Improving Listening Head Generation Performance Using Speech Representations from Self-Supervised Learning

Tamon Mikawa*, Yasuhisa Fujii†, Yukoh Wakabayashi*, Kengo Ohta‡, Ryota Nishimura* and Norihide Kitaoka*
* Toyohashi University of Technology, Japan
E-mail: {mikawa.tamon.fj, nishimura.ryota.tz, kitaoka}@tut.jp, wakayuko@cs.tut.ac.jp
† Google DeepMind, Japan
E-mail: yasuhisaf@google.com
‡ National Institute of Technology, Anan College, Japan
E-mail: kengo@anan-nct.ac.jp

***Abstract*—Appropriate, non-verbal listener behavior during conversation plays an important role in encouraging additional speaker utterances and achieving smooth turn-taking. In particular, head movements are valuable, non-verbal cues frequently used in response to the speaker's speech and movements. This study focuses on the Listening Head Generation (LHG) task, which generates listener head movements in automated dialogue systems from the speaker's voice and head movements. Through a performance comparison experiment, we verify the effectiveness of self-supervised learning (SSL) models, such as Wav2Vec 2.0 and HuBERT, for LHG tasks. While conventional LHG methods have employed traditional speech features such as log-Mel filter bank and Mel-frequency cepstral coefficients, this study confirms that utilizing latent speech representations acquired from SSL models improves objective evaluation metrics such as L2 error and the Pearson Correlation Coefficient. Our results provide important insights that will be useful for the realization of completely bidirectional multimodal dialogue systems.**

I. Introduction

Recent advances in artificial intelligence have improved the naturalness of automated dialogue systems. While these improvements are significant, achieving truly human-like communication will require a better understanding of natural human behavior during conversations and the ability to appropriately generate this linguistic and nonverbal information.

Head movements are among the nonverbal modalities frequently used in conversations and are important in facilitating smooth communication.

Many studies have investigated the relationship between head movements and speech features. Munhall et al. [1] reported that head and facial movements are strongly correlated with the amplitude and pitch of the speaker's voice, complementing the transmission of linguistic information. Graf et al. [2] conducted frequency band analysis of head movements and showed that high-frequency components are closely related to prosodic structure, with significant head movements synchronized with fundamental frequency, power, and phrase boundaries, suggesting a systematic relationship between speech structure and head movement.

Otsuka et al. [3] analyzed the communicative functions of head movements in listeners and speakers. Listener head movements were found to serve multiple functions simultaneously, providing complex feedback, while speaker head movements also serve various communicative purposes. Importantly, their study suggests that speaker head movements may influence listener head movements.

Head movement during dialog varies across cultures and languages, with frequent nodding observed during Japanese conversations. Koda et al. [4] showed that inappropriate avatar nodding patterns, such as the complete absence of nodding, or nodding following the conventions of different cultures, negatively affect the user experience, emphasizing the importance of culturally appropriate head movement generation. Against this background, we address Listening Head Generation (LHG) [5] for Japanese dialogue agents.

Previous LHG studies have mainly used English datasets, and have generated listener head movements from speaker voice and head movements using conventional speech features such as Mel-Frequency Cepstral Coefficients (MFCC) and log Mel-filterbank (FBank). However, recently proposed Self-Supervised Learning (SSL) models trained on large speech datasets have demonstrated superior performance in speech-related tasks [6]. SSL models automatically acquire general speech representations through task-independent learning. Several studies have indicated that these SSL models learn to extract latent representations containing acoustic features and higher-order prosodic and syntactic features. Lin et al. [6] reported that many SSL models outperform baselines in prosody-related tasks, effectively extracting prosodic information from speech signals. Fan et al. [7] successfully generated head movements synchronized with speech using Wav2Vec 2.0, demonstrating the effectiveness of SSL models for audio-driven head movement generation.

Based on these findings, we approach this research under the hypothesis that, in LHG tasks, speech representations from data-driven SSL models can automatically extract speech information that human-designed FBank features cannot capture,

979-8-3315-7206-8/25/$31.00

enabling more precise listener head movement generation. Therefore, this study aims to improve LHG task performance in Japanese dialogue by verifying the effectiveness of speech representations obtained from state-of-the-art SSL models. Specifically, we compare the performance of multiple SSL model-encoded speech representations with that of previously proposed methods using conventional speech features, through objective evaluation metrics. Through this verification, we demonstrate the superiority of SSL models in LHG tasks, and gain insights for improving multimodal dialogue systems.

