frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 27 October 2017
doi: 10.3389/frobt.2017.00049

Novel Speech Motion Generation by Modeling Dynamics of Human Speech Production

Kurima Sakai[1,2]*, *Takashi Minato*[2], *Carlos T. Ishi*[2] *and Hiroshi Ishiguro*[1,2]

[1]*Graduate School of Engineering Science, Osaka University, Toyonaka, Japan,* [2]*Advanced Telecommunications Research Institute International, Keihanna Science City, Japan*

We developed a method to automatically generate humanlike trunk motions based on speech (i.e., the neck and waist motions involved in speech) for a conversational android from its speech in real time. To generate humanlike movements, the android's mechanical limitation (i.e., limited number of joints) needs to be compensated for. By enforcing the synchronization of speech and motion in the android, the method enables us to compensate for its mechanical limitations. Moreover, motion can be modulated to express emotions by tuning the parameters in the dynamical model. This method is based on a spring-damper dynamical model driven by voice features to simulate the human trunk movements involved in speech. In contrast to the existing methods based on machine learning, our system can easily modulate the motions generated due to speech patterns because the model's parameters correspond to muscle stiffness. The experimental results show that the android motions generated by our model can be perceived as more natural and thus motivate users to talk longer with it compared to a system that simply copies human motions. In addition, our model generates emotional speech motions by tuning its parameters.

Keywords: humanlike motion, speech-driven system, head motion, android, emotional motion

Edited by:
Francesco Becchi, Telerobot Labs s.r.l., Italy

Reviewed by:
Felix Reinhart, Bielefeld University, Germany
Manfred Hild, Beuth University of Applied Sciences, Germany

***Correspondence:**
Kurima Sakai
kurima.sakai@atr.jp

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *13 July 2016*
Accepted: *12 September 2017*
Published: *27 October 2017*

Citation:
Sakai K, Minato T, Ishi CT and Ishiguro H (2017) Novel Speech Motion Generation by Modeling Dynamics of Human Speech Production. Front. Robot. AI 4:49. doi: 10.3389/frobt.2017.00049

1. INTRODUCTION

Humanoid robots, especially android robots, are expected to join daily human activities since they can interact with people in a humanlike manner. Androids that resemble humans are suitable for social roles that require rapport and reliability (Prakash and Rogers, 2014), and recent studies have used them as a guide at an event site (Kondo et al., 2013), a salesperson at a department store (Watanabe et al., 2015), a bystander in a medical diagnosis (Yoshikawa et al., 2011), and a receptionist (Hashimoto and Kobayashi, 2009). One critical issue in developing androids is the design of behaviors that people can accept. People tend to expect an agent's behaviors to be based on its appearance (Komatsu and Yamada, 2011); accordingly, humanlike or natural behaviors are expected from androids. Here, natural means humanlike since their appearance is very humanlike and humanlike motion matches with the androids. In other words, people tend to have negative impressions of androids when they do not show the expected humanlike motions. However, an android actually suffers from a clear technical limitation originating in its mechanics; it is not possible to produce motions that are exactly the same as those of humans. The number of

controllable joints, the range of joint motions, and achievable joint velocity are limited, and thus an android's motion is much less complex than a human's motion.

This issue can be serious for an android. To effectively exploit its humanlike characteristics, androids must be designed with motions that can foster rapport and reliability. This requires the capability to express subtle changes of facial expressions and body motions. Concerning nonverbal behavior, a human conveys information to a partner not only by gestures but also by varying such motions slightly according to emotions or attitudes. Therefore, in the design of android motions, we must solve the following issues:

- The android needs to produce humanlike motions despite its mechanical limitations (e.g., restricted degrees of freedom).
- The motions must be modulated (i.e., changeable motion properties) based on the android's internal state (i.e., expressing emotions or attitudes).

For the first issue, androids must provide humanlike impressions even though their motions are not exactly identical as those of humans. To develop a model that can generate such movements, this study focuses on the synchronization between such multimodal expressions as speech and body movement. By enforcing multimodal synchronization in an android, we hypothesize that people will feel a motion's human-likeness even though it is mechanically restricted. This idea comes from existing knowledge that the multiplicative integration of multimodal signals is an effective cognitive strategy for humans to recognize an object's material or texture (e.g., Ernst and Banks, 2002; Fujisaki et al., 2014). In the development of a motion generation model, we also need to consider how to evaluate the human-likeness of the generated motion. If we can reproduce human motions in androids, we can develop a model by minimizing the measurable differences between the generated motion and the original human motion. Unfortunately, this approach does not achieve humanlike motions due to the mechanical limitations. Therefore, this study develops a model based on subjective human evaluations; we selected model parameters whose generated motions were subjectively judged for their human-likeness. The second issue can be solved by parameterizing the motion generation. Accordingly, this study develops an analytical model with parameters that influences the characteristics of the generated motions.

This study's target is a conversational android that mainly talks with people. It needs to evoke and maintain people's motivation to talk with it by creating a sense of reliability. Moreover, it needs to give the impression that it is producing its own utterances. It should also produce movements involved in speaking (speech motion). In addition, such speech motions must be changeable based on the android's current emotion. This study proposes a method of generating speech motions by modeling the dynamics of the human's trunk (neck and waist) motions involved in speech production. The synchronization between multimodal expressions (speech and body movements) improves the naturalness of the android's motion and reduces the negative impressions of noncomplex motions caused by its mechanical limitations. People might accept that the android itself is speaking if the movements of its lips, neck, chest, and abdomen, which are normally involved in human vocalization, are produced. This approach also allows motion variation by tuning the parameters in the dynamical model. The motion can be easily modulated to express the desired emotion since the parameters can be associated with motion changes owing to various mental states. Another important issue is real-time processing. In some cases, the utterance data cannot be prepared in advance for an autonomous robot: that is, the motions cannot be generated in advance. We design motion generation to occur simultaneously when the android speaks.

To find which speech and head motion features are synchronized during speaking, we first investigates human speech motions in Section 3. In Section 4, we develop a motion generation model that enforces the synchronization found in Section 3. The naturalness of generated motion needs to be subjectively evaluated. In Section 5, a psychological experiment verifies that the proposed model can generate android speech motions that can motivate people to talk with it. Section 6 shows that speech motion can be changed to express emotions by tuning the model parameters.

2. RELATED WORKS

Many studies have tackled the issue of human motion transfer in humanoid robots while overcoming the limitation of robot kinematics. For example, Pollard et al. (2002) proposed a method to scale the joint angles and velocities of measured human motions to the capabilities of a humanoid robot. Their robot successfully mimicked the dancing motions of performers while preserving their movement styles. Even though those studies basically aimed to reproduce a human performer's motion, they could not modulate the motions and/or mix them with other motions according to a given situation. Furthermore, most studies have focused on transferring a motion's gestural meaning and not its human-likeness or changes of motion properties.

Another approach to the human-likeness of android motion is multimodal expression, that is, displaying motions synchronized with other expressions, such as speech. Salem et al. (2012) proposed that generated a pointing gesture by a communication robot that matches its utterance (for example, the robot points to a vase and says "pick up that vase"). This study showed the importance of gesture-level synchronization. Sakai et al. (2015) made a robot's head motion more natural by enforcing the synchronization of speech and motion. Their method automatically added head gestures (e.g., nodding and tilting) to the robot motions transferred from human motions, where the head gestures were synchronized with a human's speech acts. In their experiments with their method, people had more natural impressions of the robot motion than with the transferred motion. This result suggests that enforcing the synchronization of speech and motion improves the naturalness and human-likeness of an android's motion.

Some studies of virtual agents have proposed methods to automatically generate the neck movements of agents matched with their speech. Le et al. (2012) modeled the relationship between human neck movements (roll, pitch, and yaw) and the voice information (power and pitch) using a Gaussian mixture model (GMM) for real-time neck-motion generation. Other models using a Hidden Markov Model (HMM) have also been proposed: Sargin et al. (2008), Busso et al. (2005), and Foster and Oberlander

(2007). Watanabe et al. (2004) proposed a different kind of model that estimates the speaker's nodding timing based on historical on–off patterns in the voice. These methods can successfully generate humanlike motions, but they are restricted to certain situations in which a model's learning data can be collected. Generally, human speech motions depend on the relationship between the speaker and listener and the speech context (for example, in happy situations, the magnitude and speed of human speech motions might be larger than usual). To alter the agent motion to fit the given situation, data must be collected for any situation, which is impractical, unfortunately.

The situation-dependency issue of can be solved by modulating the android's motions. Jia et al. (2014) synthesized the head and facial gestures of a talking agent with emotional expressions. In their method, the nodding motion involved in the utterance of stressed syllables is modulated according to the emotion represented in a PAD model (an extension of Russell's emotion model, Russell, 1980). However, their study focused only on positive emotions. Moreover, only the amplitude of motion was modulated, although the other motion features (e.g., velocity) should be varied according to the emotion. Masuda and Kato (2010) developed a method that changed the gestural motions of a robot by associating Laban theory (Laban, 1988) with Russell's emotion model. Unfortunately, since their method produced exaggerated gestures that cannot be expressed by humans, it is unsuitable for very humanlike androids. Other researchers have studied the relationship between motion characteristics and emotion in walking (Gross et al., 2012) and kicking (Amaya et al., 1996) motions. Such studies commonly conclude that the magnitude and speech of motion vary depending on the emotion.

