for gaze estimation [131].

5) EARLY FUSION VS. LATE FUSION

Table 6 compares the performance of the *early*-fusion and *late*-fusion models. By using the same greedy grid strategy as in IV-B3, the optimal numbers of units $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$ in the *early*-fusion model were found to be (350, 200, 100, 200) for back-channel (r1) and (700, 200, 100, 200) for positive emotion (c1). Table 6 indicates that the models had comparable performance, although the *late*-fusion model slightly outperformed the *early*-fusion model in the back-channel category (r1) and the *early*-fusion model slightly outperformed the *late*-fusion model in the positive emotion category (c1). For the other frequent functions, overall, the *late*-fusion model slightly outperformed the *early*-fusion model; the *late*-fusion model provided F-scores that were 0.30 points higher on average.

TABLE 6. Performance comparison between *early*-fusion and *late*-fusion CNN models. Acc., prec., rec., and F denote the accuracy, precision, recall, and F-score, respectively. The results of the *late*-fusion model are taken from Table 3.

model	r1(back-channel)				c1(positive emotion)			
	acc.	prec.	rec.	F	acc.	prec.	rec.	F
Early	.852	.895	.868	.881	.719	.345	.683	.458
	∧	∧	∧	∧	∨	∨	∨	∨
Late	.857	.902	.869	.885	.700	.333	.721	.456

6) SVM MODEL VS. CNN MODELS

Table 7 compares the performance of the SVM model and CNN model. It shows that both models performed comparably for back-channel (r1), but the CNN model clearly outperformed the SVM for positive emotion (c1) by approximately 5 points in terms of the F-score. The SVM's best parameters obtained by a grid search were $(C, \gamma) = (2^{-9}, 2^{-9})$ for r1 and $(C, \gamma) = (2^{-9}, 2^{-3})$ for c1. For other frequent functions, the CNN models generally yielded higher F-scores (as high as 6.0 points) than those of the SVM models. These results suggest the superiority of the CNN models over the SVM models, especially when they target functions that are *difficult* in terms of low inter-coder agreement and a small amount of data.

7) DETECTING HEAD MOVEMENTS

The CNN models are designed for pre-detected head-movement intervals. For fully automatic head-function detection, it is necessary to detect head-movement intervals over entire conversations. Here, we verify the potential of head-movement detection using a CNN model having the same structure as that used for head-function detection. We trained and evaluated a CNN model (late fusion)

TABLE 7. Performance comparison between the SVM model and CNN model (*late* fusion). Acc., prec., rec., and F denote the accuracy, precision, recall, and F-score, respectively. The results of the CNN model are taken from Table 3.

model	`r1`(back-channel)				`c1`(positive emotion)			
	acc.	prec.	rec.	F	acc.	prec.	rec.	F
SVM	.846	.861	.903	.882	.612	.274	.744	.401
	$\wedge$	$\wedge$	$\vee$	$\wedge$	$\wedge$	$\wedge$	$\vee$	$\wedge$
CNN	.857	.902	.869	.885	.700	.333	.721	.456

TABLE 8. Performance of head-movement detection.

	Accuracy	Precision	Recall	F-score	Cohen's κ
CNN	.844	.842	.693	.760	.646
	$\vee$	$\vee$	$\vee$	$\vee$	$\vee$
SVM	.746	.826	.546	.657	.469

and SVM model using the same optimization strategy and cross-validation schemes as in the previous experiments. The number of units used in the CNN model was $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4) = (50, 600, 300, 100)$. Table 8 compares the performance of these models; the CNN model achieved a moderately high detection performance, as suggested by the F-score and Cohen's κ at the *substantial* level, and it clearly outperformed the SVM models. This indicates that fully automatic recognition would be possible by cascading the movement-detection model with the function-recognition models.

E. COMPARING DIFFERENT CODE AGGREGATION SCHEMES

The proposed multifunctional head-movement corpus employed the logical-sum-based code aggregation scheme (here denoted as OR) of the three coders' outputs to capture the diversity in the observers' interpretations. However, this scheme is highly susceptible to outlier codes introduced by such operational errors. To confirm the plausibility of this aggregation scheme and the code based on it, we compared it with two different code aggregation schemes, *majority vote* and *logical product*. The majority vote scheme (denoted as Maj. Vot.) is that if and only if two or more coders agreed on the existence of a function, the final code becomes positive. Otherwise, a negative decision is given to the final code. The logical product scheme (denoted as AND) is that if and only if all three coders agreed on the existence of a function, the final code is positive. Intuitively, the more coders agreed on a decision, the more reliable and consistent the final code is. Therefore, the Maj. Vot. and AND schemes can provide more reliable codes than the OR scheme, and they can hypothetically lead to better detection performance.

Table 9 compares the detection performance of the three code aggregation schemes, OR, Maj.Vot., and AND, on two functions, `r1`(back-channel) and `c1`(positive emotion). Using the Maj. Vot.-based corpus and AND-based corpus, the late-fusion CNN models were built using the parameters in Table 4. Table 9 indicates that the Maj. Vot. and AND schemes result in lower performance (approx. 2∼4 points lower in terms of the F-score) on `r1` detection and greatly decreased performance (as much as 14∼37 points lower in terms of the F-score) on `c1` detection. In Table 9, *Rate* shows the percentage of the total length of frames with positive `r1` or `c1` labels relative to the total length of head-movement intervals. Table 9 shows that the Maj. Vot. and AND schemes largely limited the number of positive samples. When using the AND scheme, the data reduction rate relative to the original OR-based data was 61% in the `r1` case and 91.5% in the `c1` case. This indicates that the coders have more diversity in their emotion interpretations (e.g., `c1`) than in their behavior interpretations (e.g., `r1`).