II. Related Work

Early research in Talking Head Generation [8], [9], [10] and LHG [11], [12] predominantly focused on rule-based approaches. Maatman et al. [12] demonstrated that users felt "being listened to" when the agent displayed appropriate head movements, such as nodding while listening, confirming the importance of listener head movements in facilitating smooth dialogue.

With the advent of deep learning, it became possible to generate more natural head movement. Ding et al. [13], [14] proposed approaches for generating speaker head movements from speech, and reported that the Bidirectional Long Short-Term Memory (BLSTM) [15] model yielded superior results. They also found that incorporating pitch and speech power in addition to FBank features enhanced the generation of head movement, suggesting the importance of comprehensive speech information in audio-driven head movement generation. Greenwood et al. [16] reported that synthesizing multiple consecutive frames together during LHG resulted in more dynamic head movements, while Zhou et al. [5] established a baseline deep learning model using LSTM for the real-time generation of listener head movements and facial expressions from the speaker's voice and visual information.

Transformer models [17] employing attention mechanisms have been widely used recently in many sequence modeling tasks, however their computational complexity increases quadratically with input sequence length, making them less suitable for real-time generation tasks like LHG. Recent motion generation approaches commonly utilize LSTM-based models, which have linear computational complexity in relation to data length, making them appropriate for sequential processing. However, Ding et al. [14] noted limitations in model size expansion. Given this background, we propose a new deep learning model for the LHG task, incorporating insights from previous research [13], [16] while using a modeling method similar to the one proposed by Zhou et al. [5]. By comparing the use of conventional FBank features with speech representations encoded by SSL models, we analyze in detail how the utilization of comprehensive speech information improves head movement during LHG.

III. Proposed Method

A. LSTM-Transformer Model

The computational complexity of Transformer models, which have been widely adopted for sequence modeling tasks, increases quadratically with input sequence length, making them unsuitable for autoregressive generation tasks requiring real-time processing. Therefore, in this study we propose a model that retains the advantages of the Transformer architecture while improving computational efficiency. Based on insights from Yu et al. [18], we replaced the self-attention mechanism in the Transformer model with a unidirectional LSTM, and eliminated the positional encoding module, as shown in Figure 1. This approach achieves computational complexity that scales linearly with input sequence length while enabling the learning of long-term dependencies through the LSTM's state retention capability. We refer to this approach as the LSTM-Transformer model. In our experiments, we set the number of blocks N to 20, and included 1,280 hidden units in each layer.

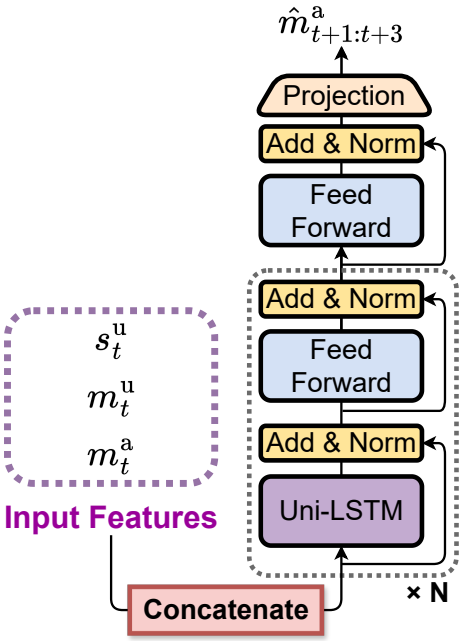

Fig. 1. Architecture of LSTM-Transformer model. We concatenate features of each modality along the feature dimension. Since the frame rate of the speech features is an integer multiple of head movement features, we combine multiple speech feature frames with one head movement frame to construct input for one generation step.

