Multimodal Attention for Lip Synthesis Using Conditional Generative Adversarial Networks

Andrea Vidal[a], Carlos Busso[a,*]

[a]*Department of Electrical and Computer Engineering, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, 75080, TX, USA*

Abstract

The synthesis of lip movements is an important problem for a *socially interactive agent* (SIA). It is important to generate lip movements that are synchronized with speech and have realistic co-articulation. We hypothesize that combining lexical information (i.e., sequence of phonemes) and acoustic features can lead not only to models that generate the correct lip movements matching the articulatory movements, but also to trajectories that are well synchronized with the speech emphasis and emotional content. This work presents attention-based frameworks that use acoustic and lexical information to enhance the synthesis of lip movements. The lexical information is obtained from *automatic speech recognition* (ASR) transcriptions, broadening the range of applications of the proposed solution. We propose models based on *conditional generative adversarial networks* (CGAN) with self-modality attention and cross-modalities attention mechanisms. These models allow us to understand which frames are considered more in the generation of lip movements. We animate the synthesized lip movements using blendshapes. These animations are used to compare our proposed multimodal models with alternative methods, including unimodal models implemented with either text or acoustic features. We rely on subjective metrics using perceptual evaluations and an objective metric based on the LipSync model. The results show that our proposed models with attention mechanisms are preferred over the baselines on the perception of naturalness. The addition of cross-modality attentions and self-modality attentions has a significant positive impact on the

[*] Corresponding author

Email addresses: `axv170003@utdallas.edu` (Andrea Vidal), `busso@utdallas.edu` (Carlos Busso)

performance of the generated sequences. We observe that lexical information provides valuable information even when the transcriptions are not perfect. The improved performance observed by the multimodal system confirms the complementary information provided by the speech and text modalities.

Keywords: Speech-driven animations, socially interactive agents, Conditional GAN, Lip movements, cross-modal attention, attention mechanism.

1. Introduction

Effective modeling of *socially interactive agents* (SIAs) require an authentic display of human-like behaviors to ensure strong user engagement. Advances in generating realistic behaviors can impact applications for movie productions, video games, and *human-computer interaction* (HCI) systems. Gestures, postures, and facial movements are key to create convincing animations that mimic human behaviors (Pelachaud et al., 2021). A method used to generate animations requires the mapping of movements performed by actors, which are transferred to the animated characters (Stone et al., 2004; Williams, 1990; Rizzo et al., 2004; Kipp et al., 2007; Neff et al., 2008). Another way to produce animations is with artists, who manually create the required movements. Although effective, these animation methods are time-consuming. An important research topic is to explore methods that can generate human-like animations with ease to alleviate this problem. Several studies have proposed to generate gestures using speech and text (Sadoughi and Busso, 2015; Ferstl and McDonnell, 2018; Zhou et al., 2018a; Ahuja et al., 2019; Ferstl et al., 2019; Sadoughi and Busso, 2019; Ahuja et al., 2020; Ferstl et al., 2020; Kucherenko et al., 2020, 2021). In the case of facial animation, previous studies have focused on the entire face (Lee et al., 1995; Deng et al., 2006b,a; Cao et al., 2013; Edwards et al., 2016; Karras et al., 2017; Zhou et al., 2018b; Richard et al., 2021a) or just the orofacial area (Brand, 1999; Luo et al., 2014; Fan et al., 2016; Pham et al., 2017; Suwajanakorn et al., 2017; Sadoughi and Busso, 2018b). A challenging aspect of generating facial animations is to model lip movements, since it is not acceptable to have a talking-head with a mismatch between the audio and lip movements. Therefore, over the years, different methods have been developed to create lip animations based on a combination of acoustic, text, and facial features. Early works used *hidden Markov models* (HMM) or *Gaussian mixture* *models* (GMM) to create lip synthesis animation (Brand, 1999; Luo et al., 2014). Advancements in deep learning have brought a breakthrough in this research field with methods such as the use of blendshapes for synthesizing lip movements (Pham et al., 2017, 2018), and the generation of lip synthesis without phoneme shapes to create articulations (Taylor et al., 2017; Chen et al., 2018a). Despite these advancements, the creation of convincing lip movements for facial animations is still a challenge.

