I-Brow: Hierarchical and Multimodal Transformer Model for Eyebrows Animation Synthesis

Mireille Fares, Catherine Pelachaud, Nicolas Obin

I-Brow: Hierarchical and Multimodal Transformer Model for Eyebrows Animation Synthesis *

Mireille Fares[1,2,4], Catherine Pelachaud[1,3,4], and Nicolas Obin[2,4]

[1] ISIR
[2] IRCAM-STMS
[3] CNRS
[4] Sorbonne University, Paris, France

Abstract. The human face is a key channel of communication in human-human interaction. When communicating, humans spontaneously and continuously display various facial gestures, which convey a large panel of information to the interlocutors. Likewise, appropriate and coherent co-speech facial gestures are essential to render human-like and smooth interactions with social agents. We propose "I-Brow", a model that produces expressive and natural upper facial gestures based on two modalities: text semantics and speech prosody. Our deep learning model is based on Transformers and convolutions. It has a hierarchical two-level encoding property: its input features are encoded, at both word and utterance levels, where an utterance corresponds to an Inter-Pausal Unit (IPU). We conduct subjective and objective evaluations to validate our approach.

Keywords: Eyebrows synthesis · Multimodality · Transformers.

1 Introduction

Nonverbal communication is the first form of communication in the lifespan of humans[20]. Before humans evolved their ability to speak and use language, they were able to communicate using their visual body gestures - their non-verbal channels of communication [20]. During speech, a variety of verbal, emotional, and conversational cues are displayed on the speaker's face. Facial gestures are consciously and unconsciously used to adjust speech, accentuate words, or mark speech pauses. [42]. Speakers render their communication expressive by blinking, moving their eyebrows and eyelids, frowning, and nose wrinkling [44].
During speech, Fundamental Frequency (F0) variations are highly correlated with eyebrow motion [5], which are the most relevant and common facial gestures employed during interactions [6]. Eyebrows can be utilized as a back-channel to

* This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

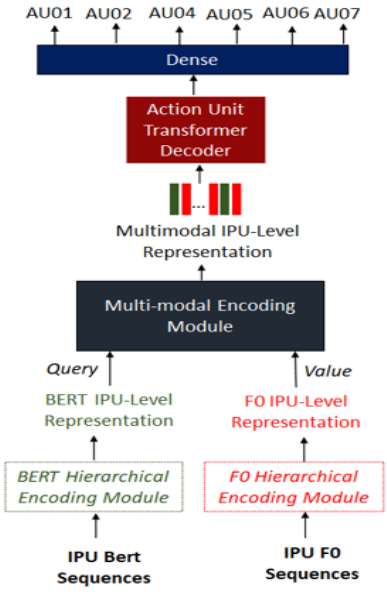

Fig. 1. "I-Brow" overall architecture. The network takes as input sequences of fundamental frequencies that correspond to one Inter-Pausal Unit (IPU) - sequence of continuous stretch of speech in one speaker's channel, delimited by a silence of more than 200ms -, and the corresponding text bert embeddings. With its hierarchical property, it encodes the input features at both word-level, and IPU-level, then generates a multimodal IPU-level representation of the input features. The network then learns to map the resulting representation to upper-facial - eyebrows - gestures.

signal the listener's level of understanding, agreement, or indicate listener's attitude towards what the speaker is saying [5]. Appropriate, expressive and human-like co-speech facial gestures are therefore an essential part of communication. To enable a smooth and engaging interaction with virtual agents, the agents' verbal behavior must be produced in conjunction with appropriate non-verbal communication [30]. In this paper we present "I-Brow" (Figure 1), a novel approach for upper facial gestures synthesis for Embodied Conversational Agents (ECA). Our model predicts expressive eyebrows and eyelids movements based on audio and text data. Upper facial movements are synthesized based on a hierarchically encoding: information at both, word-level and utterance-level - specifically Inter-Pausal Unit level - are encoded altogether. An Inter-Pausal Unit (IPU) is a continuous stretch of speech in one speaker's channel, delimited by a silence of more than 200ms, with a sequence of words that corresponds to what the speaker is pronouncing. The upper-facial gestures are predicted frame by frame. In contrast to previous works related to facial synthesis [40, 22, 18, 19, 26, 9, 37,

41, 31, 13, 4, 43, 17, 12, 29, 8], our work makes use of two modalities to allow for semantic-aware speech-driven continuous upper facial movements. Our contributions can be listed as follows: (1) acoustic and semantic features are mapped into continuous upper-facial gestures per inter-pausal unit, (2) word-level and IPU-level inputs features are encoded hierarchically, to extract important information from both word-level and IPU-level data.

2 Background and Related Work

The development of gesture synthesizing systems for virtual agents has received a lot of attention during the past years.

