MULTIMODAL GENERATION OF UPPER-FACIAL AND HEAD GESTURES WITH A TRANSFORMER NETWORK USING SPEECH AND TEXT

Mireille Fares† Catherine Pelachaud* Nicolas Obin†

* ISIR, CNRS, Sorbonne Université, Paris, France
† STMS Lab - IRCAM, CNRS, Sorbonne Université - Paris, France

ABSTRACT

We propose a semantically-aware speech driven method to generate expressive and natural upper-facial and head motion for Embodied Conversational Agents (ECA). In this work, we tackle two key challenges: produce natural and continuous head motion and upper-facial gestures. We propose a model that generates gestures based on multimodal input features: the first modality is text, and the second one is speech prosody. Our model makes use of Transformers and Convolutions to map the multimodal features that correspond to an utterance to continuous eyebrows and head gestures. We conduct subjective and objective evaluations to validate our approach.

Index Terms— Co-Speech Gestures, Visual Prosody, Embodied Conversational Agents, Transformers

1. INTRODUCTION

1.1. Context

The human face is a key channel of communication in human-human interaction. During speech, humans spontaneously and continuously display various facial and head gestures that convey a large panel of information to the interlocutors. These gestures are known as “visual prosody” [1]: facial or head movement are produced in conjunction with verbal communication. During speech, the Fundamental Frequency (F0) variations and upper-facial movements are highly correlated [2]: they are the results of linguistic and conversational choices, and eyebrow motion can also occur during pauses. Another kinematic–acoustic relation that happens during production of speech is between F0 and head motion [2]. Natural and rhythmic head movements are one of the key factors in producing natural animations [3, 4]. To enable a smooth and engaging interaction with virtual agents, the agents’ verbal behavior must be produced in conjunction with appropriate non-verbal communication [5].

1.2. Related works and our contributions

A large number of head motion generation systems have been proposed in previous works [6, 7, 8, 9, 10, 11, 12]. A variety of generative statistical models aimed to predict the multimodal behavior of a virtual agent. Hidden Markov Models (HMM) [6], Recurrent Neural Networks (RNN) [12, 7], and Dynamic Bayesian Networks (DBN) [10, 11] have been used to generate head motion from speech: Generative Adversarial Networks (GAN) have been proposed to produce facial gestures from speech [13, 9], and [14] generates continuous 3D hand gestures based on acoustics and semantics. However, most of the aforementioned approaches exploit as input one modality only, namely speech, and neglect to render their approach using semantic information. Also, most of them focus only on facial expressions while the correlation between facial expressions and head movements are crucial to produce a natural behavior. For instance, [6, 7, 8, 11] do not generate eyebrow motion along with head motion, which are both correlated to F0 [2], and therefore are correlated to each other. On the other hand, transformer networks and attention mechanisms have been recently proved to be very efficient for sequence-to-sequence modelling, with particular advances for modelling multimodal processes. For instance, Transformers were previously used for translating speech to text (ASR) [15, 16], and multimodal learning of images based on text [17].

To overcome those limitations, we propose a novel approach for upper-facial and head gestures generation based on a multimodal transformer network. Our main contributions can be listed as follows: a transformer network is presented to handle the sequence-to-sequence modelling of multimodal upper-facial and head movements at the word-level. As inputs, the transformer exploits both acoustic and semantic information, by adding BERT semantic embeddings [18] to word-level encoded F0 contours. As outputs, modelling the correlation between head motion and upper-facial gestures allows the generation of a more coherent and natural behavior of the agent [1].

2. PROPOSED ARCHITECTURE

2.1. Multimodal Input/Output Features

The upper-facial movements and head rotations are represented by mean of Action Units (AUs) as defined in the Fa-
cial Action Coding Systems (FACS) and 3D head angles.

The work presented in this paper only considers the AUs that represent eyebrows and eyelids movements which are: inner raise eyebrow AU1, outer raise eyebrow AU2, frown AU4, upper lid raiser AU5, cheek raiser AU6, and lid tightener AU7.