B. Task Modeling Method

In this study, our aim is to generate natural, real-time head movements for listening dialogue agents using multimodal information from the speaker's voice and head movements. The LSTM-Transformer model takes as input the user's speech features and head movements, as well as the agent's own head movements up to the current point in time, and autoregressively predicts the agent's head movements for the next point in time. Since the frame rate of speech features is an integer multiple of the frame rate of head movement features, for feature integration we combined multiple speech feature frames with one head movement frame to construct the input for a single generation step. Specifically, multiple speech feature frames

corresponding to the head movement frame at time t are concatenated and integrated as a multimodal input vector. This integration enables the utilization of detailed speech information within the corresponding time window for each step of head movement generation. Based on insights from Greenwood et al. [16], we adopted an approach that outputs three frames (0.24 seconds) of head movement features in a single inference during model training, enabling the model to learn dynamic changes in head movement sequences.

C. Head Movement and Speech Features

In this study, head movement features were represented as 18-dimensional vectors, comprising rotation angles ($pitch$, yaw, $roll$) and position coordinates (x, y), along with their first- and second-order dynamic features.

Regarding speech features, we compared the use of conventional FBank features with the use of latent representations obtained from Transformer-based SSL models (HuBERT and Wav2Vec 2.0 [19]) pretrained with Japanese speech. We applied masking to the self-attention modules of these SSL models to prevent access to future information, considering the autoregressive nature of the LHG task.

We examined two methods for utilizing latent representations; (a) concatenation along the feature dimension (HuBERT, Wav2Vec 2.0), which comprehensively utilizes various speech information from each layer, based on findings from previous research [6], and (b) weighted average (WA) calculation using learnable parameters (HuBERT w/ WA, Wav2Vec 2.0 w/ WA), which allows the LHG model to automatically learn the importance of each SSL model layer for efficient integration of speech information. Additionally, we evaluated HuBERT-NAR (non-autoregressive), which does not consider autoregressive behavior, to assess the impact of masking future information.

The speech features described in this section were input to the head movement generation model at integer multiples of the head movement feature frame rate. Using these speech features, we then constructed different versions of our proposed LSTM-Transformer model. For comparison with the proposed models, we also conducted experiments with a standard Transformer model using latent representations from HuBERT, which were processed using concatenation, i.e., method (a) described above.

D. Evaluation Metrics

We adopted multiple evaluation metrics in our evaluation experiments to assess and analyze the quality of the generated listener head movements from multiple perspectives. Based on previous research [20], we modified or added the following metrics:

- *Pearson Correlation Coefficient (PCC)*, which calculates the correlation between the velocities of generated head angles and ground truth (GT) head angles. This metric evaluates the appropriateness of head movement generation patterns by measuring the correspondence of velocity increase/decrease patterns.

- *F1, Precision (Pre.), Recall (Rec.), and Accuracy (Acc.)* were used to evaluate the correspondence of head movement timing between generated and GT head movements, such as nodding, using a head movement activity detector.

- *Head Movement Activity (Act.)*, which represents the proportion of head movement occurrences, such as nodding, in a head movement sequence.

Among the evaluation metrics used in previous research [20], FID, P-FID, and SI were also calculated from head angles of 25 frames (2 seconds), and L2 and Variation (Var.) were calculated from head angles of 125 frames (10 seconds). The head movement activity detector was adjusted by empirically setting thresholds for the vertical velocity of head position to detect movements such as bowing, nodding, and jerking.

IV. Experimental Details

In this section, we describe our experimental conditions in detail. Currently, there are no large, publicly available, multimodal, dialog datasets that include Japanese head movements and annotations. Therefore, we collected our own Japanese multimodal dialogue dataset for model training. This dataset consists of approximately 90 hours of one-on-one audiovisual conversations gathered via the Zoom web meeting platform, involving 24 participants (10 males and 14 females). The dataset includes head movement video and separately recorded audio of each participant. All dialogues were annotated with turn information to identify speaker and listener roles. Additionally, the dataset was split so that listeners do not overlap across the training, validation, and test sets. Note that this dataset is currently unpublished.

We extracted the features described in Section III-C from this dataset. As model inputs, we obtained three types of features at video frame time t:

- user (speaker) head movements m_t^{u},
- user speech s_t^{u}, and
- dialogue agent (listener) head movements m_t^{a}.