This study proposes multimodal models with self-attention and cross-attention mechanisms that aim to synthesize realistic, natural lip movements from acoustic and text features. Our approach starts with *generative adversarial networks* (GANs), which have the ability to generate realistic movements using the adversary-based architecture formed by a generator and a discriminator. In our formulation, the generator creates the lip movements, which are judged by the discriminator. The discriminator detects if the movements are real of fake. This formulation is improved by adding a condition as an input, creating a *conditional generative adversarial network* (CGAN), which helps the model to contextualize the generated movements. We propose to constrain the CGAN architecture with modalities that are relevant to lip articulation: text and speech. Text defines the sequence of phonemes to be generated, improving the selection of correct movements. We do not assume that text is available. Instead, we rely on transcriptions generated by an *automatic speech recognition* (ASR) system, making our implementation feasible for a broader range of SIA applications. Speech conveys the emphasis in the message, correcting the synchronization problems from the text selection. It also introduces emotional nuances that make the model more natural. We propose a CGAN per modality, constrained by either text or speech. The two CGANs are pretrained by generating lip movements exclusively with the underlying modality. We fuse the unimodal CGANs with two alternative attention-based solutions: self-attention and cross-attention mechanisms. Self-attention mechanism determines the relevant temporal information within a modality, while the cross-attention mechanism assesses which temporal information is important by considering the two modalities. Therefore, we can observe the influence of one modality on the other. The addition of the attention mechanisms not only improves the fusion of the modalities, but also allows us to study the effect of individual modalities on the lip movement generation. We visualize the generated movements using a novel approach based on blendshapes created just with lip landmarks. We map the facial landmarks of interest to the blendshape meshes. Then, we use an affine transformation to project the 3D points of interest into the 2D space of the facial landmarks. Finally, we compute the contribution of each blendshape to the lip movements using non-negative least square and *principal component analysis* (PCA). We create the animations by combining the blendshapes.

We assess the performance of our proposed attention-based models by comparing them with several unimodal baselines trained using only speech or text features. We implement these unimodal systems with either CGAN or *bidirectional long-short term memory* (BLSTM) layers. We also compared our attention-based approaches with multimodal systems implemented with CGAN and BLSTM. The multimodal CGAN baseline concatenates the representations of pre-trained CGANs conditioned by each modality without an attention mechanism. The multimodal BLSTM baseline also fuses the unimodal BLSTM representations combining acoustic and text features. We measure performance with subjective and objective evaluations, assessing the naturalness of the animated lip movements using blendshapes. We use indirect and direct comparisons to determine the best models. The indirect evaluations measure the naturalness of the animations using a 10-point Likert scale. The proposed CGANs models with attention mechanisms obtain the highest scores (5.79 for self-attention and 5.74 for cross-attention). These scores are even higher than the score provided for animations generated with the original lip sequences (5.43). The direct evaluations compare two videos generated by alternative methods. The results show that our proposed CGANs with attention mechanisms were preferred over the other models. Fusing the CGANs with self-attention and cross-attention mechanisms leads to similar performance, indicating that both of these proposed models are competitive for this task. Overall, the main contributions of this paper are:

- We propose systematic frameworks based on CGANs and alternative attention mechanisms that fuses speech and lexical information to synthesize lip movements.

- We train the text and speech-text models using ASR transcripts, which are not as accurate as human transcriptions. We obtain promising results, showing the robustness of our method.

- We propose a method to generate an animation based on blendshapes using only lip landmarks.

This paper is organized as follows: Section 2 presents previous studies related to this paper. Section 3 introduces the proposed CGAN models implemented with attention mechanisms. It also provides the methodology used to create the 3D animations with blendshapes. Section 5 shows the experimental results obtained with our methodology. Finally, Section 6 summarizes our study, describing future research directions in this area.

2. Related Work

Synthesis of lip movements is an important problem that has been broadly studied in recent years. Data driven models for the generation of other behaviors such as head motion have achieved incredible results (Sadoughi and Busso, 2019; Sadoughi et al., 2017; Sadoughi and Busso, 2018b; Busso et al., 2007), suggesting that similar approaches can be used for synthesizing lip movements. The challenge in generating lip motion is the tight synchronization needed between the configuration of the lips and the targeted lexical content. Even small errors can greatly affect the perception of naturalness in the animations. Recently, approaches have focused on the synthesis of lip movements using deep learning solutions. These methods are based on speech and/or text to synthesize realistic trajectories for lip movements. This section reviews recent advances in this area.