Physiological studies are also helpful to understand motions of situation dependency, which is, the relationship between internal states and motion characteristics. Many studies reported that psychological pressure influences the human's musculoskeletal system and produces jerky motions. For example, anxiety increases muscle stiffness (Fridlund et al., 1986). Nakano and Hoshino (2007) showed that a human's waist moves slowly in a relaxed state but jerkily in a nervous state. This also suggests that the musculoskeletal system is influenced by mental pressure. Consequently, speech motion that is dependent on internal states might be generated by simulating the dynamics of the musculoskeletal system.

In this study, we develop an analytical model based on the physical constraints (kinematic relations) of a human's speech production. We focus on the motions of the body's trunk (neck and waist) on the sagittal plane, since lateral motions usually depend on the social situation, e.g., neck yawing due to gaze aversion that depends on the conversation context (Andrist et al., 2014) as well as on the speaker's personality (Larsen and Shackelford, 1996). The motion generation is formulated by a spring-damper model that simulates musculoskeletal dynamics. The generated motion can be easily modulated by tuning the parameters of muscle stiffness to express internal states. Motions are triggered by utterances in this model. By observing a person's utterances, this study found the physical constraints between speech production and the movements of the human head and mouth (opening/closing) to implement this trigger.

3. RELATIONS BETWEEN PROSODIC FEATURES AND HEAD MOTIONS

3.1. Basic Idea of Generation Model

We generate a speech motion synchronized with an utterance to enforce multimodal synchronization to avoid unnatural motions caused by hardware limitations. Such motions should also be generated by an analytic model to easily modulate them based on the android's emotion. This section first finds the relation between prosodic features and head motions, which is not influenced by social situation. Human head movements are synchronized with prosodic features when a person speaks; in particular, voice power and pitch features have high correlation with head movements (Bolinger, 1985). However, for our current Japanese targets, the correlation between prosodic features and head movements is not very high (Yehia et al., 2002). Anatomical research has also reported that the human head moves when the mouth is opened (Eriksson et al., 1998). Since mouth openness strongly depends on vowels in Japanese, we expect to find a relationship between head movements and prosodic information that involves Japanese vowels. Here, we reveal what features of human movements are related to the three voice features (power, pitch, and vowels) when a person is speaking without interacting with anyone. Those relations underlie the model to generate context-free head motions in speaking (Section 4). These motions can be modulated to fit a social situation by turning the model parameters (Section 6).

3.2. Experimental Setup

To examine the relation between head motion and voice features, we recorded the speech of human subjects while controlling their voice features. Before the experiment, a preliminary test checked whether people could pronounce the required vowels (a, i, u, e, o) with the required pitch and power. We found that people did not move their heads at all when the power was low. Based on this result, we identified a relationship where the more loudly people speak, the larger the head movement becomes. We asked the subjects to produce loud voices in the experiments. The preliminary test also suggested that the vowel changes affect head motions when people loudly pronounced the vowels, but they had some difficulty doing so while keeping the required pitch. We prepared two tests: pitch and vowel. The former measured the head motion when the participants pronounced a set of any syllables (here we chose five vowels) with the required pitch (3 s for each syllable) to investigate the relation between the head motion and pitch. The latter measured the head motion when the participants loudly pronounced the required vowel for 3 s with any pitch to investigate the relation between the head motion and vowels.

The head movements were measured by an inertia measurement unit (IMU) (InterSense InertiaCube4) attached to the top of the head at a measurement sampling rate of 100 Hz. The participants were given the following instructions: "Clearly pronounce a high-voice-pitch 'a'" or "clearly pronounce a low-voice-pitch 'i.'" The frequency of the pitch was not fixed since they had difficulty producing the required frequency. They were also required to reset their posture forward before each pronunciation. In the pitch test, they loudly pronounced five vowels (each for 3 s)

with high, middle, and low pitch. In the vowel test, they loudly pronounced each vowel for 3 s with whatever pitch they could easily pronounce.

The participants conducted each test twice. The first trial acclimated them to speaking in the experimental room and checked that the recording system.

3.3. Results

Eleven Japanese speakers (six males, five females, average age: 22.0, SD: 0.54) participated in the experiment. They were recruited from a job-offering site for university students with different backgrounds. We removed one male from our analysis since he pronounced all the words with the same tone although we instructed him to pronounce them in high, middle, and low tones.

In the pitch test, the participants tended to maintain almost constant head posture while speaking. We calculated the average head-elevation angle during their utterances (**Figure 1**). They tended to look up when pronouncing with a high pitch and look forward or down when pronouncing with a low pitch. A one-way analysis of variance (ANOVA) with one within-subject factor revealed a significant main effect ($F(2,\ 18) = 12.84,\ p < 0.01$). Hereafter, F and p denote F statistics and significance level. Following this result, we conducted multiple comparisons by the Bonferroni method and found significant differences in the head angle as the angles increased with a greater pitch level. This means that they respectively tended to move their heads up and down when they spoke with high and low pitch.

Figure 2 shows the results of the vowel test results. The average head-elevation angle during speech was measured as well as in the pitch test. The vertical axis shows the head angle displacements and how much the head angle changed while pronouncing the vowels compared with the before pronouncing them. We categorized the vowels into two groups: wide group (a, e, o) and narrow group (i, u). We found a significant difference between the two groups ($p < 0.05$).[1] The participants greatly moved their heads when they opened their mouths. The observed motions seem to be made purely for pronouncing utterances since there was no social context; they were pronouncing meaningless syllables without any interaction. Therefore, the revealed relations might not depend on the social situations. However, the speech motions did change depending on the social situation. The generated motions of the android based on the relations must match the social situation.

Most participants in the vowel test pronounced with a middle pitch. The angle displacements in **Figure 2** are less than 5° but the head angles of middle pitch in **Figure 1** exceed 5°, because they slightly raised their head postures and moved their heads when they opened their mouths.

[1]We use the Wilcoxon rank sum test because the normality was not assumed by a Shapiro–Wilk test.

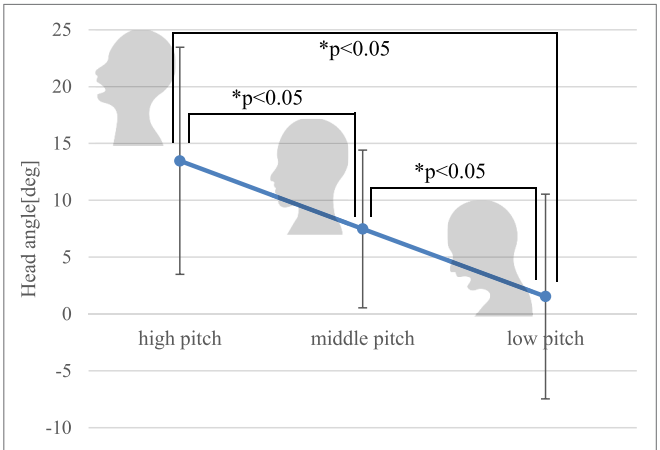

FIGURE 1 | Head postures related to voice pitch. The angle is zero when the participant is facing forward.

4. SPEECH-DRIVEN TRUNK MOTION GENERATION

4.1. Generating Smooth Motion from Prosodic Features

In this section, we develop a model to generate a context-free trunk motion. First, we made a head motion-generation model based on the above results and extended it to both trunk and neck motions. We found a strong relation between the head angle and the prosodic features in Section 3, but a simple mapping is not appropriate for generating a smooth motion, which is essential for human-likeness (Shimada and Ishiguro, 2008; Piwek et al., 2014). This is because prosodic features are sometimes intermittent and change rapidly. The model needs to generate smooth motions based on the rapidly changing discrete sound input. Concerning humanlike motions, people perceive naturalness in the second-order dynamic motions of a virtual agent (Nakazawa et al., 2009). Therefore, we use a spring-damper dynamical model (**Figure 3**; equation (1)) to generate a smooth motion from the non-smooth prosodic features

$$j\ddot{\theta}_{base} + d\dot{\theta}_{base} + k\theta_{base} = \tau(t)dir(t). \quad (1)$$

Head angle θ_{base} is driven by the external force $\tau(t)dir(t)$, where t denotes discrete time. The absolute values of force $\tau(t)$ and force direction $dir(t)$ are separately described below for convenience of explanation. The prosodic information is associated with the force $\tau(t)dir(t)$ based on the results in Section 3. For example, if the

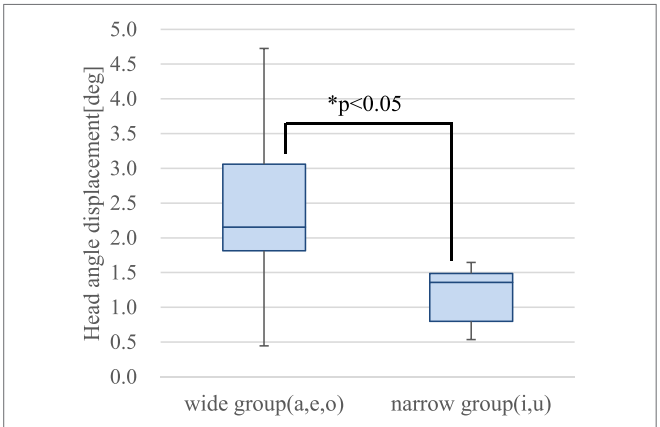

FIGURE 2 | Head displacements related to vowels.

model is given, a large upward force is made by widely opening the mouth, and a large upward neck movement is generated, following second-order dynamics. Parameters j, k, and d in the equation are equivalent to head weight, muscle stiffness, and muscle viscosity, respectively, based on using the spring-damper model of human muscle dynamics (Linder, 2000; Liang and Chiang, 2006). The meaning of each parameter is easy to grasp, and we can intuitively modulate the motion by changing the parameters. Furthermore, muscle stiffness is related to a person's internal states such as emotion and tension. We assume that we can modulate the motion to express the speaker's internal states.