Future dialogue agent head movements, which serve as a ground truth at time t, are composed of head movement features $m_{t+1:t+3}^{\mathrm{a}}$, i.e., with a temporal span of 3 steps. This is to enable the listening head generation model to model dynamic changes in head movement by having each step of the head movement features span a constant temporal duration. Head movement features were extracted from dialogue videos by detecting facial landmarks using Mediapipe [21], from which head angles and positions were obtained. This data was divided into segments with a maximum length of 11 seconds and a minimum length of 6 seconds, resulting in 50,979 segments for training, 4,188 for validation, and 8,010 for testing.

During model training, we input features at each time step using Teacher Forcing [22], while head movements were generated autoregressively during evaluation. Furthermore, since contextual information about head movements prior to the inference starting point is necessary for head movement generation, we first input feature sequences of 1 second to the model,

TABLE I

EVALUATION RESULTS. BOLDFACE RESULTS DENOTE THE BEST VALUES ACHIEVED FOR EACH METRIC. OUR PRINCIPAL PROPOSED METHODS ARE HIGHLIGHTED IN GREY.

Method	L2↓	FID (10^2)↓	Var.	SI	P-FID (10^2)↓	PCC (10^{-2})↑	F1↑	Pre.↑	Rec.↑	Acc.↑	Act.
Ground Truth (GT)	-	-	15.87	1.97	-	-	-	-	-	-	0.45
Random	35.15	11.38	18.77	1.94	11.44	−2.21	**0.59**	0.44	**0.88**	0.57	0.64
Mirror	39.33	1.89	22.27	1.95	19.36	0.97	0.54	0.56	0.51	0.67	0.41
Delay Mirror	39.65	2.11	22.57	1.89	6.52	0.49	0.52	0.55	0.49	0.67	0.35
FBank	24.36	**0.52**	11.91	2.04	**0.57**	5.69	0.48	0.63	0.39	0.68	0.19
HuBERT	**21.20**	1.00	9.66	2.08	1.05	9.44	0.45	**0.75**	0.32	**0.70**	0.23
Wav2Vec 2.0	21.87	0.89	10.00	2.06	0.97	9.56	0.44	0.68	0.32	0.69	0.13
HuBERT w/ WA	23.40	0.63	11.58	2.01	0.69	7.57	0.45	0.66	0.35	0.68	0.26
Wav2Vec 2.0 w/ WA	22.99	0.60	11.66	2.06	0.66	8.49	0.45	0.63	0.35	0.69	0.33
HuBERT-NAR	21.49	0.98	10.02	2.08	1.05	**9.79**	0.45	0.70	0.34	0.69	0.10
HuBERT (Transformer)	27.67	0.56	16.38	2.01	0.64	7.79	0.47	0.64	0.37	0.68	0.23

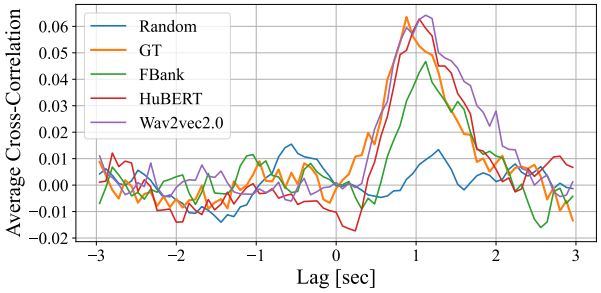

Fig. 2. Average cross-correlation between speaker and listener head motion velocity over 400 samples.

then learned and generated the subsequent head movements (m_t^{a}). Note that this initial 1-second period was excluded from the loss calculation.

We used AdamW [23] (learning rate: 5×10^{-6}, weight decay: 1×10^{-2}) as the optimizer for our experiments, with Cosine Annealing as the learning rate scheduler. We employed Huber Loss as the loss function and trained the models for 100 epochs with a batch size of 45.

We also compared the following baseline LHG methods with our proposed method:

- *Random*: Outputs randomly sampled head movement data from the training dataset.
- *Mirror*: Outputs the same head movements as the speaker.
- *Delay Mirror*: Outputs the speaker's head movements with an approximately 1.6-second delay, which is determined based on cross-correlation between the listener and speaker head movements in the training data.

