Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation
Mireille Fares, Catherine Pelachaud, Nicolas Obin

To cite this version:
Mireille Fares, Catherine Pelachaud, Nicolas Obin. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. 30th European Signal Processing Conference (EUSIPCO) 2022, Aug 2022, Belgrade, Serbia. hal-03675040

HAL Id: hal-03675040
https://hal.science/hal-03675040
Submitted on 3 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation

Mireille Fares
*ISIR, STMS
Sorbonne Université*

Catherine Pelachaud
*ISIR, Sorbonne Université*

Nicolas Obin
*IRCAM, Sorbonne Université, CNRS*

Abstract—We propose a semantically-aware speech driven model to generate expressive and natural upper-facial and head motion for Embodied Conversational Agents (ECA). In this work, we aim to produce natural and continuous head motion and upper-facial gestures synchronized with speech. We propose a model that generates these 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 and compare it with state of the art.

Index Terms—Semantically and Speech Driven Gestures, Transformers, Visual Prosody, Embodied Conversational Agents

I. INTRODUCTION

A. 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. 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].

B. Related works and our contributions

A large number of head motion generation systems has been proposed in previous works [5]–[11]. A variety of generative statistical models aimed to predict the multimodal behavior of a virtual agent. Hidden Markov Models (HMM) [5], Recurrent Neural Networks (RNN) [6], [11], and Dynamic Bayesian Networks (DBN) [9], [10] have been used to generate head motion from speech; Generative Adversarial Networks (GAN) have been proposed to produce facial gestures from speech [8], [12], [13] 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, [5]–[7], [10] 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) [14], [15], and multimodal learning of images based on text [16]. To overcome those limitations, we propose a novel approach for upper-facial and head gestures generation based on a multimodal transformer network. Our contributions can be listed as follows: 1) a transformer network operating on multi-modal input text and speech information in order to generate upper-facial and head movements, and 2) a cross-attention module that can efficiently exploit semantic and speech information. 1

The paper is organised as follows. The next section describes the proposed architecture of our model and its multimodal features. Then our objective and subjective evaluation experiments are presented including the LSTM-based Baseline Model. We finally discuss our results.

II. PROPOSED ARCHITECTURE

A. Multimodal Input/Output Features

The upper-facial movements and head rotations are represented by mean of Action Units (AUs) as defined in the Facial Action Coding System (FACS) [17] and 3D head angles. The work presented in this paper only considers the AUs that represent eyebrows 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, 1

1Video samples of our model’s gestures predictions and other related material can be found in: https://github.com/mireillefares/VAAnimation/blob/main/README.md
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 [18]. Finally, BERT word embeddings were extracted from the text transcription.

![Diagram](image)

**Fig. 1 (a): Model Architecture**

**Transformer Encoder Output**

- **Transformer Decoder Output**

**Fig. 1: Face Gesture Generation Model Architecture**

### B. 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 Fig.1 (a). 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) = (F01, . . . , 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(1) (resp. R vector Y_W,R(1)) which in turn corresponds to the sequence of values dY_W,AU(1) = (AU1, . . . , AU_N_AU) (resp. Y_W,R(1) = (R1, . . . , RN_R)) where j denotes the jth AU/R and N_W the number AU/R values corresponding to the word W.

**F0 Encoder:** As depicted in Fig.1 (a), 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.1 (b); it is similar to the one proposed in [19]. 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_W(2) are fed as inputs to the Cross-Modality Attention Module (see Fig.1 (a)). This Module has the same structure as the Transformer Decoder in [19]. 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 the other modality - the slave. This module takes X_W(2) - Text Modality - as master, and Z - Speech Modality - 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 [19], it is composed of residual connections applied around each of the sub-layers, followed by layer normalization. As depicted in Fig.1 (c), 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 ensure that the predictions for position index \( j \) depend only on the known outputs at positions less than \( j \). We use 6 Transformer Decoders, one for each AU/R. For simplicity, Fig.1 (a) only illustrates one decoder.

**AU/R decoder**: As depicted in Fig.1 (a), the Transformer Decoder outputs are concatenated together, then fed to 3 one-dimensional convolutional layers that include 64 filters, with a kernel size equal to 3, to learn the correlation between the 6 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 \( Y_{AU} \) and \( Y_{R} \).