4.2. Head-Motion Generation Based on Spring-Damper Model

This section defines the external force in equation (1) using prosodic information. The results in Section 3.3 show that pronouncing vowels with a wide opened mouth produces a large head movement. Hereafter, to express the vowel information with a continuous value, we use the value of mouth-openness value. As the mouth is opening, the force $m(t)$, which is proportional to the degree of mouth openness $\epsilon(t)$, is given to the model. $\epsilon(t)$ can be estimated from the voice, as described in Section 4.3. On the contrary, no force is given to the model while the mouth is closing, and thus the head smoothly returns to its original position by using the restoring force of the spring, as shown in equation (2). In this study, the sampling time is 10 ms. As described in Section 4.3, speaking with large power produces a large head movement. While the voice power is increasing or keeping the same level, force $p(t)$, which is proportional to it, is given to the model. As with the vowel, no force is given to the model with the power shown in equation (3). In total, the force is expressed as equation (4), where v and l are constant values (no physical meaning) to balance different scales of values (voice power and degree of mouth openness). **Figure 4** shows an example of the input force generated from the voice:

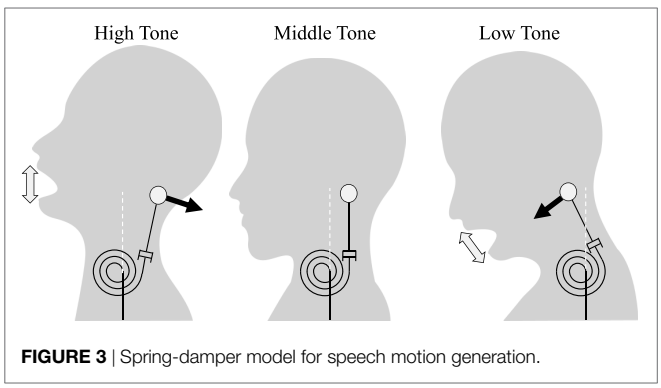

FIGURE 3 | Spring-damper model for speech motion generation.

$$m(t) = \begin{cases} \epsilon(t) & (\epsilon(t) >= \epsilon(t-1)) \\ 0 & (otherwise) \end{cases}, \tag{2}$$

$$p(t) = \begin{cases} \rho(t) & (\rho(t) >= \rho(t-1)) \\ 0 & (otherwise) \end{cases}, \tag{3}$$

$$\tau(t) = vp(t) + lm(t). \tag{4}$$

From the voice pitch results in Section 3.3, we associate the pitch with the force's direction, as shown in equation (5), where p_t denotes the state of pitch ($p_t \in \{High, Middle, Low\}$). Since the participants in the pitch test tended to change their postures more than usual, we focused on the direction of movement. **Figure 1** shows that the average head angle is a low pitch with a positive value, but we defined $dir(t)$, which became minus when the pitch was low, to enforce the relationship between the pitch and head direction.

Equation 5 means that an upward force is given to the head when the pitch is high and a head-lowering movement when the pitch is low. When the pitch is in the mid-range, no force is given

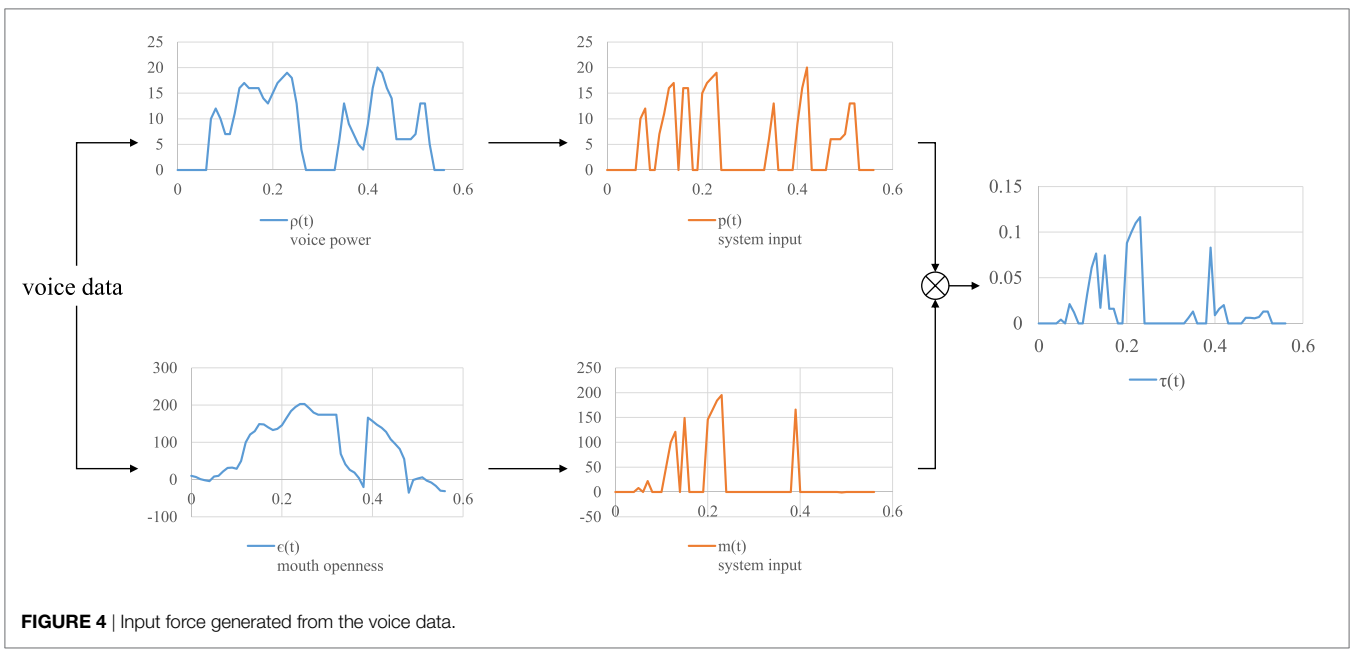

FIGURE 4 | Input force generated from the voice data.

to the model and only a restoring movement is generated. The next section describes how we classify the pitch into three groups

$$dir(t) = \begin{cases} 1 & (Head\ up) & (p_t = High) \\ -1 & (Head\ down) & (p_t = Low) \\ 0 & (Restoring\ movement) & (p_t = Middle) \end{cases}. \quad (5)$$

4.3. Prosodic Information Extraction

Fundamental frequency F0 was extracted in a 32-ms frame size every 10 ms. We used a conventional method that extracts the autocorrelation peaks of residual signals calculated by a linear predictive coding (LPC) inverse filter. Then value $\overline{F0}$ is calculated by averaging $F0$ over the last 100 ms. The pitch is classified as follows:

$$p_t = \begin{cases} High & \overline{F0} > F0_{high} \\ Low & \overline{F0} < F0_{low} \\ Middle & (\text{otherwise}) \end{cases}. \quad (6)$$

The frequency range was empirically determined since the voice's fundamental frequency depends on speaker's gender and age.

We estimated the mouth openness using a lip-motion-generating system (Ishi et al., 2012) based the voice's formant information. These methods can extract the prosodic features in real time and simultaneously create motion generation when the android produces a voice.

4.4. Trunk Motion Generation

Zafar et al. (2002) revealed that an up-and-down movement of the head also produces a front-and-back movement of the body. This suggests that a cooperative movement between the neck and waist might produce a more humanlike impression. Zafar et al. (2000) also reported that lip movement is followed by a neck movement, although with a slight delay. From these findings, we defined trunk motions (including head motion) as phase-shifted motions of θ_{base}, as shown in equation (7). act_i indicates the i-th actuator of the robot (**Figure 5**), and α_{act_i} and β_{act_i} are parameters that determine the coordination between the joints:

$$\theta_{act_i}(t) = \alpha_{act_i}\theta_{base}(t + \beta_{act_i}). \quad (7)$$

4.5. Improvement of the Model

To verify the model, we recorded the voice and motion of the speakers and compared the motions generated by our model and the originals. We tested two female speakers: Speakers M and H. We tested female voices because we used a female type of android is used in the latter experiment. The speakers read a self-introduction whose content was identical (approximately 50 s), and their trunk movements (neck and waist angles) were measured by the IMU attached to their head and torso.