TABLE 9. Comparison of the detection performance on `r1`(back-channel) and `c1`(positive emotion) using different code aggregation schemes: OR, Maj. Vot., and AND. OR denotes the logical sum (the performance values were taken from Table 3). Maj. Vot. denotes the majority voting scheme. AND denotes a logical product scheme. Rate (%) denotes the data amount, i.e., the ratio of the total length of frames labeled with `r1` or `c1` relative to the total length of the head-movement intervals.

	Scheme	Rate	Accuracy	Precision	Recall	F-score
`r1`	OR	63.3	.857	.902	.869	.885
	Maj. Vot.	45.1	.836	.849	.844	.846
	AND	24.7	.827	.965	.778	.861
`c1`	OR	17.4	.700	.333	.721	.456
	Maj. Vot.	6.66	.816	.222	.530	.313
	AND	1.48	.952	.206	.051	.082

A plausible reason for the lower performance of the Maj. Vot. and AND schemes may be the smaller amount of positive samples, which lead to poor model generalization, as discussed in IV-D1. Although the F-score of the `c1` case dropped greatly, the F-score of `r1` remained closer to that of the OR scheme. Interestingly, in `r1`, the AND scheme outperformed the Maj. Vot. scheme in terms of the F-score and precision. These results suggest that higher inter-coder agreement could be a positive factor in the detection performance in some cases. However, such a factor is not necessary to counter the adverse effect caused by the small amount of data, as seen in the low F-scores in the `c1` case.

In summary, for the given conditions in terms of the amount of conversation data and the number of coders in our present study, the proposed logical-sum-based aggregation was an appropriate choice because it can retain the diversity in the coders' interpretations and the higher detection performance compared with other data aggregation schemes, which enforce the inter-coder agreement in the code. To retain both the coders' diverse interpretations and the reliability of the corpus, as future work, it would be important to increase the amount of conversation data and the number of coders because more data leads to better generalization by the CNN models, and more coders would enable the removal of outlier codes by using majority voting.

F. SUMMARY AND DISCUSSION OF THE AUTOMATIC DETECTION OF HEAD-MOVEMENT FUNCTIONS

The results of the experiments reveal the potential of automatic recognition of head-movement functions and provide insightful perspectives. The recognition performance varied from nearly 30% to 90% in terms of the F-score, depending on the function category. The performance measures were positively correlated with the amount of data and the inter-coder agreement. The detection results had a tendency toward overdetection that added more functions to the existing functions in the corpus, which implies dense interactions between the speaker and listeners. There are several ways of interpreting the difference between the CNN's output and the corpus. First, the results could be due to the limited discriminating power of the CNN models. Since the models used in this study are considered to be baseline CNN models, there might be room for further performance improvement by using more powerful DNNs. Second, the limited amount of data restricted the CNN's performance, as indicated in Fig. 8(a) in IV-D1. Third, the poor consistency of the data due to its multifunctionality could have hampered the generalization capability of the CNN model (see Fig. 8(b) in IV-D1). This indicates that the head movements and utterance features for some functions in our corpus do not include sufficient information for training the CNN models. Thus, increasing the amount of data would be an effective strategy. In addition, more multimodal behavioral data and/or contextual data might complement our bimodal data and help to improve the performance.

There are several possible directions toward improved DNN models as follows. First, to cope with a small amount of data and improve the model generalization, a data augmentation technique for time series is considered effective [140]. Second, to capture various temporal structures such as a rapid periodic nodding and a slow single-shot nod, the multi-scale CNN structures are prospective [141]. Third, a channel attention mechanism can be a promising extension because it can adaptively weight the importance of each head-movement channel, such as the *elevation* for affirmative nodding and *roll* for doubtful tilt movements. Finally, residual structures (e.g., ResNet [142]), deeper architectures (e.g., VGG [143]), and hybrid models such as ConvLSTM [121] are worth exploring.

Another issue pertains to the limitations of the corpus. In highly interactive conversations, as in Fig. 10, the speaker and listeners actively move their heads, manifesting a rich functional spectrum. However, there is a possibility that the current corpus contains only part of the entire functional spectrum and that our CNN model detected some of the missing functions. The human coders used a coarse-to-fine strategy to reduce their workload, and this might have biased the assignments of the speech-related functions to the speaker only and the reaction functions to the listeners only. However, in actual conversations, the speaker not only makes an utterance but also reacts to the listeners' feedback signals by using head movements. Likewise, the listeners not only listen and react to the speaker but also often give short back-channel utterances with head movements. In this way, the speaker and listeners densely exchange multifunctional messages through head movements without interruption throughout the conversation. Our CNN models do not simply rely on the participant's role but rather identify multiple functionalities solely from the behavioral data. Therefore, there is a possibility that the CNN model's output could have caught such an interactive multifunctionality that the human annotators missed.

Furthermore, as mentioned in III-C, the coders' operational errors and individual differences in cognitive style might partially account for the diverging codes among the coders, which might have had some impact on the detection performance.

In addition, we confirmed that CNN models that use image-based head-pose measurements provide a level of performance comparable to that of sensor-based measurements. Although further validation is needed, the machine-learning approach for analyzing head-movement functions seems promising for various applications and communications research.