V. RESULTS

Table I shows the evaluation results for 400 samples, each 10 seconds long, generated using each LHG method. The scores of HuBERT and HuBERT-NAR are almost the same, confirming that HuBERT functions stably even when using autoregressive approaches. The SSL models achieved the best L2 and PCC results, with an approximately 68% relative improvement in PCC compared to FBank. On the other hand, FBank outperformed the proposed methods in FID, suggesting that the distribution of head movement generated using the SSL models differs more from the ground truth. Variation in head movement using SSL models is smaller than FBank and increases the gap from GT, which reflects that SSL models tend to generate more static head movement. However, for precision against GT in head movement timing, HuBERT achieved an approximately 20% improvement over FBank. These results indicate that using the output of SSL models as speech features enhances synchronization with actual listener head movement timing compared to FBank. Additionally, the Random LHG model achieved the best results in F1 score and recall, which may be because the data used to train the Random model's head movement sampling contained speakers with relatively frequent head movements, which resulted in higher recall scores.

To evaluate the synchronization of head movement between speakers and listeners further, we conducted additional verifications. Figure 2 shows the average cross-correlation between speaker and listener angular head velocities across all 400 samples, where the horizontal axis represents the delay between speaker and listener head movement. The GT results show cross-correlation peaking at approximately a 1-second delay from the speaker's head movements, a pattern consistently observed across almost all speakers in our dataset. This reflects the natural response characteristics of listener head movement to speaker head movement. The results generated using the SSL models were closer to actual human correlation patterns than those generated using FBank, suggesting that our proposed approach achieves more human-like responsiveness to speaker head movements. Although Random achieved the best F1 score, the head movements it generated did not match the movement of human listeners.

Moreover, our proposed model achieved improvements in L2, PCC, and precision performance compared to the standard Transformer model, as well as improved computational efficiency as evidenced by the real-time factor (RTF) of 0.15 for the proposed model versus 0.22 for Transformer. Meanwhile, SSL models using weighted averaging (HuBERT

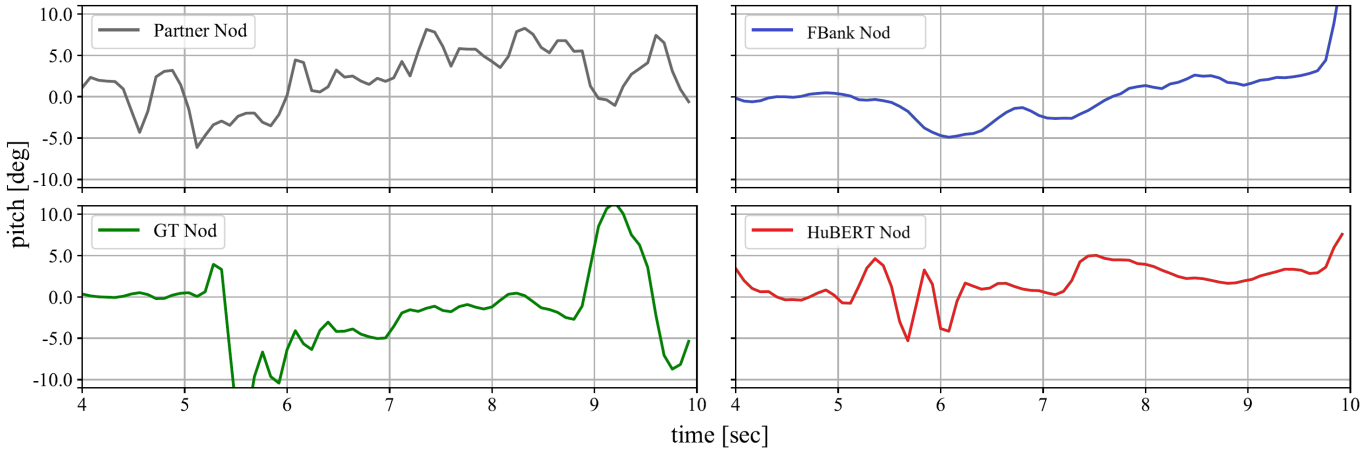

Fig. 3. Trajectory of head rotation in the nodding direction (x-axis). Partner Nod represents the speaker's head movement, while all others represent actual human or generated listener head movements. Deep, sharp valleys in the trajectory indicate nodding motions.