To test the proposed model, we controlled the android's neck joint by the model. Parameters j and d in the model were set to 0.0676 and 0.52. k was set to 0.195 ($\theta_{base} \geq 0$) and 0.065 ($\theta_{base} < 0$). We assumed that the head can move easily when it looks downward since the head posture is not defying gravity. To implement these characteristics we set a smaller k value for $\theta_{base} \geq 0$. Balance parameters v and l in equation (4) were set to 0.001 and 0.0005, where the voice power ranged approximately from 10 to 20 dB and the mouth openness ranged approximately from 0 to 200 (non-dimensional value). Thresholds $F0_{high}$ and

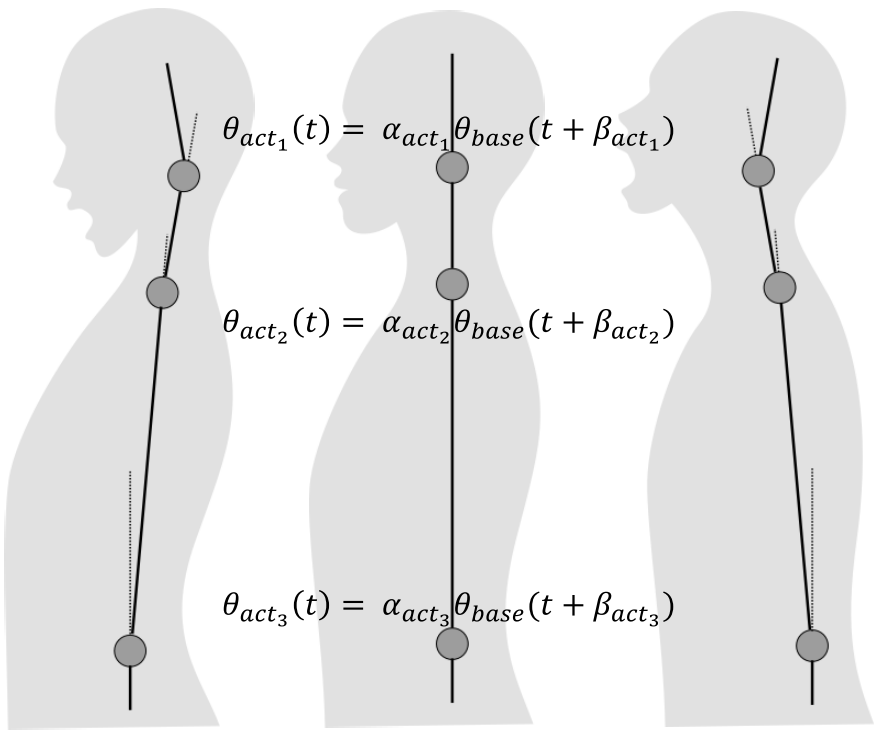

FIGURE 5 | Multi-joint control.

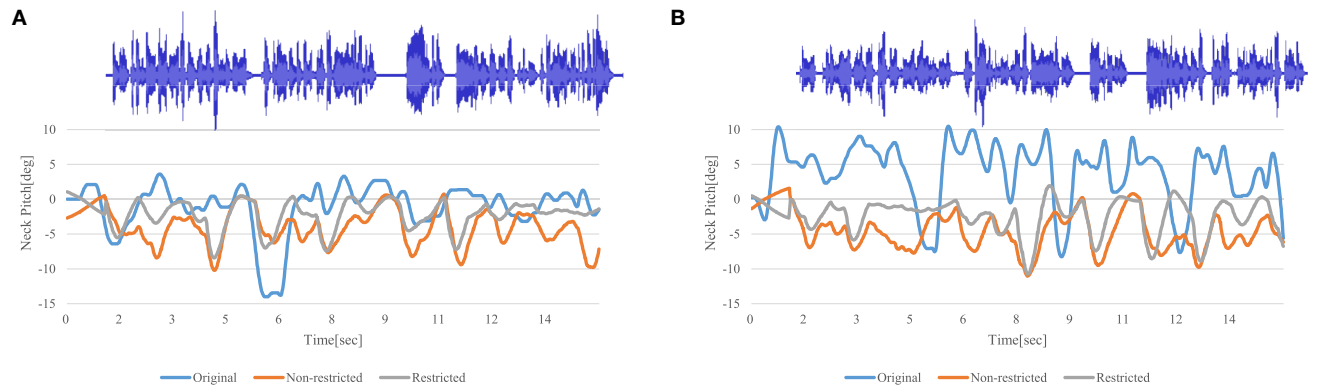

FIGURE 6 | Human head movements during speaking (top plot shows speech waveform and bottom plot shows generated head motion). **(A)** Speaker H. **(B)** Speaker M.

$F0_{low}$ in equation (6) were, respectively, set to 256 and 215 Hz. These parameters described in this session were set by trial and error. The experimenter selected the parameters in a preliminary trial for a natural neck motion that matched the speech. Mouth openness ($\epsilon(t)$) was calculated from the voice by Ishi et al.'s method (Ishi, 2005). Generated angle θ_{base} was directly used to control the neck joint (i.e., $\theta_{neck}(t) = \theta_{base}(t)$), but no other trunk joints were controlled in this experiment.

Figure 6 shows the head angles generated from two speakers' voices. Here, original indicates the original motion and non-restricted indicates the motion ($\theta_{neck}(t)$) generated by the proposed model. In **Figure 6B**, there is a clear difference between the trajectories of the original and proposed method, where the original angle was positive in most cases (Speaker M looked up during the speech), but the angle by the proposed method was negative in most cases. This is because Speaker M reset her posture again just before speaking and consequently looked a little bit up during the speech, although she was required to face forward before speaking and to keep her gaze during the speech. In this research, since we focus on the correlation between the changes of the prosodic information and the neck pitch angle, our system cannot reproduce the direction in which the original speaker looked. This study did not consider the average neck angle in its later analysis. The original motions tended to consist of large movements followed by small vibrating motions. On the other hand, the proposed model seems to generate a simple cyclic pattern, which produces a robot-like impression. We assumed that the speakers would prominently make a large head movement when they make a large voice and widely open their mouths. In other words, the amount of motion is not simply proportional to the magnitude of the prosodic features; large changes in the prosodic features produce larger movements. To clearly express this relationship, we introduce thresholds to restrict the motions shown in equations (8) and (9) by which a large driving force is only given to the model when the voice power and/or mouth openness are largely increasing. In this model, larger movements are generated when the prosodic features exceed the thresholds (p_th and m_th), and otherwise the restoring force generates small vibrating movements. The plot of restricted in **Figure 6** shows the movements by this model, where the thresholds p_th and m_th were empirically set to 1 and 10 so that large head movements prominently appear synchronized with a loud voice. With this improvement, motions similar to the original ones can be generated:

$$p(t) = \begin{cases} \rho(t) & (\rho(t) - \rho(t-1) >= p\_th) \\ 0 & (otherwise) \end{cases}, \tag{8}$$

$$m(t) = \begin{cases} \epsilon(t) & (\epsilon(t) - \epsilon(t-1) >= m\_th) \\ 0 & (otherwise) \end{cases}. \tag{9}$$

5. EVALUATION OF PROPOSED SYSTEM

5.1. Experimental Setup

First, we evaluated the impressions of the android's movements generated by the proposed method to show that it could generate natural and humanlike motions. The next section shows that the generated movements can be modulated for expressing internal states. The existing systems described in Section 2 basically reproduce the original human movements. Such methods do not enforce the synchronization between speech and motion. To show how our method improved the impressions of the androids, we compared it with a method in which the original speaker motions were reproduced in the android. This method copied the speaker's head and waist angles measured by IMU to the corresponding joint angles of the android (copy condition). Furthermore, to verify whether an unnatural impression is produced when the motion is not synchronized with the voice, we prepared another android motion by copying the original with a 1-s delay (non-synchronized condition) and two conditions for our model: with improvement of equations (8) and (9) and another without it. Then, we compared four conditions (proposed method with improvement, proposed method without improvement, copy, and non-synchronized, hereafter referred to as Proposed-I, Proposed-NI, Copy, and NoSync). We used a female type of android, named ERICA (**Figure 7**).

The participants evaluated the android's movements without any interactions (videotaped movements were used). We used the recorded human motion data and voice data mentioned in

Section 5. To verify that the method generated motions from any speaker's voice, this experiment used the data from both speakers. A comparison of **Figures 6A,B** revealed that the movement of Speaker M was larger and more frequent than that of Speaker H. Consequently, we examined the movement of different types of speakers.

The parameters in the model were the same as those used in Section 5. The head and waist motions in the Proposed-I and Proposed-NI conditions were generated from the speakers' voice data and used to control ERICA's corresponding joints (i.e., we used two joints). The neck joint was controlled as $\theta_{neck}(t) = \theta_{base}(t)$. No phase-shift between the neck and waist motions was assigned, that is, $\theta_{waist}(t) = \alpha_{waist}\theta_{base}(t)$, where $\alpha_{waist} = 0.1$. The android's lip movements were automatically generated from the voice by Ishi et al.'s method (Ishi, 2005) in all the conditions. Furthermore, we added blinking at random intervals normally distributed with a mean of 4 s and a SD of 0.5 s. **Figure 8** shows the kinematic structure of the android. We used joint 1 for blinking, joints 15 and 18 for trunk movements, and joint 13 for lip movements. The waist joint's control input was processed by a low-pass filter that moved over an average of the past 200 ms, since the time constant in the waist control is larger than in the neck control. The android's voice was provided by playing the original voice. Its movements were videotaped by covering the front waist-up image of its body (**Figure 9**), and the videos were used as the experimental stimuli.

Each participant evaluated all eight conditions (four types of motion × two types of speaker). The order of speakers was fixed

FIGURE 7 | Android ERICA.

FIGURE 9 | Video stimulus.