V. GENERAL DISCUSSION

We believe that the functional head-movement corpus introduced in this paper will be an important resource for multidisciplinary communication studies involving head movements. One interesting topic is to analyze the temporal development of mutual understanding among conversation participants. Intraturn analysis can reveal how the listener's cognitive state and responses change over time—starting with *listening*, followed by *understanding*, and finally reaching *affirmation*—by assessing the temporal changes in the functions of the listener's back-channel head movements. Interturn analysis can reveal how a one-to-one empathetic relationship between the speaker and listener evolves into a group-level consensus as the conversation progresses. Second, interactional synchrony [144] is a very important aspect to explore. By going beyond head-movement synchrony *as a whole* as studied in past works [144]–[146], our corpus makes it possible to analyze function-level synchrony, which could lead to a deeper understanding of rapport building [147], consensus building [146], and therapy [148].

An important direction of study is to expand the scope to multimodal behaviors, such as gaze, facial expressions, verbal content, prosody, and hand/body gestures. First, gaze and head movements work together to signal visual attention in eliciting and giving feedback [52], [76], [78] and in yielding and taking turns [53]. Thus, joint modeling of these two modalities should be able to provide better disambiguation of these interactional functions from others. Second, facial expressions are powerful cues for recognizing communicative functions and are both supplementary and complementary to head movements. The attitude perceived from head movements could be modulated by facial expressions; e.g., a "nod" plus a "smile" could be interpreted as "like" and "accept" rather than "agree" and "understand" [16]. Third, nonverbal content could be integrated with verbal content and back-channel functions such as newsmarkers (e.g., "really?"), change-of-activity tokens (e.g., "alright"), and assessment tokens (e.g., "great") [79]. These different verbal back-channels complement the visual back-channel function. In addition, the dialogue act would be another powerful cue for detailed classification; e.g., *affirmation* could be divided into subcategories such as agreement, acceptance, and permission, as in Poggi's typology [24], [25]. Fourth, phonetic and paralinguistic features are useful cues to the speaker's affective, attitudinal, and emotional state [149], [150]. Furthermore, it would be useful to develop a comprehensive classification of bodily movements, including those of the head, arms, hands, torso, and feet. For example, we could investigate the usage of head movements in illustrative gestures in story-telling situations.

Some methodological issues affected the building of our functional head-movement corpus. Currently, our corpus employs a decision-making rule based on the logical sum of three coders' judgments. This means there is no distinction between the cases judged 'positive' by all three coders and those judged 'positive' by only one coder. To more accurately model the diversity of observer interpretations, it would be worth considering continuous-valued codes, which can represent the probability density distribution of multifunctionality. Such continuous-valued codes could be used as weighting coefficients in the cost function for training the CNN models. Another plausible method is majority voting and/or outlier rejection applied to a larger number of coders, as we mentioned in IV-E. Moreover, the list of communicative functions in Table 1 is not comprehensive and has missing categories, such as the listener's *gesticulations*, the speaker's *clause boundaries*, and functions related to gaze shifts. The flat structure of our coding scheme allows users to add more categories or aggregate some into one category.

Our findings are based on very limited data, i.e., two groups of conversations among four Japanese females. To generalize our findings, it will be necessary to consider differences among individuals, genders, cultures, and languages. First, human behaviors are strongly influenced by personality traits; e.g., extroverted people speak more loudly and look at others more often [84]. Second, there is a gender gap in nonverbal behaviors; i.e., females are generally more expressive than males [84]. Third, there are great cultural and language differences among people; e.g., Japanese listeners nod more than American listeners when giving back-channel responses [74]. Such differences are rooted in sociocultural practices; e.g., Japanese people tend to constantly monitor each other's feelings and to try to create social bonds through coordinating their behavior [22], [74], [147]. More detailed analyses from these perspectives would be needed to see how such differences affect the relationship between the kinetics of head movements and their functions and the relation between the frequency of occurrence and the context of the particular functions. Such analyses may reveal the potential for generalization and transferability of a classification model trained for a specific culture/language to other settings.

Related to the above points, although this paper focused only on the immediate effects of head movements, head movements in conversations affect social outcomes beyond the interactions in the moment. Kita and Ide suggested that they help to build rapport among conversation participants; their suggestion is based on the hypothesis that mutual simultaneous nodding boosts positive affects [147]. Yap *et al.* found a correlation between head motion synchrony and perceived empathy by passive listeners. Ramseyer and Tschacher found that synchrony between the head motions of a therapist and client can predict the outcome of the therapy [148]. Osugi *et al.* showed that nodding and shaking of the head have an effect on the observer's perceptions in terms of likeability and approachability [151]. Wells and Petty discovered that neutral listeners tend to show a more favorable attitude toward the content of speech when they are nodding than when they are shaking their heads while listening [152]. Briñol and Petty found that nodding and shaking the head can either strengthen or weaken confidence in the message heard [153]. Tom *et al.* showed that nodding while listening increases the preference for neutral objects and that shaking the head while listening decreases the preference [154]. Therefore, as future work, it would be interesting to investigate the relationship between short-term head-movement functions and long-term social outcomes.

Furthermore, the research framework described in this paper has the potential to boost social interaction studies. Nonverbal behaviors, including head movements in conversations, are related to personality traits [8], [9], leadership [10], [11], vertical dimensions such as dominance and power [155], communication skills [6], [7], and mental states such as depression [156]. By going beyond the statistics on head movements as a whole, e.g., the frequency of nodding, our functional analysis could reveal more detailed aspects of the phenomena in these domains. Intuitively, a good speaker is more effective at attracting the listener's attention, soliciting their feedback, and making his/her positions more understandable and agreeable. Likewise, a good listener more promptly gives feedback to the speaker and leads the speaker so that he/she says what the listener wants to know. Such mutual orientation and cooperation are key in successful conversation. Thus, emergent communicative functions should be closely linked to communication skills and group performance. A detailed analysis of these functional aspects by using the corpus and automatic classification models would be beneficial to numerous disciplines, including education, organizational research, and mental healthcare.