2.1 Gesture Generation Models

Hofer et al. [17] present a speech driven head motion sequence prediction system based on Hidden Markov Models. Haag et al. [16] propose a technique for speech driven head motion synthesis that uses deep neural networks with stacked bottleneck features, along with an LSTM network. Lu et al. [23] present an approach that predicts head motion based on speech waveforms. Ahuja et al. [1] study the links between spoken language and co-speech gestures. They propose "Adversarial Importance Sampled Learning" (AISLe) which combines adversarial learning with importance sampling. Sadoughi et al. [29] propose a speech-driven system to predict hand and head motion, using a Dynamic Bayesian Network. Their model is constrained by contextual information and these constraints condition the state configuration between speech and gestures. However, their model predicts movements based on only speech. Ferstl et al. [12] use generative adversarial training to map speech to 3D gesture motion. Moreover, Kucherenko et al. [21] propose a speech and text driven gesture generation that maps speech acoustic and semantic gestures into continuous 3D gestures. Yoon et al.[39] present an automatic gesture generation model that uses the multimodal context of speech text, audio, and speaker identity to reliably generate gestures. [14] propose an approach driven by speech to produce body gestures, however their approach uses models trained on single speakers.

2.2 Facial Gestures Synthesis Models

Cao et al. [4] produce expressive facial movement synchronized with the acoustic features of input utterances. Taylor et al. [33] synthesize lower facial movements based on a deep learning approach that employs a sliding window predictor that learns nonlinear mappings from phonemes to mouth motion. Zoric et al. [43] propose a facial gesture generation system for ECAs. Lip motion is generated based on input speech signal. In their work, virtual speakers can read given input text and transform it into the appropriate speech and facial movements. Mariooryad et al. [24] model a facial animation framework based on speech to generate head and eyebrows motion using Dynamic Bayesian Networks. Ding et

al. [8] propose an animation approach that uses HMM: their statistical model maps speech prosody with facial gestures. Song et al. [31] suggest an audio-driven approach based on conditional recurrent generation network, which merges image and audio features into a recurring unit and produce facial animation by time-dependent coupling. Vougioukas et al. [36] generate videos of talking heads based on a person's image, and audio data. They also produce lip movements, that are synchronized with speech, as well as facial expressions like blinks and eyebrow motion. Their approach is based on GAN with three discriminators. Their goal is to generate realistic expressions synchronised with speech. Suwajanakorn et al. [32] propose an approach based on Long Short-Term Memory (LSTM) for synthesizing a video of Obama's speech, they map original audio features to mouth shapes. The model could not perform well in generalizing other identities despite its good accuracy in lip synchronization. Tae-Hyun et al. [26] propose a speech-driven model trained through a large number of videos. Zhou et al. [41] synthesize random facial animation models by breaking the entanglement between audio and video. Chung et al. [18] present a speech-driven model, which integrates an auto-encoder to learn the correspondence between audio features and video data. Generated animation of their talking faces lack continuity. Duarte et al. propose an audio-driven method to synthesize facial videos [9], but the results are ambiguous. Garrido et al. [13] also propose a speech-driven approach that synthesize the speaker's face by moving the mouth shape of the speaker in the dubbing video to the target video. Karras et al. [19] propose a speech-driven real-time 3D facial animation model with low latency through the audio input. [7] also propose a novel approach for synthesizing speech-driven 3D facial animation.

The aforementioned works have focused on producing nonverbal behaviors (facial expression, head movement, gestures in particular) driven namely by speech. However, they have not considered both speech and text semantics for the production of the gestures. Only the work of Kucherenko et al. [21] is driven by speech and text, however they do not synthesize facial gestures.

3 Multimodal Data Features

We consider multiple modalities in our model. The following features were selected to be used for each modality:

Action Units features - Upper-facial gestures are represented by Action Units (AUs) that are predefined in the Facial Action Coding Systems (FACS)[10]. AUs that represent eyebrows and eyelids movements are AU1 (inner raise eyebrow), AU2 (outer raise eyebrow), AU4 (frown), AU5 (upper lid raiser), AU6 (cheek raiser), and AU7 (lid tightener). Action units are continuous values of intensities ranging from 0 (lowest) to 5 (highest). Continuous AU intensities were quantized to generate a finite range of discrete values. This step was applied to reduce the model size and energy consumption, as recommended by [15].

Audio features - The audio feature that we are considering in our model is the prosodic feature: the Fundamental Frequency F0. F0 values are continuous

values of frequency ranging from 85 to 180 Hz for the vocal speech of an adult male speakers. The values of an adult female speakers range from 165 to 255 Hz [3, 34]. F0 sequences were similarly quantized as AUs, to generate a finite range of discrete values. ***Text features*** -Text is a sequence of word. The dataset we have used included BERT embeddings for each word.

4 Training and Testing Dataset

TED (Technology, Entertainment, Design) conferences are conferences where speakers share their main research and expertise with their audience. Each speaker has a unique communicative style, a specific presentation topic, with a main goal to captivate his or her audience. We trained and tested our model on TED dataset [11], containing preprocessed AUs, F0s, and BERT embeddings of shots of 200 TED videos. These shots were filtered such as the speakers' face and head are visible and close to the camera. The average length if these videos is 13 minutes (minimum length is 1 min, and maximum length is 47 mins). The frame rate is 24 FPS, and the total number of IPUs is 266,000. We shuffled all the IPUs, then split them into: training set (80%), validation set (10%) and test set (10%).

The optimization algorithm that was used for training the model is Adam Optimizer with custom scheduling. The loss function used is the Sparse Categorical Crossentropy loss. Our test set is composed of *Speaker Dependent (SD)* data set, as well as *Speaker Independent (SI)* data set. The *Speaker Dependent (SD)* - the test set we have defined previously. It aims to evaluate to what degree the model can generalize on new IPUs said by the multiple speakers the model has seen during training. On the other hand, *Speaker Independent (SI)* include IPUs said by unseen speakers. It aims to evaluate the degree to which gestures predictions can generalize on unseen speakers.