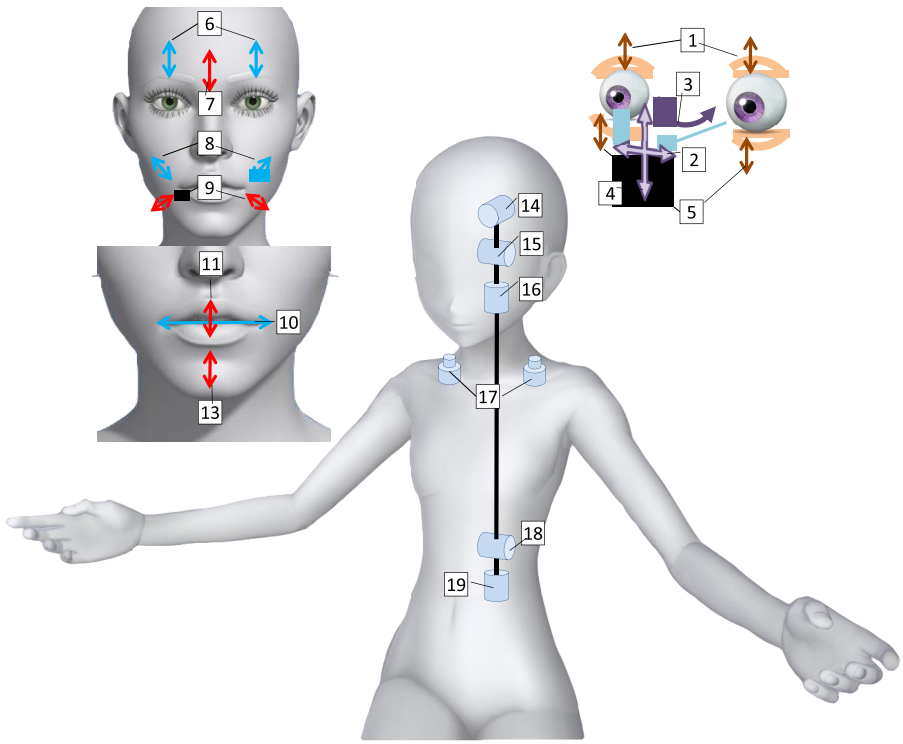

FIGURE 8 | Kinematic structure of proposed android.

(Speaker H was first) because we assumed that there would be no order effect related to the speakers. We counterbalanced the order of the four motion conditions. The participants could watch the video stimuli as many times as they liked.

5.2. Evaluation Measurements

The participants rated the naturalness and impression of each motion in the video on a 7-point Likert scale. For naturalness, the questionnaire asked whether "the neck and waist movements are natural" (naturalness). Piwek et al. (2014) revealed that natural movements improve the sense of intimacy toward agents. Then, the questionnaire asked participants to rate the statement, "I want to interact with the android," (will) as the willingness to interact with it.

5.3. Results

Fifteen Japanese speakers (12 males, 3 females, average age: 21.5, SD: 1.6) participated in the experiment. They were recruited in the same manner as described earlier. We conducted a two-way ANOVA with two within-subject factors (motion and speaker) and found no significant effect for the scores of will, but there was a significant trend of the motion factor for naturalness scores ($F(3, 42) = 2.33$, $p < 0.1$). In further examining the scores for Proposed-I and Proposed-NI scores, we found that some gave higher scores for Proposed-I, while others assigned opposite scores. This means that the effect of motion restriction in equations (8) and (9) was subjective. To comprehensively verify the effect of the proposed method, we combined the scores of Proposed-I and Proposed-NI scores into the scores of the "Proposed" condition by extracting the higher score between Proposed-I and Proposed-NI within-participants for each evaluation measurement.

A two-way ANOVA with two within-subject factors was conducted. Regarding the will score, there was a significant effect of motion factor ($F(2, 28) = 4.90$, $p < 0.05$). **Figure 10** compares the scores between the motion factors. Multiple comparisons by the Holm method revealed significant differences between conditions as Proposed > Copy ($p < 0.05$) (meaning the Proposed condition score is larger than that in the Copy condition, and the same hereafter) and Proposed > NoSync ($p < 0.05$). Regarding Naturalness, there was a significant effect of motion factor ($F(2, 28) = 6.51$, $p < 0.01$) as well as significant differences between conditions as Proposed > Copy ($p < 0.05$) and Proposed > NoSync ($p < 0.01$) (**Figure 11**).

We also found an interaction effect between motion and speaker factors for naturalness ($F(2, 28) = 5.89$, $p < 0.01$). **Figure 12** shows the average naturalness scores of six conditions. In the NoSync condition, the Speaker H score was significantly higher than that for Speaker M ($F(1, 14) = 12.64$, $p < 0.01$). In the Speaker M condition, there was a significant simple main effect for the motion factor ($F(2, 28) = 14.40$, $p < 0.01$). We conducted multiple comparisons by the Holm method and found significant differences among the conditions: Proposed > Copy ($p < 0.05$), Proposed > NoSync ($p < 0.001$), and Copy > NoSync ($p < 0.05$). The experimental results showed that the android motions generated by our model appeared more natural and motivated participants to talk more compared with copying of the human motions. This also means that our method outperformed the existing method by ideally reproducing the original human

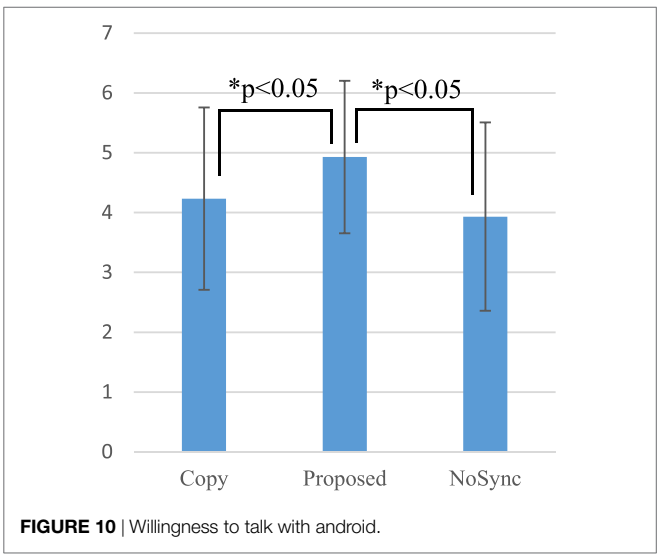

FIGURE 10 | Willingness to talk with android.

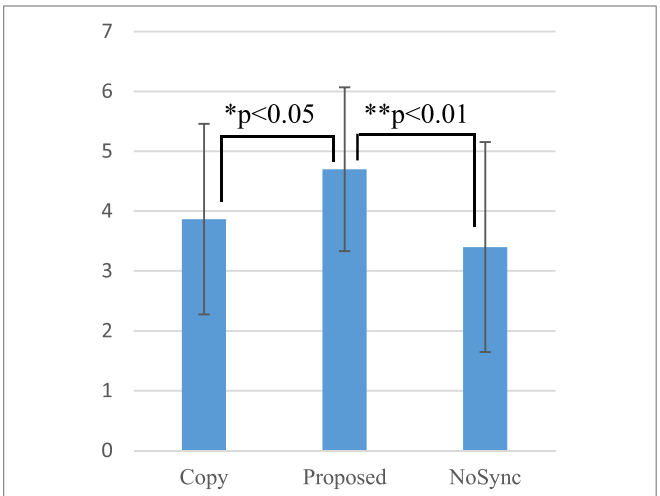

FIGURE 11 | Naturalness of neck and waist motions.

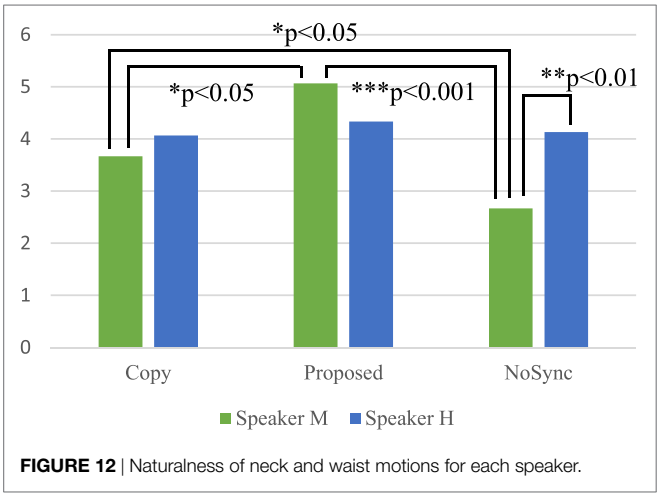

FIGURE 12 | Naturalness of neck and waist motions for each speaker.

movements. We tested two speakers and found this effect for both of them suggesting our model generated natural speech motions for various speech patterns.