We employed the minimum bimodal approach, which deals with a single person's head movements and speaking activity, without other modalities or interactional context. Although the primary stream of research is toward multimodal behaviors and interactional contexts, as mentioned above, our approach has several advantages. First, measuring the head pose is more convenient than measuring facial expressions and hand/body gestures. Sensing facial expressions requires near-front high-resolution imagery or facial electromyography (fEMG) [157]. Sensing hand and body gestures requires wide-field imagery and/or motion capture systems. On the other hand, head pose velocity can be easily measured with low-resolution image-based tracking and/or wearable accelerometers, which are available in increasing numbers of head-worn devices, such as head mounted displays (HMDs), headset microphones, and smart glasses [158]. Our preliminary experiment confirmed that our CNN models perform well when using head pose *acceleration* as input data. Since our model is context-free and relies on only a single person's behaviors, a single model built for face-to-face conversations can be applied to other situations, such as teleconferences and embodied conversational agents (ECAs), under the assumption that people tend to make similar head movements in different situations.

Our study has several applications. One is ECAs [4], [5], which in the context of our study would have two facets: reading the user's head movements [13]–[15] and synthesizing the ECA's head movements [13], [16], [17]. The former concerns 'mind-reading' of the user's attitudes and cognitive state from their head movements. An ECA can better decide what to do next if it can read the user's intention to listen, level of understanding, and desire to speak. Sensing the polarity from subtle movements of the user would be important, especially when it is negative, because the user might need help. The latter facet is important for ECAs to correctly display their intentions by performing head movements in an interpretable way. Our corpus and model would be useful for both designing an ECA's head movements and evaluating the generated movements. Not only directly evaluating the ECA's head movements but also analyzing users' reactions in terms of their head movements would be an interesting topic. Furthermore, developing and evaluating telepresence systems are both within our scope [18]. Comparing face-to-face and telepresence settings in terms of emergent head-movement functions would indicate the quality of a telepresence system. The optimal control of the head movements of telepresence robots or agents would be important for compensating and amplifying the communicative functions, depending on the communication delay and other environmental conditions.

VI. CONCLUSION

This paper presented a functional head-movement corpus and deep neural network for analyzing the multiple communicative functions of head movements in multiparty face-to-face conversations. To explore the multifunctionality of head movements, we defined a set of non-mutually-exclusive functions that cover a wide range of functions, including speech production, eliciting and giving feedback, turn management, and cognitive and affect display. Our head-movement corpus employs binary codes that represent arbitrary combinations of functions and logical-sum-based aggregations of multiple coders' observations for capturing multifunctionality as well as possible.

We created a functional head-movement corpus from four-party conversations among Japanese females. The corpus analysis revealed multifunctional structures, in which the speaker modulates multiple functions, such as emphasis, and elicits the listener's responses through rhythmic head movements while listeners express various attitudes and responses through continuous back-channel head movements. Additionally, various cognitive and affective display functions were found to be densely interrelated, and a rich functional spectrum was formed through the head movements.

We proposed convolutional neural networks (CNNs) to separately detect each head-movement function from bimodal data consisting of the head-pose velocity and the presence or absence of utterances. The experimental results were insightful and promising. The recognition performance varied from nearly 30% to 90% in terms of the F-score, depending on the function category, and had positive correlations with the amount of data and inter-coder agreement. In addition, we confirmed a tendency toward overdetection that added more functions to those originally in the corpus, which implies that dense interactions exist between the speaker and listeners. Furthermore, we found that CNN models outperform SVM models and that CNN models that use image-based tracking can provide a comparable level of performance to that of sensor-based measurements.

These results suggest that the functional head-movement corpus and CNN-based head function detectors are useful tools for understanding human communications involving head movements and for developing various applications.

ACKNOWLEDGMENT

The authors would like to thank NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, which supported us in starting and continuing this work on a joint research project between NTT and Yokohama National University. Additionally, we would like to express our gratitude to the internship program between NTT and Nagaoka University of Technology, which gave the second author the opportunity to engage in this project.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[2] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[3] K. Otsuka, "Conversation scene analysis," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 127–131, Jul. 2011.

[4] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Mag.*, vol. 22, no. 4, pp. 67–83, Oct. 2001.

[5] E. Andre and C. Pelachaud, "Interacting with embodied conversational agents," in *Speech Technology: Theory and Application*, vol. 7. New York, NY, USA: Springer, 2010, pp. 123–149.

[6] S. Okada, Y. Ohtake, Y. I. Nakano, Y. Hayashi, H.-H. Huang, Y. Takase, and K. Nitta, "Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2016, pp. 169–176.

[7] R. Ishii, S. Kumano, and K. Otsuka, "Analyzing gaze behavior during turn-taking for estimating empathy skill level," in *Proc. 19th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2017, pp. 365–373.

[8] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez, "Linking speaking and looking behavior patterns with group composition, perception, and performance," in *Proc. 14th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2012, pp. 433–440.

[9] M. Jensen, "Personality traits and nonverbal communication patterns," *Int. J. Social Sci. Stud.*, vol. 4, no. 5, pp. 57–70, Mar. 2016.

[10] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 441–456, Feb. 2018.

[11] A. Darioly and M. Mast, "The role of nonverbal behavior for leadership: An integrative review," in *Leader Interpersonal Influence Skills: The Soft Skills Leadership*. New York, NY, USA: Taylor and Francis, 2014, pp. 73–100.

[12] S. Bonaccio, J. O'Reilly, S. L. O'Sullivan, and F. Chiocchio, "Nonverbal behavior and communication in the workplace: A review and an agenda for research," *J. Manage.*, vol. 42, no. 5, pp. 1044–1074, Jul. 2016.