5 "I-Brow" Model for Speech-driven Upper Facial Gesture Generation

This section describes our proposed approach for generating upper-facial gestures from two modalities: speech acoustics and semantics. We have applied a hierarchical encoding of the input features, such that the encoding encompasses both word-level encoding as well as Inter-Pausal Unit level encoding. The overall architecture of "**I-Brow**" is depicted in Figure 1.

To build an optimized IPU-level architecture, we started by implementing an architecture that includes only word-level features. Then, we used the word-level model as a baseline to implement our IPU-level architecture. To decide on the number of tokens to consider in IPUs, we generated the distribution of words in our dataset. We observed that approximately $29,000$ IPUs contain 10 tokens which include words and pauses as well. Thus, we decided to render the size of all IPUs of our training, validation and test sets equal to 10 words. Larger IPUs were truncated, the smaller ones were padded.

The following sections describe the word-level architecture, as well as the different components of the IPU-level architecture.

5.1 Word Level Model Architecture

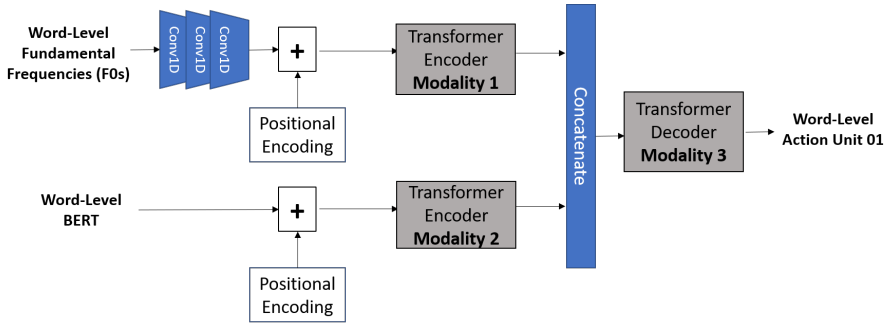

Fig. 2. word-level Network Architecture

As depicted in figure 2, the word-level architecture takes as input a sequence of Fundamental Frequencies (F0s) that corresponds to a word unit **W**, as well as the corresponding BERT embedding of the same **W**. The sequences of F0s are first passed through three one dimensional convolutional layers, to produce a representation of F0 contours. These layers include 64 filters, with a kernel size equal to 3. Positional encoding is then applied on the resulting vector, which is then given as input to a Transformer encoder. The Transformer Encoder has 4 encoding layers, with 4 attention heads. It has the same architecture as the one that was first proposed in [35].

On the other hand, positional encoding is applied on BERT embeddings, and the result is fed to another Transformer Encoder that has the same parameters as the previously described one. The outputs of both Transformer Encoders are concatenated, and then fed to a Transformer Decoder, which produces the corresponding word-level action units.

The Transformer Decoder has 4 decoding layers, with 4 attention heads, and has the same architecture as the one in [35]. For simplicity, figure 2 only illustrates the whole word-level architecture that predicts one Action Unit. All hyper-parameters of this architecture were chosen empirically.

5.2 IPU-level Model Architecture - "I-Brow"

The IPU-level architecture "**I-Brow**", illustrated in Figure 1, takes as input BERT and F0 features that correspond to the 10 words of the IPUs. BERT word embeddings are hierarchically encoded on both word-level and IPU-level.

In the same manner, we hierarchically encode F0s on a word-level and IPU-level. BERT and F0 IPU-level representations are then fed to our Multi-Modal Encoding Module, which produces one representation that encompasses both modalities. This final IPU representation is then fed to the Decoder Module which produces the 6 Action Units: $AU01$, $AU02$, $AU4$, $AU05$, $AU06$ and $AU07$.

Model Configuration	
4*T(1) to T(100)	Sequence of Token Representations produced by word level F0 Transformer Encoder
3*P(1) to P(100)	Positional Encoding of each token generated by the 3 layers of CONV1D
3*CONV1D	1 Dimentional Convolutional Layer
3*T(1) to T(768)	Sequence of Token Representations produced by word level BERT Transformer Encoder

Table 1. Details of the model configuration

The following sections describe each module of the IPU-level architecture. Note that the different acronyms used in the modules are summarized in Table 1.

5.2.1 Encoding Modules

5.2.1.1 F0 Hierarchical Encoding: The first module of our architecture is the F0 Hierarchical Encoding Module. This module produces one final F0 representation for the whole IPU by encoding the fundamental frequencies at the word level as well as at the IPU-level. Figure 3 illustrates the F0 hierarchical encoding architecture with 3 input words, for simplicity. Each word has a corresponding sequence of Fundamental Frequencies. The maximum length of all sequences of F0s is equal to 100 timesteps. First, F0 sequences are passed through 3 one dimensional convolution layers to extract the important features. Positional encoding is then applied on the result, and the output is fed to a Transformer Encoder which produces a sequence of F0 token representation for each word. This Transformer Encoder has the same architecture as the one in the original Transformer encoder [35]. Our encoder contains 4 encoding layers. Afterwards, we add a layer of self-attention which takes as input all F0 token representations that correspond to all words in the IPU. The output of this self attention layer is the final F0 representation for the whole IPU.