2.1. Lip Movements Driven by Speech Features

Suwajanakorn et al. (2017) proposed a method for lip synthesis based on *long-short term memory* (LSTM) for a single speaker. In this work, the generation of 18 lip landmarks relies on 25 *Mel frequency cepstral coefficients* (MFCCs). Sadoughi and Busso (2017) proposed a system to generate facial movements, including lip movements using BLSTM layers. The system takes as inputs MFCCs, which are related to the orofacial area of the face (Busso and Narayanan, 2007), and *low-level descriptors* (LLDs) from the *extended Geneva minimalistic acoustic parameter set* (eGeMAPS) (Eyben et al., 2016). The eGeMAPS features are commonly used for emotion-related recognition problems. These features were expected to provide complementary information about the underlying emotion in the sentence. Karras et al. (2017) proposed an end-to-end system, which generates a 3D facial mesh from raw audio. The end-to-end system is composed of *convolutional neural networks* (CNNs) and is divided into three parts: feature extraction, co-articulation parameters, and generation of the vertex position for the 3D mesh. This approach not only generates lip motion but also adds an emotional component to the animated mesh. Other approaches have focused on generating a more realistic talking-head with the inclusion of emotions. Pham et al. (2018) proposed a new end-to-end system trained using a spectrogram. This approach is based on CNNs, adding the temporal modeling with a *recurrent neural network* (RNN). The output of this system corresponds to the weights of the shapes to synthesize the animations. Sadoughi and Busso (2018a) hypothesized that emotion and co-articulation are related. For this reason, they proposed two multitask models, which have as a primary task the prediction of lip movements. The first model received as input MFCCs and eGeMAPS features, which were fused and used to predict lip movements, triviseme, and emotion. The second model received as input MFCC features, which are conditioned by the emotion prediction from a separate model using as input the eGeMAPS features. The second model predicted lip movements and triviseme. Sadoughi and Busso (2021) proposed an expressive lip synthesis method using a conditional sequential GAN. The generator and discriminator of this network are composed of BLSTM layers, capturing past and future information. In this work, the input noise is not only constrained by the speech features related to orofacial movements (MFCCs) but also emotional speech features such as LLDs and eGeMAPS, adding emotional information to the generated trajectories.

Lip movements have also been modeled by generating a complete facial animation that includes lip movements. Some studies rely on speech features extracted from an architecture trained for ASR tasks, hence having implicit lexical information. Cudeiro et al. (2019) presented VOCASET, which is a new dataset of 4D face scans. They proposed a new model based on convolutional layers to generate 3D facial animation in a speaker-independent fashion, which helps with the generalization of the created animation. Richard et al. (2021b) proposed a method for generating 3D facial animations, where the speech-driven facial animation is done by distinguishing which part of the audio contributes more to the movements of the upper or lower part of the face. They created two networks for this purpose. One of them combines the information of the expression and the speech. The second one is an UNet decoder that passes that information to the latent space of a particular speaker, generating the animation. Despite the promising results, one of the disadvantages of the method is the amount of data required to generate the facial animation. Chai et al. (2022) presents a network architecture based on CNNs used to extract features from Mel-spectrogram, which are followed by

RNNs to extract temporal information. This temporal information is the input of a temporal attention mechanism, which helps to emphasize important frames to generate facial animation. Fan et al. (2022) presents FaceFormer, which is based on the transformer architecture where one of the attention mechanisms is performed over the extracted speech features and the facial animation. This architecture helps the alignment between facial movements and speech. In the studies of Chai et al. (2022) and Fan et al. (2022), the model was aware of the speaker's identity. The incorporation of an attention mechanism for a 3D facial animation has been done, to our knowledge, only for speech-driven approaches.

2.2. Lip Movements Driven by Text Features

Text has also been used to synthesize lip motion. Sako et al. (2000) proposed an HMM-based text-to-audio-visual system for speech synthesis. The proposed model had two parts: auditory HMM and visual HMM. The auditory HMM received as input the text and synthesized the speech signal. The visual HMM received as input the text, but also the duration state of each phoneme to reconstruct the correct lip configuration. Minnis and Breen (2000) proposed a framework for modeling phoneme coarticulation. The approach takes a sequence of phonemes, which are concatenated. They animated the lip movements by using the trajectories generated by these concatenations. Tang et al. (2008) presented a novel framework to synthesize lip movements using text. The approach synthesizes an emotional speech signal using diphones, obtaining the duration of the phonemes. Then, the phonemes and their duration are mapped into visemes, creating the animations. Stef et al. (2018) proposed a speech animation method based on text, which was converted into *international phonetic alphabet* (IPA) symbols. The animation was realized using blenshapes created for the IPA symbols. Taylor et al. (2017) proposed a speaker independent method for synthesizing lip movements based on a *deep neural network* (DNN). The DNN received as input a sliding window of phonemes to predict the coarticulation curve. Because of the overlapping of the sliding windows, they interpolated the coarticulation curve by taking the mean. Chen et al. (2018b) proposed a lip synthesis method based on blendshapes with phoneme shapes. They built an ASR to generate the phoneme sequence, which was used to create an animation using the shapes that represent each phoneme.