5.4. Discussion

Why was the speech motion of our model evaluated as more natural than the copying of the human motions? We usually do not feel unnaturalness when we see a person who barely moves her head while speaking. We usually rationalize such a situation, for example, her body is tense due to some strain. However, when we see an android whose head barely moves, we might blame the system. We assumed that people would differently attribute the reason for the movement by a human or an android. If people clearly sense the correlation between the motion and prosodic features, which are typical in human speech, we can expect to avoid the attribution of a reason for non-humanlike motions by the android. This might explain why the proposed method's motions were better than the copied motions. Humans have many joints and muscles in their bodies and always make complex and subtle body motions, which cannot be perfectly reproduced in an android due to the limitations of its degrees of freedom and the characteristics of its actuators. That is, incomplete copying of human motions produces a more negative impression than uncopied motions. The proposed method, on the other hand, might be able to compensate for such incompleteness by expressively showing the motions that are strongly related to certain voice features. A similar result was obtained in a study of speech-driven lip motion generation by an android (Ishi et al., 2011). Perhaps the synchronization between voice and motion captures the essence of human-likeness in speech motions. In the future, with our model, we will investigate which kinds of motions essentially contribute to human-likeness by modulating the model parameters. In this sense, our proposed analytical model is helpful for studying the mechanism of human–robot interaction.

We found no significant differences between the Copy and NoSync conditions, although we expected the latter to be worse (e.g., since asynchrony between voice and lip movement on decreases the human-likeness of human and virtual characters (Tinwell et al., 2015)). However, this depended on the speaker. As shown in **Figure 12**, NoSync-Speaker M had a significantly lower score in naturalness score. We inferred a strong correlation between speech and motion for Speaker M but not for Speaker H. Since the participants attributed some meaning to the delayed motions in the NoSync-Speaker H condition, and thus they did not give bad scores. This result does not negate the necessity for a temporal synchronization between speech and motion, but we need further study to determine for which kinds of speech patterns this synchronization is necessary. Kirchhof (2014) suggested that the necessity of temporal synchrony between speech and gestures becomes smaller by loosening their semantic synchrony. A semantical correspondence between speech and motion might govern the participant impressions.

Comparing **Figure 6B** with **Figure 6A**, Speaker M's motions tend to have a larger magnitude. This means that she has a higher speech-motion correlation and moves her body more. Nevertheless, Speaker M's naturalness score is lower than that of Speaker H in the copy condition although the difference is not significant. Perhaps motion copying failed to take into account the coordination throughout the entire body. Speaker M might greatly move not only her neck and waist but also other body parts to balance her entire body during speech; however, the system does not generate those motions. The participants might feel some unnaturalness in the android's entire body movements with the loss of coordination. Since the proposed method reproduces well-coordinated motions with only the neck and waist, the participants felt that they were natural, even though they were different from the original speaker's motion. In their study of humanoid motion generation, Gielniak et al. (2013) showed that the coordinated motions on multiple joints produced humanlike motions. Their method emulates the coordinated effects of human joints that are connected by muscles, and their experimental result showed that this coordination provides the impression of human-likeness in the robot motions. These results suggest that motion coordination under a physical body's constraints is critical for giving a natural impression of motions.

The preference for Proposed-I or Proposed-NI motion depended on the participants. Some believed that Proposed-NI is better since it moved more, but others felt the opposite. People tend to prefer a person who mimics them: the chameleon effect (Chartrand and Bargh, 1999). Accordingly, we infer that the participants preferred motions that resembled their own. We could choose either Proposed-I or Proposed-NI for the android motion based on the personality of the conversation partner by prior personality questionnaires or estimations based on a multimodal sensor system. Further study is needed to investigate how to generate an android's speech motion to suit the personality of the conversation partner.

6. EVALUATION OF CAPABILITY OF EMOTIONAL EXPRESSION

6.1. Purpose of Experiment

This section shows how the generated motions can be modulated to express the internal state of an android. The parameters in the proposed model correspond to muscle stiffness, which depends on the speaker's mental state, e.g., stress and emotion (Sainsbury and Gibson, 1954; Fridlund et al., 1986). In other words, perhaps we can modulate the speech motions based on the mental states by tuning the parameters of the spring-damper model shown in equation (1).

We experimentally showed that our participants properly modulated the android's motions based on its desired internal states.

6.2. Experimental Setup

Some researchers revealed that motion properties, e.g., magnitude and speed, vary by emotional states (Amaya et al., 1996; Michalak et al., 2009; Masuda and Kato, 2010; Gross et al., 2012). Consequently, we transformed independent variables j, d, and k in equation (1) to ω_0, ξ, and ϕ in equation (10) so that the participants can intuitively change the speed and magnitude of the android's motions. ω_0, which is the natural angular frequency, is equivalent to the time that a motion needs to be converged. ξ is the damping ratio. The damping property depends on ξ ($\xi > 1$: over damping, $\xi = 1$: critical damping, $\xi < 1$: damped oscillation).

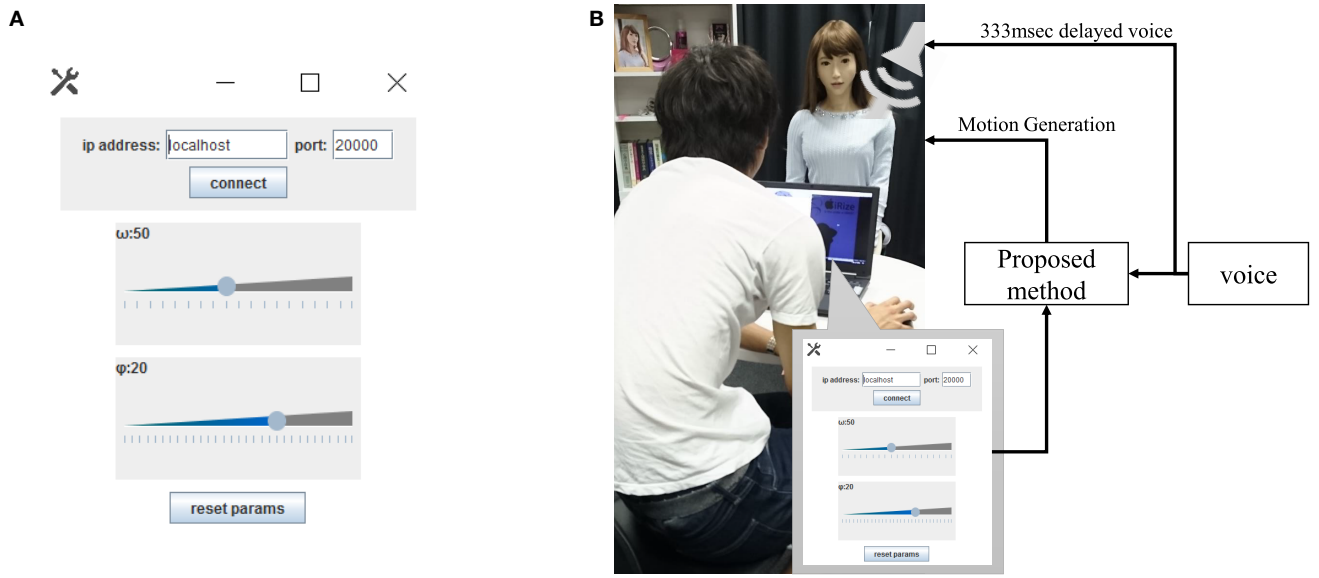

FIGURE 13 | Experimental setup. **(A)** Interface for adjusting motion parameters. **(B)** Block diagram.

ϕ is the reciprocal of inertia, and the magnitude of motion is proportional to ϕ:

$$\ddot{\theta}_{base}(t) + 2\xi\omega_0\dot{\theta}_{base}(t) + \omega_0^2\theta_{base}(t) = \phi T(t)Dir(t)$$
$$j = \frac{1}{\phi}, k = \frac{\omega_0^2}{\phi}, d = \frac{2\xi\omega_0}{\phi}. \quad (10)$$

In this experiment, we set the damping ratio ($\xi = 1$). The human head rhythmically moves in synchronization to speech. Since this movement is not damped, the damping ratio should be $\xi \approx 1$. The participants modulated ω_0 and ϕ through the graphical user interface (GUI) in **Figure 13A**. The range of ω_0 was set at 1–10 in 0.5 steps, and the range of ϕ was set to 10^0–10^2 in $10^{0.05}$ steps. The android speaks by playing a prerecorded voice while its neck and mouth move in synchrony with its voice. The voice was played from speakers on its head. We delayed the voice by 333 ms for synchronization with the motion because of the calculation delay of the prosodic features and the communication time with the android. The setting's block diagram is shown in **Figure 13B**. The motion properties were changed immediately in response to the changes in ω_0 and ϕ. The participants explored the GUI to find proper ω_0 and ϕ values for expressing the desired emotion while looking at the android's motions.

6.3. Experimental Conditions and Measurements

We chose four emotional states (happy, bored, relaxed, and tense) based on Russell's emotion model (Russell, 1980), and the participants searched for the values of ω_0 and ϕ values to feel the android's motion expressed by each emotion.

The difficulty of tuning the motions depends on the speech voice (the tuning was easy for some voices but not for the others). Therefore, the speech voice was fixed in all the emotions and for all participants. This voice sample was approximately a 1-min recording of a female experimental assistant reading a news story aloud while maintaining a neutral mental state.

The android changed its eye direction and facial expression to express its emotional state. This is because people felt difficulty exploring the parameters without facial expression in the preliminary trial. The eye movements and facial expression follow Ekman's report (Ekman and Friesen, 1981). In happy and relaxed emotions, the android smiled by lifting the angle of her mouth up and cyclically moving her eyes horizontally. In bored and tense emotions, it grimaced with her eyelids down and rolled her eyes downward.

