EasyChair Preprint

№ 15113

Listening Head Motion Generation for Multimodal Dialog System

Tamon Mikawa, Yasuhisa Fujii, Yukoh Wakabayashi,
Kengo Ohta, Ryota Nishimura and Norihide Kitaoka

September 27, 2024

Listening Head Motion Generation for Multimodal Dialog System

1st MIKAWA, Tamon
Toyohashi University of Technology
Aichi, Japan

2nd FUJII, Yasuhisa
Google DeepMind,
Tokyo, Japan

3rd WAKABAYASHI, Yukoh
Toyohashi University of Technology
Aichi, Japan

4th OHTA, Kengo
National Institute of Technology, Anan College
Tokushima, Japan

5th NISHIMURA, Ryota
Tokushima University
Tokushima, Japan

6th KITAOKA, Norihide
Toyohashi University of Technology
Aichi, Japan

***Abstract*—This paper addresses the listening head generation (LHG), i.e., an avatar head motion in dialogue systems. In face-to-face conversations, head motion is a modality frequently used by both listeners and speakers. Listeners, in particular, tend to leverage head motion along with other backchanneling cues to react to the speaker and regulate the flow of the conversation. The type of head motion during dialogues varies between cultures and individuals, which implies that head motion generation for natural communication requires considering them. Additionally, existing works for head motion generation have primarily tackled speaker head generation, with limited work on listeners. In this study, we have created a multimodal dataset of casual Japanese conversation and a scalable, real-time LHG model that adapts to individual differences in head motion. We also developed the LHG that reflects individual tendencies via fine-tuning the model. The proposed models were evaluated through subjective experiments rated by four testers. The results showed that the proposed models successfully generated natural head motion and improved the appropriateness of head motion by focusing on individual tendencies. Further analysis was conducted to compare the differences between our method and actual human motion.**

***Index Terms*—Listening Head Generation, Dialogue System, Japanese Casual Dialogue Dataset, Motion Synthesis**

I. Introduction

In recent years, with the advancement of artificial intelligence technology, dialog systems are improving and are increasingly being used for various applications. A dialog system is an automated system designed to perform specific tasks, such as providing customer service, visitor information, companionship, etc., through communication with humans. Realization of these systems requires the development of natural and advanced dialog capabilities close to those of humans. During face-to-face dialogs between humans, not only linguistic information is exchanged, but also non-linguistic information such as gaze, gestures, and prosody (e.g., intonation, stress, rhythm) are frequently used. Humans can communicate complex information to each other quickly and smoothly by simultaneously and integratively understanding and generating these varoius modalities. To implement these functions in a dialog system, it is necessary to use a computer-generated (CG) avatar as the system interface, and to present modalities other than language to the user.

Among the modalities other than language which are used by humans during interpersonal communication, head movements in particular are frequently used to convey important information. Munhall et al. [1] reported that head movements and facial expressions are strongly correlated with the amplitude and pitch of the speaker's voice, and assist in the transmission of linguistic information. Otsuka et al. [2] analyzed the communicative functions of head movements of both listeners and speakers, and reported on the functions of each head movement and the frequency of their appearance. They determined, for example, that listeners' head movements were often used to show attentive listening and agreement, and that multiple head movements functions can be employed simultaneously. They also identified the functions of speakers' head movements, such as emphasis, confirmation of understanding, and promotion of listener involvement in the dialog, and reported the possibility that a speaker's head movement can influence the head movement of listeners.

As these previous studies have shown, head motion plays a crucial role in dialog, and its use by CG avatars is expected to significantly contribute to the realization of more natural and human-like dialog systems. However, it has also been reported that there are cultural and language-related differences in the use of head motion during dialogs. For example, Koda et al. [3] reported differences in the timing and frequency of backchannels, including head motion, between native speakers of Japanese and English. Furthermore, they found that using a dialog system that provides backchannels tailored to the user's cultural background can create a sense of familiarity and increase user speaking time, emphasizing the importance of culturally adapting dialog systems for users.