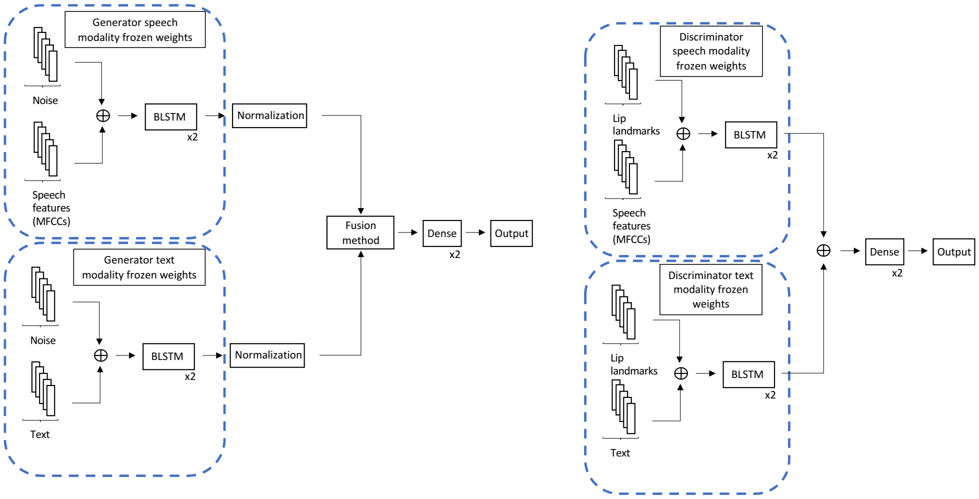
2.3. Lip Movements Driven by Speech and Text Features

Edwards et al. (2016) proposed JALI, which is a system that uses speech and text to generate lip-synchronized facial animation. Speech was used to align the phonemes with the audio. Then, the system generated the animation mapping the phonemes to visemes. The sequence of visemes depends on the jaw and lip movements of each phoneme. Fan et al. (2016) proposed a method for generating lip motions based on speech and text using a BLSTM model. They use MFCCs as speech features and triphones as text features. They synthesize a video using 48 facial landmarks. Liu et al. (2020) proposed an autoencoder and a regressor based on CNNs to generate talking faces using speech and lexical information. The autoencoder generates representations for the upper right, upper left, and lower parts of the face. These representations are used as a ground truth by the regressor, which takes as input the phoneme representation, which is aligned with the speech.

2.4. Relation to Previous Studies

This work introduces a methodological framework for synthesizing lip movements using audio and lexical information. We propose two CGAN-based architectures with self-attention and cross-attentions to capture the relation between acoustic and lexical features while creating the lip movement trajectories. Unlike previous studies that use speech and text features (see Sec. 2.3), we train our models using transcriptions from an ASR system, which are less accurate compared to human transcriptions. Even with noisy transcriptions, the models achieve strong performance. We use 3D blendshapes to perform the animation, which simplifies the animation process because the animation is done only using the lip landmarks generated by the models.

The most similar study to this paper is the work of Sadoughi and Busso (2021), which proposed a CGAN-based architecture, constraining the lip movements by the acoustic features. We build our model starting from this architecture, proposing major changes to improve its performance. The main contributions of this study with respect to Sadoughi and Busso (2021) are (1) the use of lexical information, in addition to speech, to constrain the generated lip movements, and (2) the use of self-attention and cross-attention to fuse the lexical and acoustic information. These contributions greatly improve the performance of the system by explicitly indicating the predicted sequence of phonemes, which is tightly synchronized with the acoustic features using the proposed fusion approaches. Furthermore, we render the animation using blendshapes, which is an important improvement over the study of Sadoughi and Busso (2021), which was limited by using *facial action parameters* (FAPs).



(a) Generator Speech-Text

(b) Discriminator Speech-Text

Figure 1: Proposed network architectures based on *conditional generative adversarial networks* (CGANs). The text and speech modalities are fused in the generator using the attention-based models described in Section 3.3. The symbol ⊕ represents the concatenation operation.

3. Proposed Approach

We propose attention-based fusion methods of unimodal CGANs trained with acoustic and lexical features to generate realistic lip movements. Figure 1 shows the proposed architecture. This section describes the proposed methods, including a general description of the proposed CGAN (Sec. 3.1), the unimodal blocks (Sec. 3.2), and the proposed attention-based fusion method of the modalities (Sec. 3.3). The lip trajectories are animated with our proposed blendshape approach that takes as input the facial landmarks created by our method (Sec. 3.4).