[13] D. Heylen, E. Bevacqua, C. Pelachaud, I. Poggi, J. Gratch, and M. Schröder, "Generating listening behaviour," in *Handbook of Emotion-Oriented Technologies*. London, U.K.: Springer, 2010.

[14] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," in *Proc. 11th Int. Workshop Intell. Virtual Agents*, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds. Berlin, Germany: Springer, 2011, pp. 68–79.

[15] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. T. Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 165–183, Apr. 2012.

[16] E. Bevacqua, "Computational model of listener behavior for embodied conversational agents," Ph.D. dissertation, Univ. Paris, Paris, France, 2009.

[17] C. Oertel, J. Lopes, Y. Yu, K. A. F. Mora, J. Gustafson, A. W. Black, and J.-M. Odobez, "Towards building an attentive artificial listener: On the perception of attentiveness in audio-visual feedback tokens," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2016, pp. 21–28.

[18] K. Otsuka, "Behavioral analysis of kinetic telepresence for small symmetric group-to-group meetings," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1432–1447, Jun. 2018.

[19] D. Sirkin, G. Venolia, J. Tang, G. Robertson, T. Kim, K. Inkpen, M. Sedlins, B. Lee, and M. Sinclair, "Motion and attention in a kinetic videoconferencing proxy," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2011, pp. 162–180.

[20] R. Jakobson, "Motor signs for 'Yes' and 'No,'" *Lang. Soc.*, vol. 1, no. 1, pp. 91–96, Apr. 1972.

[21] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, Jan. 1969.

[22] S. K. Maynard, "On back-channel behavior in japanese and english casual conversation," *Linguistics*, vol. 24, no. 6, pp. 1079–1108, 1986.

[23] D. Heylen, "Head gestures, gaze and the principles of conversational structure," *Int. J. Humanoid Robot.*, vol. 3, no. 3, pp. 241–267, Sep. 2006.

[24] I. Poggi, F. D'Errico, and L. Vincze, "Types of nods. the polysemy of a social signal," in *Proc. 7th Conf. Int. Lang. Resour. Eval.*, 2010, pp. 2570–2576.

[25] I. Poggi, F. D'Errico, and L. Vincze, "68 nods. But not only of agreement," in *Proc. 68th Zeichen Für Roland Posner. Ein Semiotisches Mosaik. (68 Signs Roland Posner. Semiotic Mosaic) (Stauffenburg Verlag, Tübingen, Germany)*, 2012, pp. 1–14.

[26] S. Gries, "Polysemy," in *Handbook of Cognitive Linguistics*. Berlin, Germany: De Gruyter Mouton, 2015, pp. 472–490.

[27] P. Ekman and W. V. Friesen, *Unmasking Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.

[28] D. McNeill, *Hand Mind*. Chicago, IL, USA: Univ. of Chicago Press, 1996.

[29] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[30] P. Ekman and W. V. Friesen, *Facial Action Coding System: Manual*. Paul Ekman Group LLC., 2002.

[31] R. L. Birdwhistell, *Kinesics and Context: Essays on Body Motion Communication*. Philadelphia, PA, USA: Univ. of Pennsylvania Press, 1970.

[32] M. Włodarczak, H. Buschmeier, Z. Malisz, S. Kopp, and P. Wagner, "Listener head gestures and verbal feedback expressions in a distraction task," in *Proc. Interdiscipl. Workshop Feedback Behaviors Dialog*, 2012, pp. 1–4.

[33] S. Kousidis, Z. Malisz, P. Wagner, and D. Schlangen, "Exploring annotation of head gesture forms in spontaneous human interaction," in *Proc. Tilburg Gesture Res. Meeting (TiGeR)*, 2013, pp. 1–4.

[34] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner, "Are you sure you're paying attention?'-Uh-huh'communicating understanding as a marker of attentiveness," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2011, pp. 2057–2060.

[35] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose, "Head movement correlates of juncture and stress at sentence level," *Lang. Speech*, vol. 26, no. 2, pp. 117–129, Apr. 1983.

[36] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, "Kinematics of head movements accompanying speech during conversation," *Hum. Movement Sci.*, vol. 2, nos. 1–2, pp. 35–46, Jun. 1983.

[37] U. Hadar, T. J. Steiner, and F. Clifford Rose, "Head movement during listening turns in conversation," *J. Nonverbal Behav.*, vol. 9, no. 4, pp. 214–228, 1985.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, MA, USA: MIT Press, 2016.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[40] K. Jokinen, C. Navarretta, and P. Paggio, "Distinguishing the communicative functions of gestures," in *Proc. Mach. Learn. Multimodal Interact.*, 2008, pp. 38–49.

[41] C. Navarretta and P. Paggio, "Classification of feedback expressions in multimodal data," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2010, pp. 318–324.

[42] P. Paggio and C. Navarretta, "The danish NOMCO corpus: Multimodal interaction in first acquaintance conversations," *Lang. Resour. Eval.*, vol. 51, no. 2, pp. 463–494, Jun. 2017.

[43] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[44] S. O. Slim, A. Atia, M. Elfattah, and M.-S. M. Mostafa, "Survey on human activity recognition based on acceleration data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 3, pp. 84–98, 2019.

[45] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 3995–4001.

[46] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in *Proc. ACM Int. Symp. Wearable Comput. (ISWC)*, 2017, pp. 158–165.

[47] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. 35th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1533–1540.

[48] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Services*, 2014, pp. 197–205.

[49] C. A. Ronao and S.-B. Cho, "Deep convolutional neural networks for human activity recognition with smartphone sensors," in *Proc. Int. Conf. Neural Inf. Process.*, 2015, pp. 46–53.