Our research is currently focused on realizing a dialog system that can communicate smoothly with users in Japanese, while providing them with an atmosphere of cultural familiarity. Compared to conversations among English speakers, Japanese dialog features more backchannels, including more head move-

ments, such as nodding [3], therefore the ability to precisely model these head movements is crucial for realizing a natural, multimodal Japanese dialog system.

The goal of this study is to construct a deep learning-based, listening head motion generation model. To achieve this, we first collect one-on-one multimodal dialog data from Japanese speakers. We then propose a novel model that generates real-time head response movements, based on the multimodal cues obtained from users' speech and head movements during the recorded conversations.

II. Related Work

Early studies on generating head motion and gestures [4, 5] focused on rule-based approaches. Maatman et al. [5] organized the behavior of listeners during conversations based on psycholinguistic factors, and proposed a system that generates listener body movement in real-time based on the correlation between non-semantic features of the speaker's voice and body movements, and the listener's intended response. They reported that their proposed system made users feel they were being "listened to" by the nodding and shaking of the avatar's head while it listened to the user's utterances. Huang et al. [6] constructed a backchannel generation model for rapport formation using a Conditional Random Field (CRF), and obtained higher ratings in a subject experiment compared to a rule-based model. Since the advent of deep learning, it has become possible to generate more natural head movement during dialog compared to previous rule-based methods. Ding et al. [7, 8] proposed a method to generate speaker head motion during speech based on the speaker's voice. They compared the performance of deep learning models with different architectures, and found that the Bidirectional Long Short-Term Memory (BLSTM) model demonstrated superior results. They also investigated using audio features for head motion generation and found that using a log Mel-scale filter-bank (FBank) yielded the best performance. Alexanderson et al. [9] used a diffusion model to generate more natural gestures during avatar speech by synchronizing motion generation with audio of its speech.

More recently, research on Talking Head Generation (THG) [10, 11, 12], which generates sequences of head motions and facial landmarks synchronized with speech, has garnered significant attention. Wang et al. [13] proposed a method for generating facial expressions synchronized with speech, realizing the generation of rich expressions through THG. Greenwood et al. [14] succeeded in synthesizing more dynamic head motion by generating multiple consecutive frames in a single step during the staged generation of head motion from speech. On the other hand, researchers have also focused on generating head motion and facial expressions during listening, known as Listening Head Generation (LHG) [15]. Liu et al. [16] proposed CustomListener, a framework that enables control of the listener's behavior through text prompts, however they did not address real-time head motion generation in response to speech, or integration of their framework into a dialog system. Zhou et al. [17] constructed the ViCo dataset, which is specifically designed to generate listener head motion. They proposed a baseline model that uses the speaker's speech and visual cues about the speaker as input, and then generates listener head motion and facial expressions in real-time using an LSTM-based model. In fact, many of the studies cited here utilize LSTM-based models, which have also demonstrated high performance in many other motion generation tasks [18]. However, as reported by Ding et al., there are limits when scaling up LSTM-based models for motion generation. Therefore, based on the findings of Yu et al. [19], we propose a model that replaces the masked multi-head attention module of the Transformer [20] with a Uni-Directional LSTM, and incorporates an attention module for integrating multiple modalities with different frame rates. Our proposed model has the same basic structure as Transformer and our modifications deepen the model. The parameter size of the proposed model is about 0.5B, which is significantly larger than the 248K of Zhou's model [17], thus it possesses higher expressive ability.

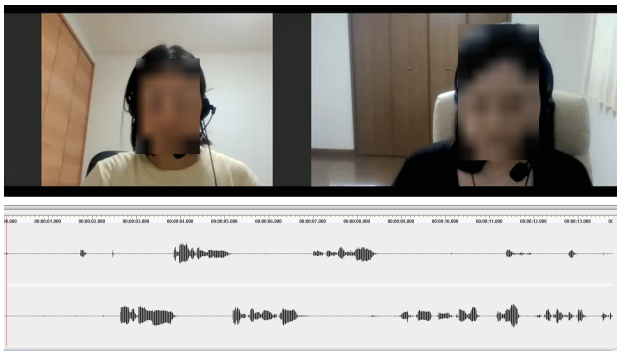

Fig. 1: Example of data from our Japanese dialog dataset.