3.1. Generative Adversarial Networks

Generative adversarial networks (GANs) were presented by Goodfellow et al. (2014). This framework consists of two networks trained with an adversarial loss: a generator $G(\cdot)$ and a discriminator $D(\cdot)$. The idea behind

GAN is to solve a minimax problem. The generator tries to create samples that resemble the training data distribution to fool the discriminator. The input of the generator is a noise signal, and its output is the generated sample following the distribution of the real data. The discriminator has to decide whether a sample comes from real data or from data created by the generator. The formulation for GAN is:

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D(G(z)))\right], \tag{1}$$

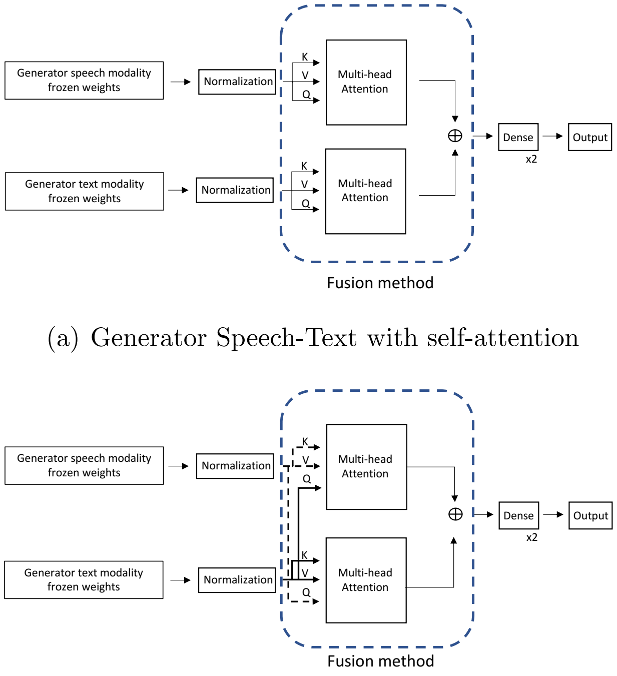
where x is the sample with probability distribution p_{data}, and z represents a noise signal with probability distribution p_z. This adversarial method has been used in multiple tasks, such as generating images (Reed et al., 2016), learning feature representations (Chang and Scherer, 2017), and detecting data anomalies (Qiu et al., 2019, 2022). The success of this methodology on several fields has led to the creation of several methods derived from the GAN framework. One of them is called Conditional GANs. Mirza and Osindero (2014) proposed a variation of this framework creating the *conditional Generative adversarial network* (CGAN). Unlike GANs, CGANs take an extra input as a condition to constrain the model, adding more information to improve the sample generated by the generator. The cost function is defined as

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x|y)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D(G(z|y)))\right], \tag{2}$$

where y represents the extra information fed to the CGAN network. For our task, the CGAN framework offers the advantage of constraining the generation of the samples with acoustic or lexical features (or both).

3.2. CGAN-Based Models for Speech and Text

As shown in Figure 1, we build two separate CGANs with the same architecture. One was constrained with lexical features and the other with acoustic features. The generator concatenates the noise and the features from the modality that constrains the model. We build the CGANs with two BLSTM layers, each of them implemented with 128 nodes. The discriminator follows the same architecture of the generator with two BLSTM layers, each of them implemented with 64 nodes. The BLSTM layers provide temporal modeling to capture the dynamic relationship between the modalities and the lip trajectories. For the text modality, the model is conditioned by the underlying sequence of phonemes. We represent the aligned sequence of phonemes by a one-hot-encoding as input to the model. For the speech modality, the noise is conditioned by the MFCCs features extracted from the speech. The acoustic information not only synchronizes speech with the phonetical units, but also introduces emphasis and emotional nuances that are important during natural human interactions. As a result, we expect better results by combining both modalities.



(a) Generator Speech-Text with self-attention

(b) Generator Speech-Text with cross-attention

Figure 2: Proposed fusion approaches using self-attention (Fig. 2(a)) and cross-attention (Fig. 2(b)). The symbol ⊕ represents the concatenation operation.