5.2.1.2 BERT Hierarchical Encoding: The second module of our architecture is the BERT Hierarchical Encoding Module, and it is illustrated in Figure 4. This module produces one BERT representation for the 10 words of the IPU. The inputs word-level BERT embeddings are initially represented by a 768 vector each. Afterwards, we add a layer of self-attention which takes as input all BERT

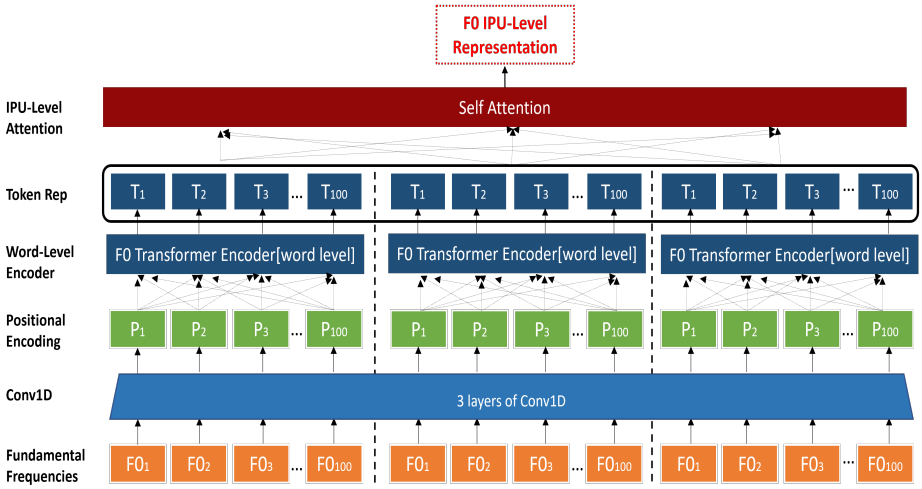

Fig. 3. F0 Hierarchical Encoding Module

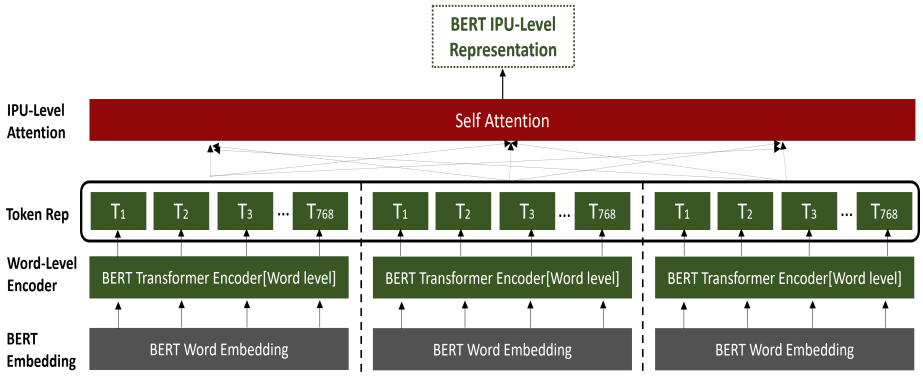

Fig. 4. BERT Hierarchical Encoding Module

token representations that correspond to all words in the IPU. The output of this self attention layer is the final BERT representation for the whole IPU.

5.2.1.3 Multi-Modal Encoding Module: The third module of our architecture is the Multi-Modal Encoding module, which is depicted in Figure 5: this module takes as input the two IPU-level representations of BERT and F0s, and produces one final representation that encompasses both features. First, both BERT and F0 embeddings are passed to a layer of additive (Bahdanau) attention [2]: we consider the BERT embedding to be the query, and F0 embedding to be the value. The output of the Query-Value attention is then passed to a 1D Global Average Pooling layer. On the other hand, we also add a 1D Global Average

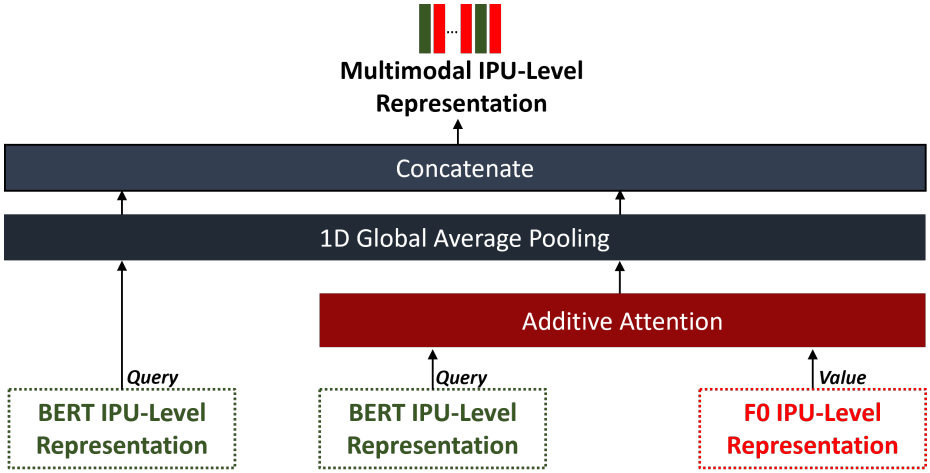

Fig. 5. Multi-modal Encoding Module

Pooling layer to the initial IPU-level BERT embedding (the query). The final multimodal IPU-level representation is the concatenation of the results.