III. Dataset Construction

To realize smoother and more natural automated dialog, we created a multimodal dataset of casual Japanese conversations for generating and analyzing the head motion tendencies of Japanese listeners, in order to build a dialog system adapted to Japanese users. The video and audio data was collected from one-on-one, casual conversations between Japanese speakers recorded on Zoom, obtained via crowdsourcing. Figure 1 shows a screenshot from one of the Zoom videos and a sample of the collected audio data. The participants wore headsets to better hear their partner's voice, and engaged in one-on-one dialog while looking at their conversation partner's face on the screen. By having each speaker use a microphone and recording each speaker's voice separately, without their dialog partner's voice, we were able to collect high quality recordings of each individual's voice with little distortion.

In this research and in future studies, we plan to focus on the head movements of dialog participants, therefore it was important to capture their head movement as accurately as possible in order to clarify the relationship between physical head movement patterns and the content of the dialog. To achieve this, we filmed the participants as they faced their monitor screen during the dialogs, with the area from their shoulders to the top of their head visible in the videos. Furthermore, to achieve

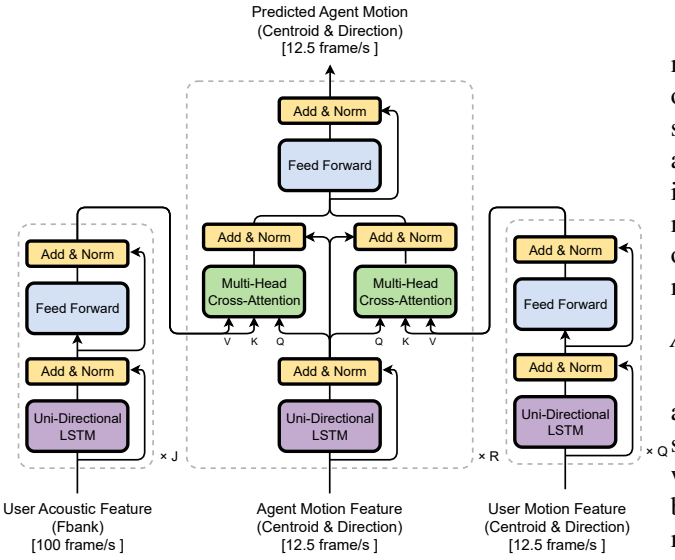

Fig. 2: Architecture of our proposed LSTM-Transformer model.

more accurate analysis of 'pure' dialog activity, we prohibited behaviors such as eating or drinking, looking away from the screen, leaving one's seat, and any other actions involving large body movements. During recording, the dialog participants freely conversed about one, pre-specified topic during each session. The 15 topics were "eating and food," "fashion," "travel," "sports," "manga and games," "housework," "school," "smartphones," "part-time jobs," "animals," "weather," "goals and future plans," "manners," "living environment their homes," and "the future of Japan."

The data collected from each dialog session consists of each participant's voice, frontal head shot video of each participant, and monaural audio of both dialog participants as recorded by Zoom. Each dialog session was approximately 10 minutes long. The dataset contains 478 dialog sessions featuring 25 different speakers, with a total duration of about 90 hours. The audio data recorded separately by each participant's microphone was saved in a 16-bit, PCM format with 16 kHz sampling. The Zoom video data of each participant was saved in MPEG4 format with a resolution of 720 × 1280 and a frame rate of 25 fps. The individually recorded audio and Zoom audio were synchronized using autocorrelation of audio power.

Human annotators also performed speaker annotation of the synchronized video and audio data, identifying which participant was speaking and which was listening at any point during the dialog. The annotators referred to the casual conversation data and annotated the time intervals during which each dialog participant was considered to be 'holding the floor'. However, we assumed that the floor was not always held exclusively, and that both dialog participants might speak simultaneously, especially before and after speaker changes. In other words, we allowed for intervals of overlapping speech.