Head rotations have three degrees of freedom, represented by the Euler angles roll, pitch, and yaw. They are represented by RX, RY, and RZ which are the rotation of the head with respect to the X, Y, and Z axes. In this paper, F0 values were extracted from the speech signal at a 5ms audio frame rate, linearly interpolated between unvoiced segments, and clipped to the range of 50 to 550Hz which represents the F0 range of human speech. Since F0, AU intensities (AU) and head rotations (R) are continuous, they were quantized to produce a finite set of discrete values to reduce the model size and energy consumption.

Finally, BERT word embeddings were extracted from the text transcription.

2.2. Face Gesture Generation Model Architecture

The proposed architecture aims at mapping multimodal speech and text feature sequence into continuous facial and head gestures. This problem is treated as a multimodal sequence-to-sequence problem, for which a transformer network operating at the word level is presented as illustrated in Figure 4. The inputs and outputs of the transformer network consist of one feature vector for each word W of the input text sequence. The input text corresponds to an Inter-Pausal Unit (IPU) which is a sequence of words separated by silent pauses longer than 0.2 seconds. F0 input sequence as well as AU intensities AU and head rotations R output sequences have a variable length. We set the maximum F0 input sequence length to 100, and the maximum AU/R output length to 124. Shorter sequences were padded to the maximum length, and longer ones were truncated. In order to handle continuous flow of input and output information with different timing, the transformer is wrapped with a F0 encoder module at its input and to a AU/R decoder at its output. The objective of the F0 encoder is to encode the continuous F0-values at the word level and the objective of the AU/R decoder is to reconstruct the continuous values for AU and R from a word-level encoding. Each spoken word W is represented by a word-level F0 embedding vector \( X_W^{(1)} \) which at its turn corresponds to the sequence of F0-values \( X_W^{(1)} = (F0_1, \ldots, F0_{N_W}) \), where \( N_W \) is the number of F0-values corresponding to the spoken word W; and \( X_W^{(2)} \) is the BERT word embedding vector corresponding to the spoken word W - including silences. Silences less than 0.2 secs may belong to IPUs. They do not have a contextual BERT embedding, hence, we replaced them by a comma - ",". Each spoken word W is also represented by a word-level AU vector \( Y_{W,AU}^{(j)} \) (resp. R vector \( Y_{W,R}^{(j)} \)) which in turn corresponds to the sequence of values \( Y_{W,AU}^{(j)} = (AU_{1}^{(j)}, \ldots, AU_{N_W}^{(j)}) \) (resp. \( Y_{R}^{(j)} = (R_{1}^{(j)}, \ldots, R_{N_W}^{(j)}) \)) where \( j \) denotes the \( j \)th AU/R and \( N_W \), the number AU/R-values corresponding to the word W.

**F0 Encoder:** As depicted in Fig.1, for each W, three one-dimensional convolutional layers are applied to project the input F0 sequence \( X_W^{(1)} \) into a word-level representation of F0 contours covering local context of F0 variations. These convolutional layers include 64 filters, with a kernel size equal to 3. The generated output vector of the latter layers \( X_W^{(1)} \) is then fed as an input to the Transformer Encoder.

**Transformer Encoder:** The transformer encoder architecture is depicted in Fig.2(a); it is similar to the one proposed in [22]. In our work, it is composed of a stack of \( N = 4 \) identical layers. Each layer has two sub-layers: the first one is a multi-head self attention mechanism with 4 attention heads, and the second one is a position-wise fully connected feed-forward network. As the original transformer encoder, we employ a residual connection around each of the 2 sub-layers, followed by layer normalization.
Cross-Modality Attention Module (CMAM): The output of the transformer encoder \( Z \), as well as \( X^{(2)}_W \) are fed as inputs to the Cross-Modality Attention Module (see Fig.1). This Module has the same structure as the Transformer Decoder in [22]. It generates a representation that can take into account both modalities text and speech. The representation learning is done in a master/slave manner, where one modality - the master - is used to highlight the extracted features in another modality - the slave. This module takes \( X^{(2)}_W \) as master, and \( Z \) as slave. Thus, it performs cross-attention such that the attention mask is derived from text modality, and is harnessed to leverage the latent features from the speech modality.