5.2.2 Decoding Module. The Decoding Module is depicted in Figure 1. It takes as input the output of the Multi-Modal Encoding Module, which is the multi-modal IPU-level representation. This representation is fed to 6 different Action Unit Transformer decoders, which are the same ones used in the word-level architecture. The outputs of the decoders are then passed to a Dense layer, which in turn generates the continuous values of the 6 Action Units for all the words of an IPU. For simplicity, Figure 1 only illustrates a Transformer Decoder for one AU.

5.3 Training and Testing Procedures

We trained and tested our model using the TED dataset. We split the data into 3 sets: training set, validation set and test set. The training set is composed of 80% of the IPUs from the dataset. The remaining IPUs were then split into validation set (10%) and testing set (10%). The optimization algorithm that was used for training is Adam Optimizer, with custom scheduling. After data quantization, we constructed two dictionaries of discrete values corresponding to AU and F0.The loss function used is the Sparse Categorical Crossentropy loss. Our test set is composed of *Speaker Dependent* data, as well as *Speaker Independent* data: the *Speaker Dependent* data include the IPUs said by speakers that the model has seen during training. On the other hand, *Speaker Independent* data correspond to the IPUs said by speakers that the model did not see during training.

6 Evaluation Measures

In this section we describe the objective and subjective measures we used in our experiments.

6.1 Objective Measures

The objective metrics used to evaluate the produced animation are (1) Root Mean Squared Error (RMSE), and (2) Pearson Correlation Coefficient (PCC).

We also assess the Action Units Activity. Since the problem of evaluating Action Units Activity is very similar to Voice Activity Detection (VAD) evaluation problem, we used some metrics that are commonly used in VAD. These metrics were proposed by Freeman et al. [27], and they are widely used to evaluate the performance of a Voice Activity Detector. We considered an Action unit as "Activated" whenever its value is greater than a threshold 0.5. Otherwise, we considered it as "Non-Activated".

The Action Units Activity Detection metrics that we considered in our objective evaluation are defined thereafter:

- `Activation Hit Rate (AHR)`: percentage of predicted AU activation with respect to ground truth. If AHR (%) is greater than 100%, it means that the model is predicting more activation than the amount of activation that is in the ground truth. Otherwise, it means that there are less activation in the prediction than in ground truth.
- `Non-Activation Hit Rate (NHR)`: percentage of predicted non-activity with respect to ground truth. If NHR (%) is greater than 100%, it means that the model is predicting more non-activation than the amount of non-activation that is in the ground truth. Otherwise, it means that there are less non-activation in the prediction than in ground truth.

6.2 Subjective Measures

To investigate human perception of the facial gestures produced by our model, we conduct an experimental study. We make use of Prolific [28], a crowd sourcing website.

6.2.1 Experimental Design. We assess the *naturalness*, *expressivity*, *coherence* and *human-likeness* of the virtual agent's generated upper facial gestures. We base our study on the recommendations proposed in [38], by adapting them to facial gesture generation instead of hand gesture generation. More specifically, we asked the following questions: (1) In which video the agent's eyebrows align most with what it is saying ? (2) In which video the agent's eyebrows movements look more natural ? (3) In which video the agent's eyebrows movements are more appropriate ? (4) In which video the agent's eyebrows movements are more expressive ? (5) In which video the agent's eyebrows movements are more synchronized with speech ?

The questions were listed in a random order for each pair of videos. The agent's lower facial gestures were hidden as shown in Figure 6, to prevent the participants from getting distracted by these gestures.

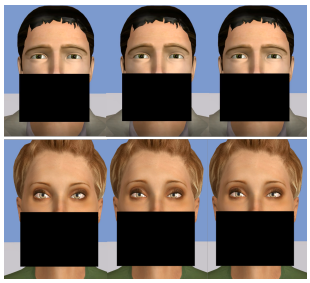

Fig. 6. The lower facial gestures of the Virtual Agents were hidden to prevent participants from getting distracted when evaluating the upper facial gestures

6.2.2 Attention Checks. We add attention check at the beginning of our perceptual evaluation, to filter out inattentive participants. These attention checks include 4 heavily distorted videos (audio and video quality). Participants are asked to report the videos where they experience sound/videos problems. The participants that do not report all 4 videos are excluded automatically from the study.

6.2.3 Experimental Procedure. The perceptual study is done by 30 participants, recruited on Prolific [28]. One requirement to be able to participate to the study is that participants must be fluent in English and above 18 years old. The study is composed of two parts:

1. In the first part, we present 5 sets of pairs of videos. Each pair is composed of two videos of the virtual agent saying a sequence of words that corresponds to one Inter-Pausal Unit. One video uses the *Speaker Dependent* AUs that are produced by our model, and the other one uses the AUs extracted from TED videos which serve as ground truth. For each pair of videos, participants are asked to answer the 6 questions listed in section 6.2.1.
2. The second part of the study is a comparative study in which we present another 5 sets of 3 videos of the agent saying a sequence of words that corresponds to a given inter-pausal unit. The first video uses the AUs from ground truth and is used as a comparison baseline for the participants. They are asked to compare the baseline video with two other videos: one that uses the AUs predicted by our model, and another one that uses the AUs of another IPU. The goal is for the participants to sort the 3 videos by selecting the video that resembles most the ground truth data.