IV. Proposed Method

In this study, our goal is to generate avatar head response motion during listening, in real-time, using multimodal cues obtained from the user's speech and head movement while speaking. Therefore, we constructed a deep learning model that autoregressively generates the agent's (listener's) head motion in future frames based on the user's (speaker's) speech and head motion up to a certain point (frame). To investigate which type of model is more suitable for generating the listener's head motion, we constructed and compared two models.

A. Motion and Acoustic Feature

Regarding the features we used to train our model, for acoustic features we used log Mel-scale filterbank (FBank). The sampling frequency of the speech data used in the experiments was 16 kHz, and short-time Fourier transform was performed by dividing the data samples into 25 ms frames using a Hamming window with a 10 ms overlap for each frame, similar to previous research [8]. The FBank was initially set to 26 dimensions, then the logarithmic power of speech was added to make it 27 dimensions. First- and second-order time differences of these 27-dimensional features were also used, resulting in 81-dimensional features per frame. Here, the calculation of the time-difference features is different from the Δ features often used in speech research, since our features do not refer to information in future feature frames; instead, they are calculated as the simple difference between the current frame and the previous frame, as shown in the following equation:

$$df_t = f_t - f_{t-1} \tag{1}$$

where df_t represents the difference features at time t, and f_t denotes the feature frame at time t.

To extract the features of head motions, we used Mediapipe [21] to detect facial landmarks and calculate the rotation angles (Euler angles) of three degrees of freedom: pitch, yaw, and roll, from the coordinates of each landmark. In addition, the centroid coordinates of the facial landmarks were used as the three degrees of freedom for the face position coordinates. By adding the first- and second-dynamic features to these six degrees of freedom, as in the calculation of the acoustic features, we finally obtained 18-dimensional features for each frame.

B. Model Architecture

The listener head movement generation model used in this study has a structure similar to Transformer, as shown in Figure 2. Features of the user's speech, user's head motion, and agent's head motion at aligned time points of each frame are used as inputs, and the model then autoregressively generates the agent's head motion for the following frame, by producing a head motion sequence. Although the frame rates of the speech features and head motion features differ, the use of Cross-Attention allows the integration of data with different sequence lengths. During model training, three frames (equivalent to 0.24 s) of head motion features are output in a single inference step, and the loss is calculated. This is done because there is

little variation in the head motion features between frames. If the model were to infer one frame at one step, it might learn to simply output the actions from the immediately preceding step without much change. The method described above forces the model to consider variations in the head motion sequence when generating future motions. In our experiments, the number of stacked blocks J, Q, and R, shown in Figure 2, was set to 12, and the number of units in each layer was set to 1,024 dimensions. The Feed Forward Layer has a bottleneck structure with 256 units at the bottleneck, and uses ReLU as the activation function. Our proposed model will hereafter be referred to as "LSTM-Transformer".

V. Experiments

We trained the LSTM-Transformer model described in Section IV to generate the head motion of a virtual agent when participating in the dialog as a listener, in real-time, based on the speech and head motion of the human speaker. The model was trained using our multimodal dataset of casual, Japanese conversation, which is described in Section III. We then evaluated the model experimentally and analyzed its performance. When we identify the virtual agent as the listener here, it means that they do not 'hold the floor'. The avatar's status as the listener was determined using the dataset's annotation information, described in Section III, which identifies each participant as either the speaker or the listener at any given point in the dialog. We extracted 45 dialog sessions from the casual conversation dataset for use as the evaluation dataset, and the remaining data was divided into training and validation sets at a ratio of 9:1, respectively.

First, the features described in Section IV were extracted from the same dialog dataset used in the experiment, in order to obtain user motion, user speech, agent motion, and future agent motion. The future agent motion is the teacher data used during model training, while the other features are the data input to the model. Since the LSTM-Transformer generates head motion in real-time, the targeted agent head motion consists of feature sequences that are one frame ahead of the current agent motion feature sequence. This data is divided into segments with a duration of at least 6 seconds and up to 11 seconds, for use in model training and inference. After extracting these data segments from our dialog dataset, we obtained 21,536 segments for use as the training set.