[50] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2014, pp. 298–310.

[51] D. Heylen, "Challenges ahead: Head movements and other social acts during conversations," in *Proc. Joint Symp. Virtual Social Agents*, 2005, pp. 45–52.

[52] E. Z. McClave, "Linguistic functions of head movements in the context of speech," *J. Pragmatics*, vol. 32, no. 7, pp. 855–878, Jun. 2000.

[53] S. K. Maynard, "Interactional functions of a nonverbal sign head movement in japanese dyadic casual conversation," *J. Pragmatics*, vol. 11, no. 5, pp. 589–606, Oct. 1987.

[54] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol. 57, pp. 209–232, Feb. 2014.

[55] W. S. Condon and W. D. Ogston, "A segmentation of behavior," *J. Psychiatric Res.*, vol. 5, pp. 221–235, Sep. 1967.

[56] A. T. Dittmann and L. G. Llewellyn, "Body movement and speech rhythm in social conversation," *J. Personality Social Psychol.*, vol. 11, no. 2, pp. 98–106, 1969.

[57] P. Bull and G. Connelly, "Body movement and emphasis in speech," *J. Nonverbal Behav.*, vol. 9, no. 3, pp. 169–187, 1985.

[58] A. Kendon, "Some uses of the head shake," *Gesture*, vol. 2, no. 2, pp. 147–182, Dec. 2002.

[59] M. Boholm and J. Allwood, "Repeated head movements, their function and relation to speech," in *Proc. Workshop Multimodal Corpora, Adv. Capturing, Coding Analyzing Multimodality*, 2010, pp. 6–10.

[60] S. Duncan and G. Niederehe, "On signalling that it's your turn to speak," *J. Exp. Social Psychol.*, vol. 10, no. 3, pp. 234–247, May 1974.

[61] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, vol. 25. Berlin, Germany: De Gruyter Mouton, 1980.

[62] M. Harness Goodwin and C. Goodwin, "Gesture and coparticipation in the activity of searching for a word," *Semiotica*, vol. 62, nos. 1–2, pp. 51–75, 1986.

[63] U. Hadar, T. J. Steiner, and F. C. Rose, "The relationship between head movements and speech dysfluencies," *Lang. Speech*, vol. 27, no. 4, pp. 333–342, Oct. 1984.

[64] V. H. Yngve, "On getting a word in edgewise," in *Proc. 6th Chicago Linguistic Soc.*, 1970, pp. 567–577.

[65] S. Duncan and D. Fiske, *Face-to-Face Interaction: Research, Methods, and Theory*. New York, NY, UDA: Taylor & Francis Group, 1977.

[66] M. F. Schober and H. H. Clark, "Understanding by addressees and overhearers," *Cognit. Psychol.*, vol. 21, no. 2, pp. 211–232, Apr. 1989.

[67] H. Clark, *Using Language*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[68] J. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *J. Personality Social Psychol.*, vol. 79, no. 6, pp. 941–952, 2000.

[69] J. B. Bavelas and N. Chovil, "Visible acts of meaning: An integrated message model of language in face-to-face dialogue," *J. Lang. Social Psychol.*, vol. 19, no. 2, pp. 163–194, Jun. 2000.

[70] L. Cerrato and M. Skhiri, "Analysis and measurement of head movements signalling feedback in face-to-face human dialogues," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2003, pp. 43–52.

[71] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *J. Memory Lang.*, vol. 50, no. 1, pp. 62–81, Jan. 2004.

[72] C. Goodwin, *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY, USA: Academic, 1981.

[73] H. Aoki, "Some functions of speaker head nods," in *Embodied Interaction: Language and Body in the Material World*. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 93–105.

[74] S. K. Maynard, "Conversation management in contrast: Listener response in japanese and American English," *J. Pragmatics*, vol. 14, no. 3, pp. 397–412, Jun. 1990.

[75] Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, and K. Shirai, "Analysis of head movements and its role in spoken dialogue," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, vol. 4, 1996, pp. 2167–2170.

[76] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *J. Commun.*, vol. 52, no. 3, pp. 566–580, Sep. 2002.

[77] A. Terrell and B. Mutlu, "A regression-based approach to modeling addressee backchannels," in *Proc. 13th Annu. Meeting Special Interest Group Discourse Dialogue (SIGDIAL)*, 2012, pp. 280–289.

[78] G. Ferre and S. Renaudier, "Unimodal and bimodal backchannels in conversational english," in *Proc. SEMDIAL (SaarDial) Workshop Semantics Pragmatics Dialogue*, Aug. 2017, pp. 20–30.

[79] R. Gardner, *When Listeners Talk*. Amsterdam, The Netherlands: John Benjamins Publishing Company, 2001.

[80] E. A. Schegloff, "Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences," in *Analyzing Discourse: Text and Talk*. Washington, DC, USA: Georgetown Univ. Press, 1982.

[81] H. M. Rosenfeld and M. Hancks, "The nonverbal context of verbal listener responses," in *The Relationship of Verbal and Nonverbal Communication*. The Hague, The Netherlands: Mouton Publishers, 1980, pp. 206–293.

[82] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *J. Semantics*, vol. 9, no. 1, pp. 1–26, 1992.

[83] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 154.

[84] M. Argyle, *Bodily communication—2nd ed.* London, U.K.: Routledge, 1988.

[85] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image Vis. Comput.*, vol. 31, no. 2, pp. 203–221, Feb. 2013.

[86] Z. Malisz and M. Karpiński, "Multimodal aspects of positive and negative responses in polish task-oriented dialogues," in *Proc. 5th Int. Conf. Speech Prosody*, 2010.