7 Objective Evaluation Results

We present in this section the evaluation we perform on the full architecture and two baselines.

7.1 "I-Brow" Model

We first conduct the experiments on our "**I-Brow**" model driven by both modalities text and speech. The generated metrics which are illustrated in Table 2, reflect the performance of our model with respect to the continuous upper facial gestures data, as well as the Action Unit Activity Detection. The Speaker Dependent RMSE and PCC scores for the different Action Units indicate that the error rate between ground truth and predictions is low, and that AUs 1, 2, 4, and 6 are correlated with the ground truth. The Speaker Dependent Action Unit Activity Detection metrics reflect that the model is capable of detecting the activation of AUs for at least 50% of the time for AU01, AU02, AU04, and AU07. The percentages of predicted non-activity are higher than 100% which means that the model predicted more non-activation than the amount of non-activation in the ground truth.

	I-Brow Model (SD)				**S-Brow Baseline Model (SD)**				**T-Brow Baseline Model (SD)**				**I-Brow Model (SI)**			
	RMSE	**PCC**	**AHR**	**NAHR**	**RMSE**	**PCC**	**AHR**	**NAHR**	**RMSE**	**PCC**	**AHR**	**NAHR**	**RMSE**	**PCC**	**AHR**	**NAHR**
AU01	0.491	0.150	61.538	108.280	0.749	0.002	0.120	134.849	0.766	-0.130	8.062	141.326	0.759	0.150	105.847	97.481
AU02	0.220	0.129	51.632	104.234	0.422	0.000	0.032	108.826	0.512	-0.097	16.279	113.317	0.251	0.100	40.234	103.272
AU04	1.160	0.015	50.666	125.517	1.270	-0.924	0.010	201.467	0.806	0.056	26.524	139.379	0.752	0.008	138.840	87.413
AU05	0.158	-0.058	0.000	100.000	0.372	0.029	0.000	103.053	0.328	-0.023	0.000	105.736	0.097	0.045	0.000	100.000
AU06	0.284	0.112	0.000	111.000	0.637	0.010	0.000	137.541	0.632	-0.115	6.803	117.276	0.727	0.009	0.000	164.630
AU07	0.684	-0.195	56.716	118.954	1.307	-0.994	0.000	223.804	1.077	-0.109	10.860	159.822	0.744	-0.034	70.990	119.880

Table 2. Objective Evaluation Results. Objective evaluation results of objective metrics for (1) I-Brow model tested with SD set, (2) S-Brow baseline model tested with SD set, (3) T-Brow baseline model tested with SD set, and (4) I-Brow model tested with SI set

Metrics were generated for speakers that the model has seen during training (Speaker Dependent), as well as for other speakers that the model did not see during training (Speaker Independent). Speaker Independent results show that the model is capable of generalizing predictions for speakers not seen during training. Figure 7 depicts examples of Speaker Dependent predictions of AUs 1, 4 and 7 over one IPU.

7.2 Baseline Models

In this section, we evaluate the importance of each input modality to our model by individually ablating them, and ending up with 2 different variants of the model.

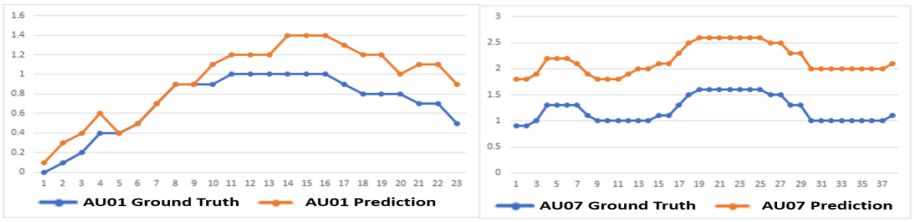

Fig. 7. Full Architecture Speaker Dependent Predictions

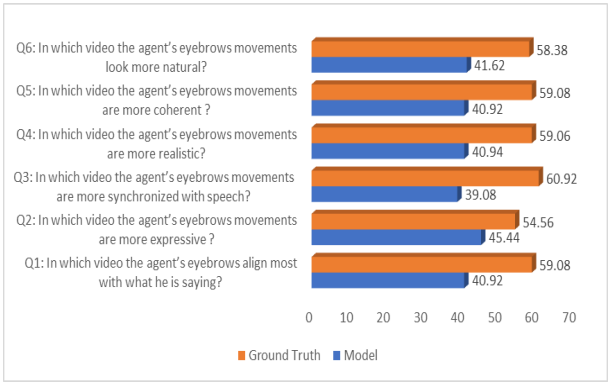

Fig. 8. Results of the first part of the perceptual study

7.2.1 "S-Brow" Baseline. The first baseline model "**S-Brow**" is a variant of **I-brow** but with an ablation of the text modality. We repeated the same experiments performed on "**I-Brow**", but this time with only speaker dependent data. Results are shown in Table 2. We can observe that RMSE errors got higher scores while PCC scores are lower. AU Activity Detection scores show that the model is less capable of detecting AU Activity.