During model training, agent motion features extracted from dialog dataset were input to the model, while during evaluation, agent head motion was generated autoregressively. Additionally, since contextual information about head motion prior to the inference starting point is necessary for generating agent head motion, the first 1 second of the feature sequence in each segment was used to warm up the model. During warm up, head motion was not generated, and during training, loss was not calculated.

We employed AdamW as the optimizer in our experiment, with a learning rate of 5×10^{-6} and a weight decay of 10^{-2}. Cosine Annealing was utilized for learning rate scheduling, and Huber Loss was used as the loss function. We set the number of epochs to 100 and the batch size to 32 during training.

Two models were used in our subjective evaluation experiments: a pre-trained model constructed as described above, and the same model fine-tuned with data from a specific listener, which was obtained from the dataset using the pre-trained model. Our motivation for using the second model was to test the hypothesis that by learning listener-specific idiosyncrasies in head motion, which vary from person to person, the model would be able to generate more accurate listener head motion. For this listener-specific fine-tuning, we obtained 3,792 data segments for model training from one, frequently appearing, dialog participant.

During our evaluation experiments, four human raters evaluated 20 listener head motions generated by each model and oracle motions (ground truth), respectively, resulting in 80 evaluation samples per model in total. The evaluation was performed by displaying video and audio of a human speaker, alongside video of the head motion of a human listener replicated by CG avatars, as well as the head motions of the CG avatars generated by each of the two models. Evaluators were asked to assess the naturalness and appropriateness of each listener, as well as signifying which listener's head motion they preferred. Since the proposed method does not generate body movement, the visualizations using CG avatars did not reflect head location information during the evaluation experiments, so only information regarding head angle was incorporated. The evaluations were conducted using the raters' mean opinion scores (MOS), as well as their overall preferences between the listener head motion responses of the human and those of head motions generated by each model. In the MOS evaluation, the naturalness of the head motion itself, and the appropriateness of the reaction to the speaker's behavior and speech, were evaluated using a 5-point scale (1 being the worst and 5 the best) for each of the two evaluation criteria. The average of the scores of the four evaluators was then computed to obtain the mean opinion score. In the overall preference evaluation, the evaluators select which of the listener head motions (human, pre-trained model, or pre-trained model with listener-adapted fine tuning) they thought was best.

VI. Result

Figure 3 shows 1.5 s of a human listener's head motion during casual conversation, and the listener head motion of an avatar generated using the proposed method without fine tuning. The gray waveform shown at the bottom of the figure is the speaker's corresponding speech waveform. While the human listener repeatedly performs an upward head motion (jerk) in response to the speaker's utterance, the head motion generated by the model results in a downward head motion (nod) during the time interval indicated by the blue box, which is slightly after the speech audio is input, confirming that the head motion was generated in response to the user's speech.

Table I shows the results of the MOS and overall preference evaluations by the four evaluators. First, in the MOS evaluation of naturalness and appropriateness, there was no

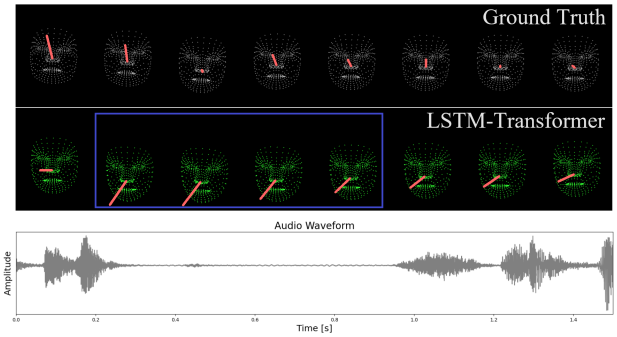

Fig. 3: Comparison of 1.5 s of human listening (top) and of motion generated using the proposed model without fine-tuning. The red lines represent the orientation of the faces. The blue box shows the time interval when avatar nodding occurred. Waveform at bottom represents speaker's voice.