[87] T. Stivers, "Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation," *Res. Lang. Social Interact.*, vol. 41, no. 1, pp. 31–57, Mar. 2008.

[88] I. de Kok and D. Heylen, "Analyzing nonverbal listener responses using parallel recordings of multiple listeners," *Cogn. Process.*, vol. 13, no. 2, pp. 499–506, Oct. 2012.

[89] S. Livingstone and C. Palmer, "Head movements encode emotions during speech and song," *Emotion*, vol. 16, no. 3, pp. 365–380, 2016.

[90] A. Adams, M. Mahmoud, T. Baltrusaitis, and P. Robinson, "Decoupling facial expressions and head motions in complex emotions," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 274–280.

[91] A. Samanta and T. Guha, "On the role of head motion in affective expression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2886–2890.

[92] H. Sacks, E. Schegloff, and G. Jefferson, "A simple systematic for the organisation of turn taking in conversation," *Language*, vol. 50, pp. 696–735, Dec. 1974.

[93] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *J. Personality Social Psychol.*, vol. 23, no. 2, pp. 283–292, 1972.

[94] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, "The timing of shifts of head postures during conservation," *Hum. Movement Sci.*, vol. 3, no. 3, pp. 237–245, Sep. 1984.

[95] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[96] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena," *Lang. Resour. Eval.*, vol. 41, nos. 3–4, pp. 273–287, Dec. 2007.

[97] E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *Proc. Interspeech*, 2005, pp. 813–816.

[98] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing interpersonal empathy via collective impressions," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 324–336, Oct. 2015.

[99] R. J. Passonneau and B. Carpenter, "The benefits of a model of annotation," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 311–326, Dec. 2014.

[100] H. Rohde, A. Dickinson, N. Schneider, C. N. L. Clark, A. Louis, and B. Webber, "Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task," in *Proc. 10th Linguistic Annotation Workshop Conjunct ACL (LAW-X)*, 2016, pp. 49–58.

[101] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes,'" in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 40–45.

[102] A. Kapoor and R. W. Picard, "A real-time head nod and shake detector," in *Proc. Workshop Percetive User Interfaces (PUI)*, 2001, pp. 1–5.

[103] J. R. Terven, J. Salas, and B. Raducanu, "Robust head gestures recognition for assistive technology," in *Proc. Mex. Conf. Pattern Recognit. (MCPR)*, 2014, pp. 152–161.

[104] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A conversation robot using head gesture recognition as para-linguistic information," in *Proc. RO-MAN. 13th IEEE Int. Workshop Robot Hum. Interact. Commun.*, Sep. 2004, pp. 159–164.

[105] S. Martin, C. Tran, A. Tawari, J. Kwan, and M. Trivedi, "Optical flow based head movement and gesture analysis in automotive environment," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 882–887.

[106] S. Tschechne, G. Layher, and H. Neumann, "A biologically inspired model for the detection of external and internal head motions," in *Proc. Artif. Neural Netw. Mach. Learn. (ICANN)*, 2013, pp. 232–239.

[107] P. Lu, X. Huang, X. Zhu, and Y. Wang, "Head gesture recognition based on Bayesian network," in *Proc. Iberian Conf. Pattern Recognit. Image Anal. (IbPRIA)*, 2005, pp. 492–499.

[108] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Contextual recognition of head gestures," in *Proc. 7th Int. Conf. Multimodal Interfaces (ICMI)*, 2005, pp. 18–24.

[109] K. Otsuka, H. Sawada, and J. Yamato, "Automatic inference of cross-modal nonverbal interactions in multiparty conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI)*, 2007, pp. 255–262.

[110] H. Wei, P. Scanlon, Y. Li, D. S. Monaghan, and N. E. O'Connor, "Real-time head nod and shake detection for continuous human affect recognition," in *Proc. 14th Int. Workshop Image Anal. Multimedia Interact. Services (WIAMIS)*, Jul. 2013, pp. 1–4.

[111] G. Galanakis, P. Katsifarakis, X. Zabulis, and I. Adami, "Recognition of simple head gestures based on head pose estimation analysis," in *Proc. The 4th Int. Conf. Ambient Comput., Appl., Services Technol. (AMBIENT)*, 2014, pp. 88–96.

[112] W. Tan and G. Rong, "A real-time head nod and shake detector using HMMs," *Expert Syst. Appl.*, vol. 25, no. 3, pp. 461–466, Oct. 2003.

[113] A. Kanaujia, Y. Huang, and D. Metaxas, "Emblem detections by tracking facial features," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, p. 108.

[114] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[115] Y. Chen, Y. Yu, and J.-M. Odobez, "Head nod detection from a full 3D model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 528–536.

[116] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.

[117] S. Bor Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1521–1527.

[118] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[119] H. Ç. Akakın and B. Sankur, "Robust classification of face and head gestures in video," *Image Vis. Comput.*, vol. 29, no. 7, pp. 470–483, Jun. 2011.

[120] M. Oyabu, D. Zrinscak, and K. Otsuka, "Automatic recognition of human head gestures in conversations using deep learning," (In Japanese), in *Proc. 31st Annu. Conf. Jpn. Soc. Artif. Intell.*, 2017.

[121] M. Sharma, D. Ahmetovic, L. A. Jeni, and K. M. Kitani, "Recognizing visual signatures of spontaneous head gestures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 400–408.

[122] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artif. Intell.*, vol. 171, nos. 8–9, pp. 568–585, Jun. 2007.

[123] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez, "Using self-context for multimodal detection of head nods in face-to-face interactions," in *Proc. 14th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2012, pp. 289–292.

[124] K. Bousmalis, L.-P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 746–752.