The participants sat in front of the android and adopted the parameters ω_0 and ϕ using the GUI shown in **Figure 13A**. For each emotion preset, they searched for the parameter values that produced the android's desired emotional expression. The order of the four emotion conditions was counterbalanced among the participants.

To evaluate the degree of easiness of this modulation, we measured how long it took to modulate the motions. To the evaluate the degree of satisfaction with the modulation, the participants also answered a 7-point Likert scale questionnaire (1: unsatisfactory, ~4: not sure, ~7: satisfactory). Based on satisfactory scores, we wanted to evaluate whether android motions were produced as the participants expected. High scores denote successfully produced motions.

6.4. Results

Twelve subjects (six males, six females, average age: 20.4, SD: 1.0) participated in the experiment. They were recruited in the same manner for the above two experiments.

Table 1 shows the elapsed needed time to modulate the motions. The voice used in this experiment was about 1 min, and the results revealed that the participants fixed the parameters within 3–5 repetitions. They did not take that much time for

motion tuning, even though the results were not compared with those of other methods. **Figure 14** shows the average degree of satisfaction with the modulated motions. Because the average scores of all the emotion exceeded four, the participants felt the modulated motions expressed the desired emotion. Even tough some experienced difficulty tuning the motion due to incongruity between the manner of speaking (speech in neutral emotion) and the emotion, they successfully produced the desired emotional motion. A statistical test revealed that the average scores of happy, relaxed, and tense were significantly higher than four ($p < 0.01$), but not bored ($p < 0.1$).[2] Expressing the bored motion was difficult with the trunk motion.

Figure 15 shows examples of modulated motion (time series of head angle) and corresponding voice data. There is a tendency for the magnitude and speed of the motions to increase in the order of bored, relaxed, happy, and tense. This result shows that the modulated motion properties are different based on the emotional states.

[2]We used *t*-test if the normality was assumed by the Shapiro–Wilk test; otherwise the Wilcoxon rank sum test was used.

TABLE 1 | Elapsed time to modulate the motions [s].

	Happy	Bored	Relaxed	Tense
Average	305	294	188	278
SD	205	266	132	213

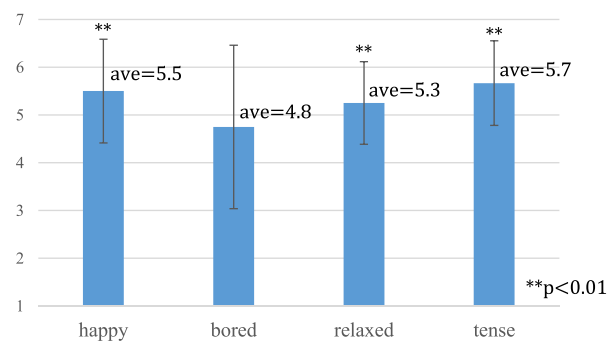

FIGURE 14 | Degree of satisfaction with the modulated motions.

6.5. Discussion

One the other hand, the existing learning-based systems need to additionally collect human emotional motion data when we generate the android's emotional expressions that are not included in the learning data. The proposed model has an advantage with facility in the motion development.

Figure 14 shows the participants were not sufficiently satisfied by the generated motions of bored. Such motions were expressed by small trunk movements, but they might have thought that just making small motions is insufficient for the bored state. They wanted to add a facial expression or a voice to distinctively express boredom but they could not; therefore, their degree of satisfaction fell. To precisely express the emotional states in the android, the facial expression, loudness of voice and speech rate must be involved in the motion modulation. In the future, our system needs to implement this idea.

The speech voice was fixed in the experiment, but the speech pattern is usually changed owing to the speaker's emotional state. For example, the voice usually becomes louder and faster in the tense and happy states and lower and slower in the bored and relaxed states. This is similar to the relation between the magnitude and speed of the motions and the emotional state shown in **Figure 15**. A tense state brings larger and faster motions and a relaxed state brings smaller and slower motions. This suggests that the motion and speech patterns are similarly changed by the emotional state, such as the magnitude, and the speed of voice and motion become larger and faster in the tense and happy states and smaller and slower in the bored and relaxed states. If this relation generally holds, the proposed model can generate emotional speech motions based on emotional voices without tuning the parameters. Furthermore, developing a system of automatic emotional voice and motion generation is possible by integrating our model with the emotional voice synthesization (Murray and Arnott, 2008). Future work should reveal the relation for this system. Here, the relation might depend on the age and gender of

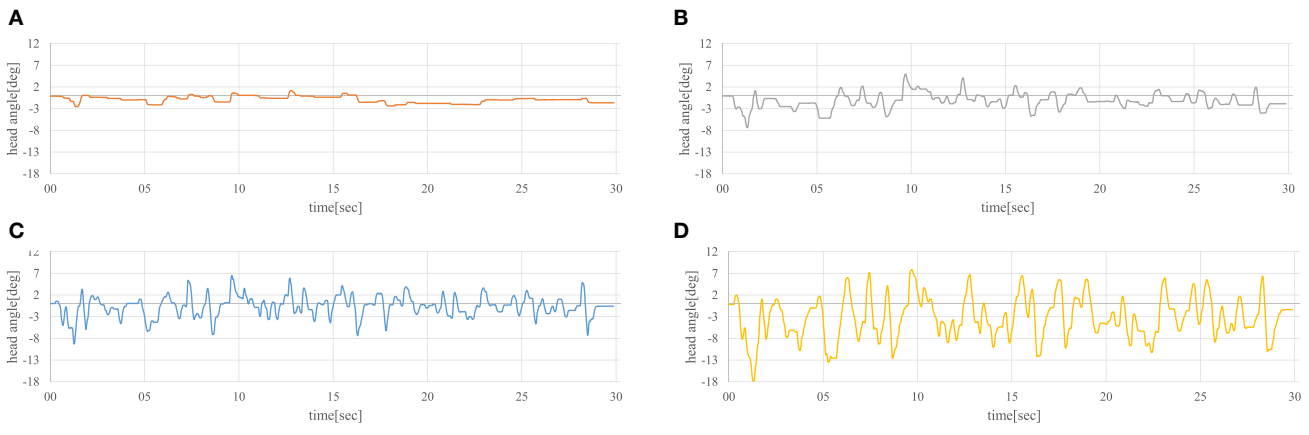

FIGURE 15 | Examples of emotional motion modulated by the participants. **(A)** Bored motion, **(B)** relaxed motion, **(C)** happy motion, and **(D)** tense motion.

the speaker since a gender difference exists in non-verbal behaviors (Frances, 1979). For example, male speakers might express more emotional motions. To use our model in different types of androids, such dependencies must be investigated in future work.

7. FUTURE WORKS

People tend to anticipate some properties of a speaker's motion when they hear her voice, for example, a vigorous movement is expected from a cheerful voice. We found fewer differences between the proposed and copy methods in the Speaker H comparison (**Figure 12**) because the participants might have expected much movement from her voice. The proposed model generates the motion based on pitch, power, and vowel information, but we also expect other voice features to be related. Identifying the relation between voice features and expected motion is critical. Moreover, human movement has some randomness that should be implemented in the method. This study tested the proposed method for Japanese speech, but the idea of synchronization between motions and voice features could more effectively work for English speech, since English speakers show a higher correlation between speech and head movement than Japanese speakers (Yehia et al., 2002). Furthermore, the results here were only obtained from university students. Although elderly people prefer a human-like to a robot-like appearance, younger adults prefer the opposite (Prakash and Rogers, 2014). Therefore, the influence of the android's appearance and the participants' age must also be investigated in future work.

We evaluated our proposed method for subjective impressions but not for behavioral aspects. The participant behavior or attitude toward the android might change if they had a feeling of humanlikeness about it. For instance, gaze behavior (Shimada and Ishiguro, 2008) and posture are influenced by the relationship between conversation partners. In future work, a behavioral evaluation is necessary to scrutinize the method's effects.

8. CONCLUSION

We developed a method that automatically generates humanlike trunk motions of a conversational android from its speech in real time. By enforcing the speech motions that are strongly related to prosodic features, the method compensates for the negative impressions caused by incomplete copying of human motions from which conventional methods suffer. By simulating human trunk movement based on a spring-damper dynamical model, the motion can be modulated based on the android's internal states. Our experimental results show that the android motions generated by the model appear natural and motivate people to talk more with the android, even though the generated motions are different from the original human motions. The results also suggest that the model can generate humanlike motions for any speech pattern. The additional capability of enforcing multimodal synchronization must be applicable to other types of humanoid robots and other modalities than speech and motion. Such possibilities must be investigated in future work.

ETHICS STATEMENT

The study was approved by the Ethics Committee of the Advanced Telecommunications Research Institute International (Kyoto, Japan). In the beginning of the experiment, we explained about this study to the subjects who were university students and received informed consent from them. We used a job-offering site for university students, and the subjects were recruited.

AUTHOR CONTRIBUTIONS

KS proposed the idea of this speech-driven trunk motion generating system, build the experiment system, conducted the experiment, analyzed the result, and wrote this article. TM designed the experiment, analyzed and evaluated the result. CI worked in the system development, especially sound signal analysis. HI proposed the basic idea to generate humanlike movements under a mechanical limitation of the android (i.e., limited number of joint).

FUNDING

This research was supported by the Japan Science and Technology Agency, ERATO, ISHIGURO symbiotic Human-Robot Interaction Project, Grant Number JPMJER1401.