TABLE I: Results of subjective evaluation using 5-point mean opinion scores and overall evaluator preference.

Model	***Naturalness***	***Appropriateness***	***Preference*** [%]
Ground Truth	3.53	3.90	70.00
Proposed w/o FT	3.23	2.74	11.25
Proposed w/ FT	3.10	2.88	18.75

significant difference between the evaluations of the proposed model and those of the proposed model fine-tuned using listener adaptation. In the naturalness evaluation, the ground truth (human listener) achieved the best results, but both of the proposed methods achieved similar scores. On the other hand, in the appropriateness evaluation there was a significant difference of more than 1 point between the ground truth and the proposed methods. In the preference evaluation, a large difference was observed between preference for the ground truth head movement (70 %) and for the movement generated using the proposed methods. These results suggest that while the proposed method can generate head movements with naturalness characteristics close to those of humans, there is still significant room for improvement in appropriateness in terms of timing, since the generated head movement was sometimes perceived as being inappropriate by humans.

We also investigated differences between the human and synthesized head motion. Figure 4 shows cross-correlations of the listener head motion generated by the human, the pre-trained model, and listener-adapted (fine-tuned) model, in relation to the speaker's corresponding head motion. We examined the temporal cross-correlation coefficient for the norm of head motion angular velocity to determine the extent of the speaker's influence on each of the listener's head motions. We investigated 25 data points, with each line in the graphs corresponding to one data point. For the human listener, no relationship was observed with respect to the speaker's head motion. On the other hand, in the head motion generated using the pre-trained and listener-adapted models, there is a pattern of positive correlation at around 0 s, confirming that head motion generated using the proposed methods is influenced by the speaker's head motion. It is likely that the large difference

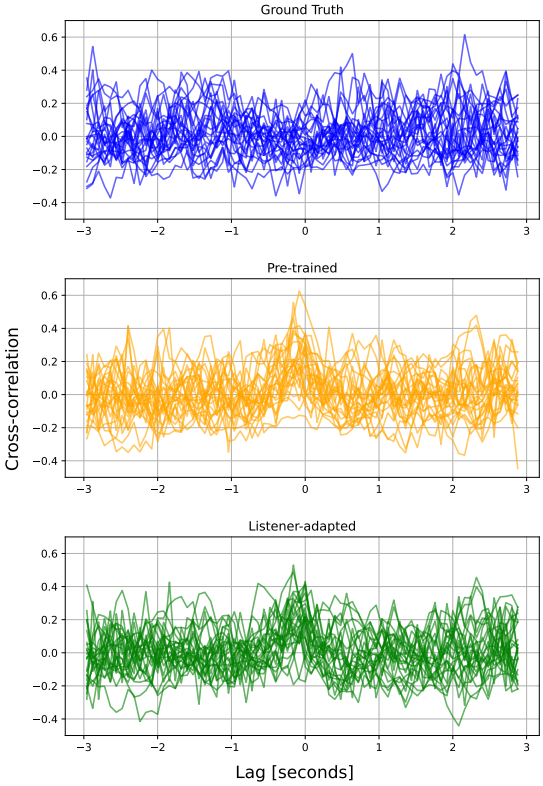

Fig. 4: Cross-correlation between listening heads and talking heads.

observed between the scores of the proposed methods and the ground truth in the appropriateness evaluation (Table I) was due to the generated head motion being needlessly influenced by the speaker's head movements, resulting in head motion that occurs at times that do not align with human expectations.