[125] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Feature selection of facial displays for detection of non verbal communication in natural conversation," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops*, Sep. 2009, pp. 1985–1992.

[126] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychol. Rev.*, vol. 97, no. 3, pp. 315–331, 1990.

[127] A. Mehrabian and J. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

[128] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[129] Polhemus. (2019) *Fastrak: The Workhorse 6Dof Motion Tracker That Set the Standard in Tracking*. [Online]. Available: https://polhemus.com/motion-tracking/all-trackers/fastrak

[130] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.

[131] K. Otsuka, K. Kasuga, and M. Köhler, "Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks," in *Proc. Int. Conf. Multimodal Interact. (ICMI)*, 2018, pp. 191–199.

[132] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[133] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 689–692.

[134] J. B. Yang. (2016). *CNN-Timeseries*. [Online]. Available: https://github.com/sibosutd/cnn-timeseries

[135] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 467–482.

[136] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[137] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[138] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.

[139] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977.

[140] B. Kenji Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," 2020, *arXiv:2007.15951*. [Online]. Available: http://arxiv.org/abs/2007.15951

[141] C. Avilés-Cruz, A. Ferreyra-Ramírez, A. Zuñiga-López, and J. Villegas-Cortéz, "Coarse-fine convolutional deep-learning strategy for human activity recognition," *Sensors*, vol. 19, no. 7, p. 1556, Mar. 2019.

[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[143] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[144] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 349–365, Jul. 2012.

[145] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall, "Movement coordination during conversation," *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e105036.

[146] Y. Miao Sin, Robin, Y. Inoue, S. Miura, J. Kwon, K.-I. Ogawa, and Y. Miyake, "Head motion synchrony in the process of consensus building: A comparison between native English and Japanese speakers," in *Proc. Int. Conf. Complex Med. Eng.*, 2015, pp. 31–36.

[147] S. Kita and S. Ide, "Nodding, aizuchi, and final particles in Japanese conversation: How conversation reflects the ideology of communication and social relationships," *J. Pragmatics*, vol. 39, no. 7, pp. 1242–1254, Jul. 2007.

[148] F. Ramseyer and W. Tschacher, "Nonverbal synchrony of head- and body-movement in psychotherapy: Different signals have different associations with outcome," *Frontiers Psychol.*, vol. 5, p. 979, Sep. 2014.

[149] J. Laver, *Principles of Phonetics* (Cambridge Textbooks in Linguistics). Cambridge, U.K.: Cambridge Univ. Press, 1994.

[150] R. Ogden, "Phonetics and social action in agreements and disagreements," *J. Pragmatics*, vol. 38, no. 10, pp. 1752–1775, Oct. 2006.

[151] T. Osugi and J. I. Kawahara, "Effects of head nodding and shaking motions on perceptions of likeability and approachability," *Perception*, vol. 47, no. 1, pp. 16–29, Jan. 2018.

[152] G. L. Wells and R. E. Petty, "The effects of over head movements on persuasion: Compatibility and incompatibility of responses," *Basic Appl. Social Psychol.*, vol. 1, no. 3, pp. 219–230, Sep. 1980.

[153] P. Brińol and R. Petty, "Overt head movements and persuasion: A self-validation analysis," *J. Personality Social Psychol.*, vol. 84, no. 6, pp. 1123–1139, 2003.

[154] G. Tom, P. Pettersen, T. Lau, T. Burton, and J. Cook, "The role of overt head movement in the formation of affect," *Basic Appl. Social Psychol.*, vol. 12, no. 3, pp. 281–289, Sep. 1991.

[155] J. A. Hall, E. J. Coats, and L. S. LeBeau, "Nonverbal behavior and the vertical dimension of social relations: A meta-analysis," *Psychol. Bull.*, vol. 131, no. 6, pp. 898–924, 2005.

[156] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 478–490, Oct. 2018.

[157] G. Read, "Facial electromyography (EMG)," in *The International Encyclopedia of Communication Research Methods*. Hoboken, NJ, USA: Wiley, 2017.

[158] L.-H. Lee and P. Hui, "Interaction methods for smart glasses: A survey," *IEEE Access*, vol. 6, pp. 28712–28732, 2018.

KAZUHIRO OTSUKA (Member, IEEE) received the B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University, in 1993 and 1995, respectively, and the Ph.D. degree in information science from Nagoya University, in 2007. He joined the NTT Human Interface Laboratories, Nippon Telegraph and Telephone Corporation, in 1995. From April 2001 to March 2020, he was consecutively a Research Scientist, Senior Research Scientist/Supervisor, and Distinguished Researcher with NTT Communication Science Laboratories. Since April 2020, he has been an Associate Professor with the Faculty of Engineering, Yokohama National University. His current research interests include multimodal data analysis, nonverbal communication, and social telepresence. He was awarded the IAPR International Conference on Image Analysis and Processing Best Paper Award in 1999, the Outstanding Paper Awards of the ACM International Conference on Multimodal Interfaces (Interaction) in 2007, 2012, and 2014, the IEICE KIYASU-Zen'iti Award 2010, the Innovative Technologies Special Award from the Japanese Ministry of Economy, Trade and Industry in 2012, and others. He is a member of the IEICE and the IPSJ.

MASAHIRO TSUMORI graduated from the National Institute of Technology (Kosen), Nara College, and enrolled in the Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology, in 2017. He received the B.E. degree from the Nagaoka University of Technology, in March 2019. He is currently pursuing the master's degree with the Graduate School of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology. From October 2018 to February 2019, he participated in the internship program at NTT Communication Science Laboratories, where he worked on this paper. His current interests include deep neural networks and image processing.