REFERENCES

Amaya, K., Bruderlin, A., and Calvert, T. (1996). "Emotion from motion," in *Graphics Interface*, Vol. 96, Toronto, 222–229.

Andrist, S., Tan, X. Z., Gleicher, M., and Mutlu, B. (2014). "Conversational gaze aversion for humanlike robots," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, 25–32.

Bolinger, D. (1985). *Intonation and Its Parts: Melody in Spoken English*. Stanford: Stanford University Press.

Busso, C., Deng, Z., Neumann, U., and Narayanan, S. (2005). Natural head motion synthesis driven by acoustic prosodic features. *Comput. Animat. Virtual Worlds* 16, 283–290. doi:10.1002/cav.80

Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect. *J. Pers. Soc. Psychol.* 76, 893–910. doi:10.1037/0022-3514.76.6.893

Ekman, P., and Friesen, W. V. (1981). The repertoire of nonverbal behavior: categories, origins, usage, and coding. *Nonverbal Commun. Interact. Gesture* 57–106.

Eriksson, P.-O., Zafar, H., and Nordh, E. (1998). Concomitant mandibular and head-neck movements during jaw opening-closing in man. *J. Oral Rehabil.* 25, 859–870. doi:10.1046/j.1365-2842.1998.00333.x

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi:10.1038/415429a

Foster, M. E., and Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Lang. Resour. Eval.* 41, 305–323. doi:10.1007/s10579-007-9055-3

Frances, S. J. (1979). Sex differences in nonverbal behavior. *Sex Roles* 5, 519–535. doi:10.1007/BF00287326

Fridlund, A. J., Hatfield, M. E., Cottam, G. L., and Fowler, S. C. (1986). Anxiety and striate-muscle activation: evidence from electromyographic pattern analysis. *J. Abnorm. Psychol.* 95, 228. doi:10.1037/0021-843X.95.3.228

Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., and Nishida, S. (2014). Audiovisual integration in the human perception of materials. *J. Vis.* 14, 1–20. doi:10.1167/14.4.12

Gielniak, M., Liu, K., and Thomaz, A. (2013). Generating human-like motion for robots. *Int. J. Robot. Res.* 32, 1275–1301. doi:10.1177/0278364913490533

Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-Shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Mov. Sci.* 31, 202–221. doi:10.1016/j.humov.2011.05.001

Hashimoto, T., and Kobayashi, H. (2009). "Study on natural head motion in waiting state with receptionist robot SAYA that has human-like appearance," in *Proceedings of 2009 IEEE Workshop on Robotic Intelligence in Informationally Structured Space*, Nashville, TN, 93–98.

Ishi, C. T. (2005). Perceptually-related F0 parameters for automatic classification of phrase final tones. *IEICE Trans. Inform. Syst.* 88, 481–488. doi:10.1093/ietisy/e88-d.3.481

Ishi, C. T., Liu, C., Ishiguro, H., and Hagita, N. (2011). "Speech-driven lip motion generation for tele-operated humanoid robots," in *Auditory-Visual Speech Processing*, Volterra, 131–135.

Ishi, C. T., Liu, C., Ishiguro, H., Hagita, N., Robotics, I., and Labs, C. (2012). *Evaluation of Formant-Based Lip Motion Generation in Tele-Operated Humanoid Robots*. Vilamoura: IEEE, 2377–2382.

Jia, J., Wu, Z., Zhang, S., Meng, H. M., and Cai, L. (2014). Head and facial gestures synthesis using PAD model for an expressive talking avatar. *Multimed. Tools Appl.* 73, 439–461. doi:10.1007/s11042-013-1604-8

Kirchhof, C. (2014). "Desynchronized speech-gesture signals still get the message across," in *International Conference on Multimodality*, Hongkong.

Komatsu, T., and Yamada, S. (2011). Adaptation gap hypothesis: how differences between users' expected and perceived agent functions affect their subjective impression. *J. Syst. Cybern. Inform.* 9, 67–74.

Kondo, Y., Takemura, K., Takamatsu, J., and Ogasawara, T. (2013). A gesture-centric android system for multi-party human-robot interaction. *J. Hum.Robot Interact.* 2, 133–151. doi:10.5898/JHRI.2.1.Kondo

Laban, R. V. (1988). *The Mastery of Movement*. Princeton Book Co. Pub.

Larsen, R. J., and Shackelford, T. K. (1996). Gaze avoidance: personality and social judgments of people who avoid direct face-to-face contact. *Pers. Individ. Dif.* 21, 907–917. doi:10.1016/S0191-8869(96)00148-1

Le, B. H., Ma, X., and Deng, Z. (2012). Live speech driven head-and-eye motion generators. *Vis. Comput. Graph.* 18, 1902–1914. doi:10.1109/TVCG.2012.74

Liang, C.-C., and Chiang, C.-F. (2006). A study on biodynamic models of seated human subjects exposed to vertical vibration. *Int. J. Ind. Ergon.* 36, 869–890. doi:10.1016/j.ergon.2006.06.008

Linder, A. (2000). A new mathematical neck model for a low-velocity rear-end impact dummy: evaluation of components influencing head kinematics. *Accid. Anal. Prev.* 32, 261–269. doi:10.1016/S0001-4575(99)00085-8

Masuda, M., and Kato, S. (2010). "Motion rendering system for emotion expression of human form robots based on Laban movement analysis," in *Proceedings of the 19th International Symposium in Robot and Human Interactive Communication*, Viareggio, 324–329.

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., and Schulte, D. (2009). Embodiment of sadness and depression–gait patterns associated with dysphoric mood. *Psychosom. Med.* 71, 580–587. doi:10.1097/PSY.0b013e3181a2515c

Murray, I. R., and Arnott, J. L. (2008). Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Comput. Speech Lang.* 22, 107–129. doi:10.1016/j.csl.2007.06.001

Nakano, A., and Hoshino, J. (2007). Composite conversation gesture synthesis using layered planning. *Syst. Comput. Japan* 38, 58–68. doi:10.1002/scj.20532

Nakazawa, M., Nishimoto, T., and Sagayama, S. (2009). "Behavior generation for spoken dialogue agent by dynamical model," in *Proceedings of Human-Agent Interaction Symposium (in Japanese)*, Tokyo, 2C-1.

Piwek, L., McKay, L. S., and Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition* 130, 271–277. doi:10.1016/j.cognition.2013.11.001

Pollard, N. S., Hodgins, J. K., Riley, M. J., and Atkeson, C. G. (2002). "Adapting human motion for the control of a humanoid robot," in *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, Vol. 2, Washington, DC, 1390–1397.

Prakash, A., and Rogers, W. A. (2014). Why some humanoid faces are perceived more positively than others: effects of human-likeness and task. *Int. J. Soc. Robot.* 7, 309–331. doi:10.1007/s12369-014-0269-4

Russell, J. A. (1980). A circumplex model of affect. *Personal. Soc. Psychol.* 39, 1161–1178. doi:10.1037/h0077714

Sainsbury, P., and Gibson, J. G. (1954). Symptoms of anxiety and tension and the accompanying physiological changes in the muscular system. *J. Neurol. Neurosurg. Psychiatr.* 17, 216–224. doi:10.1136/jnnp.17.3.216

Sakai, K., Ishi, C. T., Minato, T., and Ishiguro, H. (2015). "Online speech-driven head motion generating system and evaluation on a tele-operated robot," in *Proceedings of the 24th International Symposium in Robot and Human Interactive Communication*, Kobe, 529–534.

Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., and Joublin, F. (2012). Generation and evaluation of communicative robot gesture. *Int. J. Soc. Robot.* 4, 201–217. doi:10.1007/s12369-011-0124-9

Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2008). Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE. Trans. Pattern. Anal. Mach. Intell.* 30, 1330–1345. doi:10.1109/TPAMI.2007.70797

Shimada, M., and Ishiguro, H. (2008). "Motion behavior and its influence on human-likeness in an android robot," in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington, DC, 2468–2473.

Tinwell, A., Grimshaw, M., and Nabi, D. A. (2015). The effect of onset asynchrony in audio visual speech and the uncanny valley in virtual characters. *Int. J. Mech. Robot. Syst.* 2, 97–110. doi:10.1504/IJMRS.2015.068991

Watanabe, M., Ogawa, K., and Ishiguro, H. (2015). "Can androids be salespeople in the real world?" in *Proceedings of the ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Seoul, 781–788.

Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R. (2004). InterActor: speech-driven embodied interactive actor. *Int. J. Hum. Comput. Interact.* 17, 43–60. doi:10.1207/s15327590ijhc1701_4

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *J. Phon.* 30, 555–568. doi:10.1006/jpho.2002.0165

Yoshikawa, M., Matsumoto, Y., Sumitani, M., and Ishiguro, H. (2011). "Development of an android robot for psychological support in medical and welfare fields," in *Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics*, Phuket, 2378–2383.

Zafar, H., Nordh, E., and Eriksson, P.-O. (2000). Temporal coordination between mandibular and head-neck movements during jaw opening-closing tasks in man. *Arch. Oral Biol.* 45, 675–682. doi:10.1016/S0003-9969(00)00032-7

Zafar, H., Nordh, E., and Eriksson, P.-O. (2002). Spatiotemporal consistency of human mandibular and head-neck movement trajectories during jaw opening-closing tasks. *Exp. Brain Res.* 146, 70–76. doi:10.1007/s00221-002-1157-y

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.