VII. Conclusion

In this study, we created a multimodal dataset of casual dialogs, consisting of one-to-one conversations between Japanese speakers. This data was then used to generate multimodal listener head motion based on both the speaker's speech and head motion during their utterances. We generated this simulated listener head motion using a proposed LSTM-Transformer head motion generation model based on the Transformer architecture. Our experimental evaluation of two variants of the proposed model demonstrated that the proposed method can generate human-like listener head motion, however modeling the timing of these head motions remains a challenge. Previous research [3] has shown that backchannels such as head movement occur at linguistic boundaries. In future research, we are considering improving the appropriateness of head movements timing by leveraging linguistic modalities through real-time speech recognition technologies. In our preference evaluation, the head movements generated using the listener-adapted (fine-tuned) version of the proposed model were preferred over those generated using regular model. In future research, we will

explore latent representations that can express individual head motion characteristics, for more controllable motion generation.

Finally, the type of subjective evaluation conducted in this study is costly in terms of time and money, which is a serious obstacle to research on head motion generation. Unfortunately, as Mittal et al. [22] have pointed out, low-cost evaluation metrics commonly used in motion generation, such as mean squared error (MSE), have low correlations with subjective human evaluations, and there are currently no other evaluation metrics which can replace human evaluation. However, in the field of speech synthesis, researchers have constructed deep learning models to predict human perceptual evaluations, and have used them to train speech synthesis models [23], thereby improving the perceived quality of synthesized speech. Therefore, as a future research goal, we are considering developing a method of automated evaluation that is capable of modeling human perception of the quality of synthesized head motion.

References

[1] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility head movement improves auditory speech perception," *Psychological science*, vol. 15, pp. 133–7, 03 2004.

[2] K. Otsuka and M. Tsumori, "Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks," *IEEE Access*, vol. 8, pp. 217 169–217 195, 2020.

[3] T. Koda, H. Kishi, T. Hamamoto, and Y. Suzuki, "Cultural study on speech duration and perception of virtual agent's nodding," pp. 404–411, 2012.

[4] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," *ACM SIGGRAPH*, vol. 2001, pp. 477–486, 08 2001.

[5] R. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," *Intelligent Virtual Agents*, pp. 25–36, 09 2005.

[6] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," pp. 68–79, 09 2011.

[7] C. Ding, P. Zhu, L. Xie, D. Jiang, and Z.-h. Fu, "Speech-driven head motion synthesis using neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2303–2307, 2014.

[8] C. Ding, P. Zhu, and L. Xie, "Blstm neural networks for speech driven head motion synthesis," 2015.

[9] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 44:1–44:20, 2023.

[10] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, jul 2017.

[11] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, vol. 128, 05 2020.

[12] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," pp. 1–10, 2020.

[13] J. Wang, Y. Zhao, L. Liu, T. Xu, Q. Li, and S. Li, "Emotional Talking Head Generation based on Memory-Sharing and Attention-Augmented Networks," pp. 2–6, 2023.

[14] D. Greenwood, S. Laycock, and I. Matthews, "Predicting Head Pose from Speech with a Conditional Variational Autoencoder," pp. 3991–3995, 2017.

[15] M. Zhou, Y. Bai, W. Zhang, T. Zhao, and T. Mei, "Responsive listening head generation: A benchmark dataset and baseline," 2021.

[16] X. Liu, Y. Guo, C. Zhen, T. Li, Y. Ao, and P. Yan, "Customlistener: Text-guided responsive interaction for user-friendly listening head generation," pp. 2415–2424, June 2024.

[17] M. Zhou, Y. Bai, W. Zhang, T. Zhao, and T. Mei, "Responsive listening head generation: A benchmark dataset and baseline," 2021.

[18] A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," pp. 1153–1162, 2021.

[19] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," pp. 10 809–10 819, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017.

[21] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," 2019.

[22] T. Mittal, Z. Aldeneh, M. Fedzechkina, A. Ranjan, and B.-J. Theobald, "Naturalistic head motion generation from speech," 2023. [Online]. Available: https://arxiv.org/abs/2210.14800

[23] Y. Choi, Y. Jung, Y. Suh, and H. Kim, "Learning to maximize speech quality directly using mos prediction for neural text-to-speech," *IEEE Access*, vol. 10, pp. 52 621–52 629, 2022.