w/ WA, Wav2Vec 2.0 w/ WA) show results similar to FBank across all metrics. This suggests that the weighted averaging method cannot sufficiently utilize information that is useful for head movement generation timing. When observing the actual learned weights, we can see that although there is a slight bias towards the input and final layers, they are nearly uniform. This indicates that helpful information is not concentrated in specific layers of the SSL model but instead distributed across various layers.

Finally, Figure 3 shows examples of head nodding trajectories during a segment of dialogue, including samples of actual speaker and listener head movements, as well as head movements generated using FBank and HuBERT speech features. When using SSL models such as HuBERT and Wav2Vec 2.0, nodding tends to appear at a timing closer to actual head movements. In contrast, when using FBank, the generated head movement frequently occurred with different timing than human head movement. This may be why head movement generated using SSL models synchronizes more closely with actual listener head movements and exhibits response patterns more similar to speaker head movements, as shown in Figure 2 and Table I.

VI. Conclusion

In this study, we evaluated the effectiveness of SSL models for the LHG task, with the aim of enhancing the quality of generated head movement by utilizing latent speech representations from HuBERT and Wav2Vec 2.0 models instead of using conventional FBank features. The results of our evaluation showed that the SSL models outperformed traditional methods in terms of L2 and PCC performance, with HuBERT achieving an approximately 20% improvement in head movement timing correspondence compared to the FBank model. Cross-correlation analysis revealed that the SSL models more accurately reproduced the natural response patterns observed in human dialogues, however FID and Variation results showed that the SSL models tended to generate more restrained head movement.

Our proposed LSTM-Transformer LHG model improved RTF by approximately 32% compared to a standard Transformer model, showing its potential for real-time dialogue system applications. Additionally, although we used a Japanese dataset in this study, our proposed method is data-driven and can be applied to other languages such as English. Future work will focus on fine-tuning the SSL models to increase naturalness and enhance head movement diversity.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23H00493.

References

[1] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[2] H. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 396–401.

[3] 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.

[4] T. Koda, H. Kishi, T. Hamamoto, and Y. Suzuki, "Cultural study on speech duration and perception of virtual agent's nodding," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, ser. IVA'12, Santa Cruz, CA: Springer-Verlag, 2012, pp. 404–411, ISBN: 9783642331961.

[5] M. Zhou, Y. Bai, W. Zhang, T. Yao, T. Zhao, and T. Mei, "Responsive listening head generation: A benchmark dataset and baseline," in *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVIII*, Tel Aviv, Israel: Springer-Verlag, 2022, pp. 124–142, ISBN: 978-3-031-19838-0.

[6] G.-T. Lin et al., "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1104–1111.

[7] C. Cai et al., "Speak: Speech-driven pose and emotion-adjustable talking head generation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] 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.

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

[10] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–10.

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

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

[13] C. Ding, P. Zhu, L. Xie, D. Jiang, and Z.-h. Fu, "Speech-driven head motion synthesis using neural networks," 2014, pp. 2303–2307.

[14] C. Ding, P. Zhu, and L. Xie, "Blstm neural networks for speech driven head motion synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[15] H. Sak, A. Senior, and F. Beaufays, *Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition*, 2014. arXiv: 1402.1128 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1402.1128.

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

[17] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017.

[18] W. Yu et al., "Metaformer is actually what you need for vision," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 809–10 819.

[19] K. Sawada et al., "Release of pre-trained models for the Japanese language," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, May 2024, pp. 13 898–13 905. [Online]. Available: https://aclanthology.org/2024.lrec-main.1213.

[20] E. Ng et al., "Learning to listen: Modeling non-deterministic dyadic facial motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 395–20 405.

[21] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, *Real-time facial surface geometry from monocular video on mobile gpus*, 2019. arXiv: 1907.06724 [cs.CV].

[22] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. arXiv: 1711.05101 [cs.LG].