**Transformer Sub-Layers and Hyperparameters**: the transformer encoder and decoders have attention sub-layers, and contain fully connected feed-forward networks 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 \( d_{\text{model}} = 64 \). 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 [19].

### III. Experiments

**A. Material and Experimental Setups**

We trained it on a subset of the TED dataset collected in [20], containing preprocessed AUs/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%). There are two test conditions: SD (Speaker Dependent) and SI (Speaker Independent). The SD condition aims to assess to what extent the model can generalize on new sentences pronounced by a speaker seen during training - training set included multiple speakers. The SI condition aims to assess the extent to which gestures predictions can be extrapolated to unseen speakers. 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 Adam optimizer with \( \beta_1 = 0.9, \beta_2 = 0.98 \) and \( \epsilon = 10^9 \). We used a learning rate scheduler as in [19], 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. The total number of the model’s parameters is 2051133.

**B. 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. [21], to evaluate the performance of Voice Activity Detector (VAD) systems. We also considered 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. NAHR is the same but for non-activity. We assessed the full model using the SD test set. We additionally evaluated its capacity to generalize on the 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.

**C. Comparing to LSTM-based Baseline Model**

We compared our approach to a sequence to sequence LSTM-based model [20] to predict upper-face movements based on speech and text. This LSTM model [20] was developed only for predicting eyebrows movements (AUs). We extended it to include head movements (R). This extended model consists of mapping sequences of F0 and the Bert embedding that correspond to a Word (W), to the sequences of (AUs/R) of that corresponding W. It employs 2 layers of Bidirectional LSTMs to encode the concatenation of word-level F0 contours and Bert embeddings. The hidden internal states are then transmitted to 6 decoders to produce the corresponding AUs and R for a given input. The decoders are followed by Dense Layers with Softmax Activation. The hyperparameters of the extended LSTM model are as follows: in both encoder and decoders, the first layer of bidirectional LSTM has 200 units, and the second one has 100 units. The activation function used in these layers is LeakyReLU with \( \text{alpha} = 0.01 \). We trained and tested this model using the same training, validation and test sets described in Section III - A. We trained it on 300 epochs, using a batch size \( = 128 \). We used Root Mean Squared Propagation (RMSProp) optimizer, and Categorical Cross Entropy Loss. We used the same metrics described in III.B to compare this SOTA Baseline model to our Transformer-based model.

**D. Subjective Evaluation**

To investigate human perception of the facial gestures produced by our model, we conducted two different experimental...
TABLE I: Objective Evaluation: comparison of proposed transformer model vs. SOTA baseline model

<table>
<thead>
<tr>
<th>AU01</th>
<th>AU02</th>
<th>AU04</th>
<th>AU05</th>
<th>AU06</th>
<th>AU07</th>
<th>RX</th>
<th>RY</th>
<th>RZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0229</td>
<td>0.0287</td>
<td>0.0372</td>
<td>0.0231</td>
<td>0.0601</td>
<td>0.10</td>
<td>0.16</td>
<td>0.24</td>
<td>0.30</td>
</tr>
<tr>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>98.0</td>
<td>100.0</td>
<td>99.0</td>
<td>100.0</td>
<td>99.3</td>
<td>98.0</td>
<td>10.0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>101.0</td>
<td>100.0</td>
<td>130.0</td>
<td>100.0</td>
<td>102.6</td>
<td>104.1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>0.20</td>
<td>-0.002</td>
<td>0.53</td>
<td>0.50</td>
<td>0.48</td>
<td>0.33</td>
<td>0.53</td>
<td>0.97</td>
<td>0.22</td>
</tr>
<tr>
<td>-0.012</td>
<td>34.15</td>
<td>60.31</td>
<td>21.12</td>
<td>33.11</td>
<td>20.21</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>42.56</td>
<td>107.08</td>
<td>121.85</td>
<td>135.95</td>
<td>135.05</td>
<td>131.52</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

The first study was done by 35 participants, and consisted of 8 videos: 4 were simulated with our SI model of IPUs. We considered 4 conditions: 4 videos (condition E) were simulated using the gestures extracted from TED videos, which serve as ground truth; 4 videos of the virtual agent were simulated using the LSTM-based Baseline model of SD predictions (condition SOTA). The remaining 4 videos were produced using predicted gesture animation of IPUs with the sound of other IPUs (condition M). The second study was conducted by 55 participants. The goal of the second study was to evaluate our model when simulated with SI data, and therefore its capability to generalise to new speakers. It included 8 videos: 4 were simulated with our model’s SI predictions, and 4 using SI gestures extracted from SI set (described in III.A), which serve as ground truth. For each video in both studies, participants were asked to rate the 5 factors, namely naturalness, coherence, human-likeness, synchronization, and alignment of the virtual agent’s gestures on a 1 to 7 likert scale [23]. The questions were listed in a random order. The agent’s 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.