Figure 1 shows the architecture of the generator and discriminator of the CGAN used for this study. The proposed models use the same configuration for the generator and discriminator, but they differ in the fusion method, which is only implemented on the generator (Sec. 3.3). Figure 1(b) shows the discriminator model, which fuses the pretrained discriminator of each modality. We freeze the parameters of the discriminator for each modality and extract features from the second BLSTM layer, which are concatenated. Then, they are passed through two dense layers implemented with 128 and 64 units, followed by the output layer. Figure 1(a) illustrates the generator model, which comprises the pretrained CGAN models for the acoustic and lexical features. We freeze the generators of each modality and extract the features of the second BLSTM layer as our feature representation. These features are normalized by dividing them by its 98% percentile value. The normalized features are processed by the fusion module. Then, the processed features pass through two dense layers, each of them implemented with 256 and 128 units, respectively. The output is the 2D coordinates of the 20 facial landmarks used for the animation (Fig. 5).

3.3. CGAN Attention-Based Fusion Models

We propose two alternative attention-based fusion methods, which are shown in Figures 2(a) and 2(b). After the normalization block (Fig. 2(b)), the feature representations for both modalities pass through an attention-based model. We use two alternative attention mechanisms: self-attention and cross-attention. We use the multi-head attention presented by Vaswani et al. (2017),

$$\text{MultiHeadAttention}\left(Q, K, V\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_H\right) W^O, \tag{3}$$

where Q, K, and V represent the query, key, and value matrices respectively. These matrices are $N \times d_k$, where N is the number of frames and d_k is the feature dimension. W^O is the projection matrix of the output, $\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$ where Attention is the scaled dot-product attention (Vaswani et al., 2017). W_i^Q, W_i^K, W_i^V represent the projection matrices for query, key, and value matrices respectively with $i \in \{1, \ldots, H\}$ the number of heads. We set the number of heads to one.

3.3.1. Self-Attention Fusion Method

Our first fusion approach is the self-attention mechanism, which is independently implemented for each modality. Figure 2(a) illustrates the generator model with self-attention as a fusion method. The query, key, and value features from Equation (3) come from the same modality, as shown in Equations (4) and (5).

$$\text{MultiHeadAttention}\left(Q_{\text{text}}, K_{\text{text}}, V_{\text{text}}\right), \tag{4}$$

$$\text{MultiHeadAttention}\left(Q_{\text{speech}}, K_{\text{speech}}, V_{\text{speech}}\right). \tag{5}$$

This fusion method determines which temporal information is relevant within a modality for the synthesis of lip movements. It also provides insights to increase the interpretability of the models by showing which frames are important for the fusion. We refer to this method as CGAN speech-text with self-attention.

3.3.2. Cross-Attention Fusion Method

The second model to fuse the modalities in the generator is the cross-attention mechanism. Figure 2(b) shows the generator model with this fusion method. This approach has been successfully used in multimodal fusion problems (Goncalves and Busso, 2022; Tsai et al., 2019). In contrast to the self-attention mechanism, cross-attention integrates both modalities while estimating the attention. The key and value vectors come from one modality, while the query vector comes from the other modality. Equations (6) and (7) show this cross modality formulation.

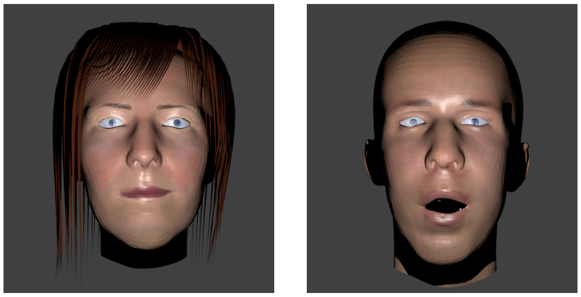
$$\text{MultiHeadAttention}\left(Q_{\text{speech}}, K_{\text{text}}, V_{\text{text}}\right), \tag{6}$$

$$\text{MultiHeadAttention}\left(Q_{\text{text}}, K_{\text{speech}}, V_{\text{speech}}\right). \tag{7}$$

In addition to studying which temporal information is relevant while combining the modalities, this fusion method leverages how much a modality influences the temporal relevance of the other modality. For example, if the relevant information for the speech modality is in the future and relevant information for the text modality is in the past, this fusion method can attend to the past and future for both modalities. However, the level of attention could be different for each modality. We refer to this method as CGAN speech-text with cross-attention.

3.4. Facial Animation using Facial Landmarks

In our previous speech-driven animation studies (Mariooryad and Busso, 2012; Sadoughi and Busso, 2017; Busso et al., 2005; Sadoughi and Busso,



(a) Female animation (b) Male animation

Figure 3: Examples of the female and male models for our animations using blendshapes. The proposed approach takes facial landmarks and generates the weights of the blendshapes to animate the sequence.