7.2.2 "T-Brow" Baseline. The second baseline model "**T-Brow**" is also a variant of "**I-Brow**" but with an ablation of the speech modality. Same experiments were done after removing the speech modality while keeping the text modality. Results are shown in Table 2. We can notice the worsened performance of the model over the different measures.

8 Perceptual Evaluation Results

We conducted our perceptual study to investigate human perception of the upper facial gestures that are produced by our model. The six evaluation measures that we consider are: agent's *naturalness*, *human-likeness*, *expressiveness*, the *synchronization* of its gestures with speech, as well as the *alignment* of what

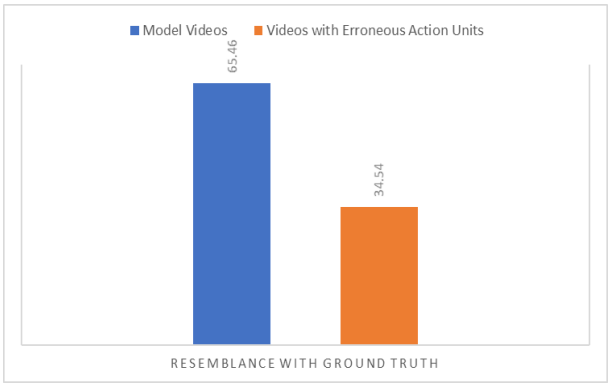

Fig. 9. Results of the second part of the perceptual study

it is saying with respect to its facial gestures (see Section 6.2.1). Note that 6 participants out of 30 did not pass the attention checks, thus we did not consider their participation in our study. Prolific suggested 6 other participants to participate in the experiments and they had successfully passed the attention checks. The results of the first part of the perceptual evaluation are presented in figure 8. This part of the evaluation aims to compare the AUs produced by our model, to the AUs of the ground truth. Results show that there was a preference for the ground truth motion facial gestures; however our model did still quite well: around 40% of the participants chose our model videos over the ground truth videos, when answering each question. This means that our model's Action Units values are close to the ground truth data. 41.62% of the participants find that the eyebrow movements look more natural in our model's videos than in the ground truth videos. Around 41% of the participants find that the agent's eyebrows movements are more coherent and more realistic in the videos produced by our model. Around 41% of the participants found that the agent's eyebrows align more with what it is saying in the videos produced by our model than the ones of ground truth. 39% of participants found that the AUs produced by our model are more synchronized than the ground truth Action Units. The results of the second part of the study are illustrated in figure 9. This part aims to compare our model's produced AUs with respect to ground truth, as well as with erroneous AUs (AUs of other IPUs). Results show that participants had a preference towards our model: 65% of participants showed that our model videos resemble more the ground truth videos than the videos where we added the Action Units of other IPUs.

9 Discussion

Objective Evaluation results show that our model is capable of generating speaker dependent upper facial gestures, and is able to predict AUs Activity/Non-Activity.

Our model performs better with the two inputs modalities by comparing its performance against the one modality (speech or text) architectures. Speaker Independent results inform us that the model is able to generalize predictions for speakers that it did not see during training. However, results could be improved by training the model on a larger number of TED videos. Subjective Evaluation results show that the videos that were simulated using our prediction model have similar qualities as the ground truth videos simulated on a Virtual Agent. Indeed, 40% of the participants reported that its eyebrows movement looks more natural, coherent, realistic, and aligned with what it is saying compared to the eyebrows movement in the ground truth videos. It means that our model videos look rather similar and resemble the Ground Truth. This conclusion was also validated by the second part of the subjective evaluation.

10 Conclusions and Future Work

This paper explored the use of hierarchy in convolutional and Transformer-based models for upper facial gesture synthesis. We started by proposing a word-level upper facial gesture synthesis model that predicts upper-facial gestures given text and speech inputs with a word-level segmentation. Then, we proposed an architecture that encodes the multimodal inputs and predicts the upper facial gestures for one inter-pausal unit. This architecture hierarchically encodes the input modalities - semantics and speech - at the word-level and IPU-level. The two encodings are then combined through the Multi-Modal Encoding Module which generates a multimodal IPU/utterance representation. This representation is then sent to six Transformer decoders, which in turn predict the six Action Units that correspond to the upper facial muscles. To the best of our knowledge, this is the first data-driven model that generates upper facial gestures based on the speech and text modalities while taking into account word-level and IPU-level features. It is also the first approach that employs Transformers with CNNs for this task, and processes sequences as whole rather than token by token. Through an objective and subjective studies, we further found that a multimodal encoding combining semantic and acoustic features is efficient for upper-facial gesture generation tasks. This paper shows the usefulness of the basic Transformer architecture for upper facial gesture generation. For future work, it would be beneficial to explore the effectiveness of the proposed model when applied to other behavior generation tasks such as head movements. Future work also involves testing the model if it is capable to reproduce behavior expressivity. In this case, we could consider adding the following voice quality features which are useful for expressiveness as the work in [25] suggests: Jitter, Shimmer, Harmonic to Noise Ratio (HNR), and Hammarberg index (Hamml). We also plan to train our model on a bigger training set, to render the model able to better generalize predictions for Speaker Independent data.