Transformer Decoder: The decoder is composed of \( N = 4 \) identical decoding layers, with \( 4 \) attention heads. Similar to the one proposed in [22], it is composed of residual connections applied around each of the sub-layers, followed by layer normalization. As depicted in Fig.2(b), the self-attention sub-layer in the decoder stack is modified to prevent positions from attending to subsequent positions. The output predictions which are offset by one position, and this masking ensures that the predictions for position index \( j \) depend only on the known outputs at positions less than \( j \). We use 9 Transformer Decoders, one for each \( AU/R \). For simplicity, Fig. 1 only illustrates one decoder.

AU/R decoder: The Transformer Decoder outputs are concatenated together, then fed to 3 one-dimensional convolutional layers to learn the correlation between the 9 output features, and therefore the correlation between facial and head movements. Finally, a Dense layer with a Softmax activation function is applied on each of the outputs, to convert the outputs to predicted next-token probabilities. The final output sequences are \( Y_{W,AU}^{(j)} \) and \( Y_{W,R}^{(j)} \).

Transformer Sub-Layers and Hyperparameters: The transformer encoder and decoders have attention sub-layers, and contain fully connected feed-forward network which are applied to each position separately and identically. Similarly to other sequence to sequence models, we use learned embeddings to convert the input tokens and output tokens to vectors of dimension \( d_{model} = 64 \). All sub-layers and embedding layers therefore use this dimension. The inner feed-forward layers are of dimension \( d_{ff} = 400 \). Positional encodings are applied to the inputs of the transformer encoder and decoders. They have the same dimension as the embeddings, so that they can be added together. We use sine and cosine functions, similar to [23].

3. EXPERIMENTS

3.1. Material and Experimental Setups

We trained it on a subset of the TED dataset collected in [23], containing preprocessed \( AU/R \), F0s, and BERT embeddings of filtered shots where speakers’ face and head are visible and close to the camera. Our subset consists of the features of 200 videos. Videos vary between 2 and 25 minutes, with a frame rate of 24 FPS, the total numbers of IPUs is 919, and of words is 62307. We shuffled all the IPUs, then split them into: training set (80%), validation set (10%) and test set (10%). The testing set is considered a Speaker Dependent (SD) set: it contains IPUs said by the speakers the model saw during training. We also created another small set of Speaker Independent (SI) from TED dataset, composed of data from 2 other speakers that are not in the 200 videos used for training. Each training batch contained 128 pairs of word embeddings, F0 sequences, and their corresponding AUs, RX, RY and RZ. The loss function used is the categorical cross-entropy between predicted and actual values. We used the Adam optimizer [24] with \( \beta_1 = 0.9, \beta_2 = 0.98 \) and \( \epsilon = 10^{-9} \). We used a learning rate scheduler as in [22], with warmup steps = 4000. We applied a \( dropout = 0.1 \) to the output of each sub-layer of the transformer, and to the sums of the positional encodings in the Transformer encoder and decoder stacks. All features values were normalized between 0 and 1.

3.2. Objective Evaluation

To assess the quality of the generated gestures, we used the following measures: Root Mean Squared Error (RMSE), Pearson Correlation Coefficient (PCC), Activity Hit Ratio (AHR) and Non-Activity Hit Ratio (NAHR). AHR and NAHR were proposed by Freeman et al. [25], to evaluate the performance of Voice Activity Detector (VAD) systems. We also consider them since evaluating AU activity looks similar to VAD evaluation. We considered an AU as “Activated” when its value is greater than 0.5, otherwise it is “Not-Activated”. AHR is the percentage of predicted AU activation with respect to ground truth. If it is greater than 100%, it means that the model is predicting more activation than the amount of activation that is in the ground truth. NAIHR is the same but for non-activity. We assessed the full model using SD test set. We additionally evaluated its capacity to generalize on SI set. To evaluate the different parts of our architecture, we conducted an ablation study as follows: (1) speech ablation, (2) text ablation, (3) CMAM ablation, and (4) AU/R decoder ablation. The ablation study was evaluated using RMSE metric.