2021), we have relied on FAPs (Ostermann, 1998), which were used to synthesize the animation. This approach lacks the flexibility needed to synthesize realistic animations. Therefore, we create our own prototype for the animation where we only specify the value of the predicted facial landmarks. For this purpose, we use 47 blendshapes with different facial expressions from the FaceWarehouse database (Cao et al., 2014). We consider orofacial areas with different shapes. We generate the animation using those blendshapes using the open-source program Blender. Figure 3 shows examples of the animations in Blender created with our approach. We create a female model and a male model.

Our models are trained with data from different subjects, who have different facial anatomy. Therefore, we normalize the facial landmarks to reduce the high variability across subjects. The normalization relies on the method described by Eskimez et al. (2018). This method aligns the facial landmarks and removes the identity information by obtaining a mean representation across the faces. This is achieved by selecting a global reference frame with a neutral facial expression with a frontal pose. We also select a speaker reference frame with a neutral face for each subject in our dataset. The speaker reference frames are translated and rotated to match the pose of the global reference, compensating for different sizes and head orientations across speaker reference frames. For each frame in the video, we use the speaker reference frame to align and rotate its facial landmarks. This step compensates for variations in size of the faces across frames due to the distance between the subject and the camera. Then, we remove the identity of the scaled facial landmarks. For this purpose, we calculate the difference between the current frame and its speaker reference frame, capturing just the facial movements. These differences are added to the global reference frame, removing the speaker differences.

We perform the animation when we have all the facial landmarks normalized. We select the facial landmarks of a neutral face as a reference. For animation purposes, we map the facial landmarks around the mouth of the mean representation into the vertices of the neutral-shape mesh. Once we have this mapping, we project the 3D blendshapes into the 2D facial landmark space using the camera matrix estimation method used by Huber et al. (2016), which uses the *gold standard algorithm* (GSA) (Hartley and Zisserman, 2004) to compute a normalized version of the camera matrix. For this purpose, the data is normalized so that the 2D and 3D representations are translated moving their centroids to the origin of the coordinates. The 2D and 3D coordinates are scaled such that the root mean square distances between the average location of the facial landmarks and the origins are $\sqrt{2}$ and $\sqrt{3}$, respectively. Finally, it obtains the affine camera matrix, which is used to project the 3D blendshape mouth landmark into the 2D space. We can find the blendshapes that represent each frame by using the non-negative least square method, which is constrained by the Tikhonov's regularization (L2-regularization) (Bishop, 1995). The non-negative least squared assigns a weight per each blendshape. We constrain the sum of the weights to be between 0 and 1. We implement this constraint to avoid changing the size of the blendshape or creating an artifact in the animation. One of the issues with this optimization problem is that blendshapes corresponding to the articulation of phonemes with lip protrusion such as /u/ or /o/ was not correctly represented. The mouth appears just open for these phonemes. We address this issue with *principal component analysis* (PCA) followed by non-negative least square. We obtain another set of weights by reducing the dimensionality of our problem from 40 coordinates (i.e., the x and y coordinates for the 20 facial landmarks) to 8 by using PCA. PCA reduces the dimensionality of the problem by selecting the eigenvectors with the highest variances (e.g., eigenvectors associated with the highest eigenvalues). We select the eight eigenvectors that have the highest variance following the work of Suwajanakorn et al. (2017). PCA emphasizes horizontal movements of the lips, solving the issue that we originally observed for phonemes such as /u/ or /o/, which are now correctly animated. We obtain the final weights for the blendshapes to animate the sequence by averaging the two sets of weights generated by the mouth landmarks and PCA method.