References

1. Ahuja, C., Lee, D.W., Ishii, R., Morency, L.P.: No gestures left behind: Learning relationships between spoken language and freeform gestures. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 1884–1895 (2020)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Baken, R.J., Orlikoff, R.F.: Clinical measurement of speech and voice. Cengage Learning (2000)
4. Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Transactions on Graphics (TOG) **24**(4), 1283–1302 (2005)
5. Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R.: About the relationship between eyebrow movements and fo variations. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. vol. 4, pp. 2175–2178. IEEE (1996)
6. Chovil, N.: Discourse-oriented facial displays in conversation. Research on Language & Social Interaction **25**(1-4), 163–194 (1991)
7. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019)
8. Ding, Y., Pelachaud, C., Artieres, T.: Modeling multimodal behaviors from speech prosody. In: International Workshop on Intelligent Virtual Agents. pp. 217–228. Springer (2013)
9. Duarte, A.C., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., Mohedano, E., McGuinness, K., Torres, J., Giro-i Nieto, X.: Wav2pix: Speech-conditioned face generation using generative adversarial networks. In: ICASSP. pp. 8633–8637 (2019)
10. Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)
11. Fares, M.: Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 743–747 (2020)
12. Ferstl, Y., Neff, M., McDonnell, R.: Adversarial gesture generation with realistic gesture phasing. Computers & Graphics **89**, 117–130 (2020)
13. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer graphics forum. vol. 34, pp. 193–204. Wiley Online Library (2015)
14. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
15. Guo, Y.: A survey on methods and theories of quantized neural networks. arXiv preprint arXiv:1808.04752 (2018)
16. Haag, K., Shimodaira, H.: Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In: Int. Conference on Intelligent Virtual Agents. pp. 198–207. Springer (2016)
17. Hofer, G., Shimodaira, H.: Automatic head motion prediction from speech data (2007)

18. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. International Journal of Computer Vision **127**(11), 1767–1779 (2019)
19. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
20. Knapp, M.L., Hall, J.A., Horgan, T.G.: Nonverbal communication in human interaction. Cengage Learning (2013)
21. Kucherenko, T., Jonell, P., van Waveren, S., Henter, G.E., Alexandersson, S., Leite, I., Kjellström, H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 242–250 (2020)
22. Li, X., Zhang, J., Liu, Y.: Speech driven facial animation generation based on gan. Displays **74**, 102260 (2022)
23. Lu, J., Shimodaira, H.: Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder. arXiv preprint arXiv:2002.01869 (2020)
24. Mariooryad, S., Busso, C.: Generating human-like behaviors using joint, speech-driven models for conversational agents. IEEE Transactions on Audio, Speech, and Language Processing **20**(8), 2329–2340 (2012)
25. Monzo, C., Iriondo, I., Socoró, J.C.: Voice quality modelling for expressive speech synthesis. The Scientific World Journal **2014** (2014)
26. Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2face: Learning the face behind a voice. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7539–7548 (2019)
27. Ong, W.Q., Tan, A.W.C., Vengadasalam, V.V., Tan, C.H., Ooi, T.H.: Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive nonlinear filter. Entropy **19**(11), 487 (2017)
28. Palan, S., Schitter, C.: Prolific. ac—a subject pool for online experiments. Journal of Behavioral and Experimental Finance **17**, 22–27 (2018)
29. Sadoughi, N., Busso, C.: Speech-driven animation with meaningful behaviors. Speech Communication **110**, 90–100 (2019)
30. Salem, M., Rohlfing, K., Kopp, S., Joublin, F.: A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In: 2011 Ro-Man. pp. 247–252. IEEE (2011)
31. Song, Y., Zhu, J., Li, D., Wang, X., Qi, H.: Talking face generation by conditional recurrent adversarial network. arXiv preprint arXiv:1804.04786 (2018)
32. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
33. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Transactions on Graphics (TOG) **36**(4), 1–11 (2017)
34. Titze, I.: Principles of Voice Production. Prentice-Hall Inc. (1994)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
36. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision pp. 1–16 (2019)

37. Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., et al.: Photo-realistic expressive text to talking head synthesis. In: INTERSPEECH. pp. 2667–2669 (2013)
38. Wolfert, P., Robinson, N., Belpaeme, T.: A review of evaluation practices of gesture generation in embodied conversational agents. arXiv preprint arXiv:2101.03769 (2021)
39. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
40. Zhang, Y., Wang, J., Zhang, X.: Conciseness is better: recurrent attention lstm model for document-level sentiment analysis. Neurocomputing **462**, 101–112 (2021)
41. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9299–9306 (2019)
42. Zoric, G., Forchheimer, R., Pandzic, I.S.: On creating multimodal virtual humans—real time speech driven facial gesturing. Multimedia tools and applications **54**(1), 165–179 (2011)
43. Zoric, G., Smid, K., Pandzic, I.S.: Automated gesturing for embodied animated agent: Speech-driven and text-driven approaches. Journal of Multimedia **1**(1)
44. Zoric, G., Smid, K., Pandzic, I.S.: Facial gestures: taxonomy and application of non-verbal, non-emotional facial displays for embodied conversational agents. Conversational Informatics: An Engineering Approach pp. 161–182 (2007)