3.3. Subjective Evaluation

To investigate human perception of the facial gestures produced by our model, we conducted an experimental study using the virtual agent Greta [26]. We added attention checks at the beginning of our perceptual evaluation, to filter out inattentive participants. The perceptual study was done by 30 participants fluent in English and with a University degree, recruited on Prolific, a crowd sourcing website. We present 12 videos in the study: each video shows the virtual agent saying a sequence of words that corresponds to a sequence of IPUs. 4 videos (condition M) use our full model of gestures predictions. 4 other videos (condition GT) were simulated
using the gestures extracted from TED videos, which serve as ground truth. The remaining 4 videos (condition E) were produced using predicted gesture animation of some IPUs with the sound of other IPUs. For each video, participants were asked to rate the naturalness, expressivity, coherence and human-likeness of the virtual agent’s gestures on a 0 to 5 likert scale, as recommended by [27]. The questions were listed in a random order. The agent’s lower facial gestures (mouth movements) were blurred to prevent participants from getting distracted by these gestures which were not inferred by our model, and therefore focus on the model’s generated gestures.

### 4. RESULTS AND DISCUSSION

Table 1 presents the results of the objective evaluation of the model’s predictions using SD test set. The 4 objective metrics were used for the different AU/R outputs. The RMSE errors reported in the table vary between AUs and R. They are close to zero for the 6 AUs, and between 0.15 and 0.3 for R. Hence, our model produces low errors for AUs and makes higher errors when predicting R. These errors could be explained as we trained our model on a database where speakers constantly look at their audience in an amphitheater and to the fact that head rotation may be less related to text and speech that facial gestures. PCC coefficients show that all SD predicted outputs correlate well with the GT, which indicates that the AU/R gestures can be well predicted though offset may happen, as illustrated in Fig. 4. AHR and NAHR were calculated to measure the activation of AUs only, since they are not applicable for R. Fig. 4 illustrates an example of predicted AU01 activation on SD test set. Fig.4 and AHR/NAHR values in Table 1 show that the model is able to predict well Activity/Non-Activity.

Ablation studies resulted in higher RMSE errors when performing Speech and Text Ablation for some AUs/R. For instance, AU01 RMSE error is 0.0819 for full model. It increased to 0.0836 after performing text ablation and to 0.09 when performing speech ablation. On the contrary, RZ RMSE score is 0.3036 for full model, it increased to 0.3089 with text ablation, and to 0.3111 after performing speech ablation. This shows that the model learns better R with speech than text. CMAM and AU/R Decoder ablation resulted in even higher RMSE errors especially for head rotations. This constitutes experimental validation that the use of both speech and text modalities together with a cross-attention efficiently improves the generation accuracy of face gestures and head rotations. As mentioned previously, we also tested our model on a SI set using the 4 objective measures. RMSE errors are between 0.048 and 0.095 for AUs, and between 0.19 and 0.22 for R. AHR and NAHR are 100%. This means that our model is capable of generalizing its predictions on SI data.

For our perceptive study, Fig. 3 shows the mean scores obtained on the 6 factors for the proposed model M, GT, and E. For the 6 factors our model M is perceived as much more closer to the GT than the E, especially in terms of naturalness and expressiveness. We conducted post-hoc t-test analyses on each factor for the three conditions. These revealed that there were no significant difference between our model M and the GT in terms of expressiveness (p-value=0.1578) while there is significant difference otherwise (p-value<0.001). This confirms our previous observation that the proposed model successfully approximates the GT with respect to its expressiveness, and to its naturalness though not significantly.

### 5. CONCLUSION

We present a new approach for modelling upper facial and head gestures. So far we have only compared our model with GT (replicated on a virtual agent). We will compare it with other models as soon as the data these models used are made available. We plan to improve R gestures, and expand our model to make it learn different speaker styles for gesture generation.
6. REFERENCES

[1] H. P. Graf, E. Cosatto, V. Strom, and Fu Jie Huang, “Visual prosody: facial movements accompanying speech,” in IEEE Int Conf on Automatic Face Gesture Recognition, 2002, pp. 396–401.