IV. RESULTS AND DISCUSSION

Table I reports the model’s as well as the SOTA’s objective evaluation results using the same SD test set. Results reveal that RMSE errors are much smaller for M (0.0229 ≤ error ≤ 0.3) than SOTA (0.2046 ≤ error ≤ 0.9786). On the other hand, PCC coefficients show that M’s predictions (0.94 ≤ PCC ≤ 1.0) are more correlated than SOTA’s predictions (-0.0525 ≤ PCC ≤ 0.0179) to GT. AHR and NAHR were calculated to measure the activation of AUs only, since they are not applicable for R. Results show that M predicts better the activation rate AHR (AHR ≥ 98.0) than SOTA (20.211 ≤ AHR ≤ 60.314). The non-activation rate is higher for SOTA (107.084 ≤ NAHR ≤ 135.95) than for M(100 ≤ NAHR ≤ 130). This constitutes objective validation that M gives better results than SOTA in terms of error, correlation, and AUs’ activation rate. Ablation studies conducted on our model resulted in higher RMSE errors when performing Speech and Text Ablation for some AUs/R. For instance, AU01 RMSE error (0.0819 for full model) increased to 0.0836 with text ablation and to 0.09 with speech ablation. On the contrary, RZ RMSE score (0.5036 for full model) increased to 0.3089 with text ablation, and 0.3111 with speech ablation. A/U Decoder ablation resulted in even higher RMSE errors especially for head rotations. AUs and R RMSE scores increased after CMAM ablation (i.e. AU01 RMSE increased to 0.9091). This constitutes objective validation that the use of multi-modal inputs (speech and text modalities) in M improves predictions. Thus we can also conclude that CMAM module is an efficient and a key component of our model, as it improves the generation accuracy of face gestures and head rotations.

For our first perceptive study conducted on SD data, Fig. 2 (a) shows the mean scores obtained on the 5 factors for the 4 conditions: our model M, the baseline SOTA, the ground truth GT, and the error E. For the 5 factors, M is perceived as much closer to GT than SOTA and E, especially in terms of Alignment and Synchronization between speech and gestures. The mean difference between M and GT is 0.72 for Alignment and 0.74 for Synchronization (Fig.2 (a)). SOTA is perceived as the closest to E especially in terms of Coherence, and Alignment: the mean difference is 0.4 and 0.44 respectively (Fig.2 (a)). We also conducted ANOVA Between-Subjects to test whether there are significant differences between the 4 conditions along all the 5 factors. Results showed that for all factors, tests were significant (p<0.001). We additionally performed a post-hoc Fisher’s LSD Test to do pair-wise comparisons of the means between the factors of all conditions. Significant results (p<0.001) were found for
all the factors when comparing the condition GT with the 3 other conditions and when comparing our model M with SOTA and E. In particular, there were significant differences between M and SOTA in terms of the 5 factors \((p<0.007)\). This constitutes experimental validation that when used with SD data, condition M is perceived significantly better than SOTA and E for all the factors. The difference for the 5 factors between M and GT is not significant. For our second perceptive study conducted on SI data, Fig. 2 (b) shows the means scores obtained on all factors for conditions M and GT: for the 5 factors M is perceived as close to GT. As our data were not normally distributed (Shapiro test’s \(p<0.05\)), we conducted a post-hoc unpaired Wilcoxon test on each factor for the two conditions. For all tests, we could not find significant differences \((p>0.05)\). Thus our model when used with SI data and GT received similar values for the 5 factors.

V. CONCLUSION

We have presented a new approach for modelling upper facial and head gestures using Transformer model. We conducted objective and subjective evaluations that showed that our model produces animations that are close to the ground truth in term of expressivity while ensuring that speech and computed gestures are aligned and synchronized. In a next future, we plan to expand our model to learn different speaker styles for gesture generation.

REFERENCES