[2] H Yehia, T Kuratate, and E Vatikiotis-Bateson, “Linking facial animation, head motion and speech acoustics,” Journal of Phonetics, vol. 30, no. 3, pp. 555–568, 2002.

[3] Y Ding, C Pelachaud, and T Artieres, “Modeling multimodal behaviors from speech prosody,” in Int WS on Intelligent Virtual Agents, 2013, pp. 217–228.

[4] L Chen, G Cui, C Liu, Z Li, Z Kou, Y Xu, and C Xu, “Talking-head generation with rhythmic head motion,” in Eu Conference on Computer Vision, 2020, pp. 35–51.

[5] M Salem, K Rohlfing, S Kopp, and F Joublin, “A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction,” in RoMan. IEEE, 2011, pp. 247–252.

[6] G Hofer and H Shimodaira, “Automatic head motion prediction from speech data,” in Interspeech, 2007.

[7] K Haag and H Shimodaira, “Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis,” in Int. Conf on Intelligent Virtual Agents, 2016, pp. 198–207.

[8] J Lu and H Shimodaira, “Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder,” arXiv preprint arXiv:2002.01869, 2020.

[9] K Vougioukas, S Petridis, and M Pantic, “Realistic speech-driven facial animation with gans,” Int Journal of Computer Vision, pp. 1–16, 2019.

[10] S Mariooryad and C Busso, “Generating human-like behaviors using joint, speech-driven models for conversational agents,” IEEE Trans on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2329–2340, 2012.

[11] N Sadoughi and C Busso, “Speech-driven animation with meaningful behaviors,” Speech Communication, vol. 110, pp. 90–100, 2019.

[12] S Wang, L Li, Y Ding, C Fan, and X Yu, “Audio2head: Audio-driven one-shot talking-head generation with natural head motion,” preprint arXiv:2107.09293, 2021.

[13] T Karras, T Aila, S Laine, A Herva, and J Lehtinen, “Audio-driven facial animation by joint end-to-end learning of pose and emotion,” ACM Trans on Graphics (TOG), vol. 36, no. 4, pp. 1–12, 2017.

[14] T Kucherenko, P Jonell, S van Waveren, G E Henter, S Alexandersson, I Leite, and H Kjellström, “Gesticulator: A framework for semantically-aware speech-driven gesture generation,” in ACM Int Conf on Multimodal Interaction, 2020, pp. 242–250.

[15] O Hrinchuk, M Popova, and B Ginsburg, “Correction of automatic speech recognition with transformer sequence-to-sequence model,” in ICASSP. IEEE, 2020, pp. 7074–7078.

[16] A Mohamed, D Okhonko, and L Zettlemoyer, “Transformers with convolutional context for asr,” arXiv preprint arXiv:1904.11660, 2019.

[17] S Yao and X Wan, “Multimodal transformer for multimodal machine translation,” in Association for Computational Linguistics, 2020, pp. 4346–4350.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[19] P Ekman and EL Rosenberg, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.

[20] Y Guo, “A survey on methods and theories of quantized neural networks,” arXiv preprint:1808.04752, 2018.

[21] G Rehm and T Declerck, “Language technologies for the challenges of the digital age,” in GSCL, 2018.

[22] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin, “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.

[23] M Fares, “Towards multimodal human-like characteristics and expressive visual prosody in virtual agents,” in Int Con on Multimodal Interaction, 2020, pp. 743–747.

[24] D P Kingma and B J Adam, “A method for stochastic optimization. arxiv,” Learning, 2014.

[25] W Q Ong, A W C Tan, V V Vengadasalam, C H Tan, and T H Ooi, “Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive nonlinear filter,” Entropy, vol. 19, no. 11, 2017.

[26] C Pelachaud, “Greta: a conversing socio-emotional agent,” in ACM SIGCHI Int WS on Investigating Social Interactions with Artificial Agents, 2017, pp. 9–10.

[27] P Wolfert, N Robinson, and T Belpaeme, “A review of evaluation practices of gesture generation in embodied conversational agents,” arXiv preprint arXiv:2101.03769, 2021.