4. Experimental Settings

4.1. Database

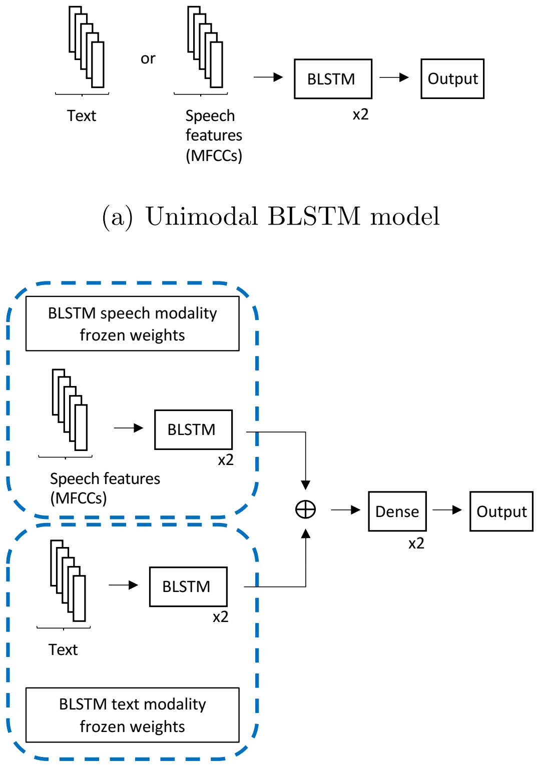
The proposed models are trained and tested using the MSP-Face database (Vidal et al., 2020), which was collected at The University of Texas at Dallas. The MSP-Face corpus is a collection of online videos of people talking in front of a camera. They discussed diverse topics such as video games, feelings, opinions, and experiences during their depression. Each video was manually segmented, creating speaking turns with a duration between 3 and 10 seconds. The characteristics of those videos are single speaker, frontal face, clear speech, and no background music. In total, we collected $27,325$ video segments, which approximately correspond to 70 hours of video data. This database provides a variety of emotional displays from multiple speakers. The number of speakers in this database is 491 speakers. Although our study does not use emotional labels, the corpus was annotated with emotional descriptors using perceptual evaluations in a crowdsourcing platform. We manually select a subset of videos with frontal faces. Then, we exclude videos if we are not able to automatically obtain facial landmarks for 85% of the frames in a video. Lip landmark detection is not always reliable, regardless of the library used to extract them. Some of the extracted frames could present some issues, where, for example, landmark on the internal side of the lips may not be well tracked. Frontal faces are often better. We curated a sub-set of the MSP-Face for our task for this reason since we need to minimize the error propagated by the lip landmarks. With this data selection, we use a set of $6,994$ videos, which correspond to 17.9 hours. The size of this set is larger than the training set used in many of the studies that aim to generate human behaviors (Sadoughi and Busso, 2021; Richard et al., 2021a; Fan et al., 2016). These videos were partitioned into the train (78%), and development (22%) sets, using speaker independent sets (i.e., no overlap of speakers in the sets). We re-segment original videos from the MSP-Face corpus to create the test set used to evaluate our models. These videos are 30 seconds long, and are not included in either the train and development sets. The longer duration of the test videos allows our evaluators to clearly perceive the synchronization between the generated lip movements and the lexical content. For more information about this corpus, the readers are referred to the work of Vidal et al. (2020).

The videos selected from the MSP-Face database do not have the same frame rate. We fixed the frame rate to 29.97 across all the videos for the extraction of the features. If the video frame rate does not match with the previously mentioned rate, we interpolated or extrapolated the facial landmarks depending on the case. We extract three features from the videos: speech features, facial landmarks, and phonemes. The speech features are obtained using the Python library Librosa (McFee et al., 2019). We extract the signal from the videos, formatting the audio into a mono channel with a sample rate equal to 16 KHz. From the audios, we extract 25 MFCCs with a window size of 25 ms and a stride of 33.3 ms to match the fixed frame rate of the videos. From the video, we extract the facial landmarks using the DLib library (King, 2009), which gives us 68 facial landmarks. This study only uses 20 landmarks corresponding to the lip area (Fig. 5). The third set of features is the lexical content, which is represented by the sequence of phonemes that are aligned with the audio. We do not rely on manual transcriptions. Instead, we obtain automatic transcriptions using the ASR included in the Microsoft Indexer solution. The ASR transcriptions are processed with common modifications such as changing numbers and symbols into words (e.g., $ into dollar). After these changes, we align the audio and the ASR transcriptions using the Montreal Forced Aligner (McAuliffe et al., 2017). We use 77 phonemes plus 4 silence markers, representing the lexical content with an 81-dimensional one-hot vector.

For training and testing our models, we use a sliding window with 15 frames and a stride of one frame to transform the extracted features per video into a window sequence. This processing strategy creates overlapping between the windows. As a result, each frame will be determined by 14 frames from the past and 14 frames from the future, modeling the co-articulatory movements.

4.2. Baselines

We use several baselines to evaluate the models. We implement baselines using BLSTM layers, which is a common framework in speech-driven lip movement generation (Fan et al., 2016; Sadoughi and Busso, 2018a). We use two unimodal baselines and one multimodal baseline implemented with BLSTMs. The unimodal baselines for speech (BLSTM Speech) and text (BLSTM Text) follow the same architecture, which includes two BLSTM layers, each of them implemented with 128 nodes and a 40D output layer



(a) Unimodal BLSTM model

(b) BLSTM Speech-Text

Figure 4: BLSTM-based baselines for the unimodal and multimodal frameworks. The symbol ⊕ represents the concatenation operation.

(Fig. 4(a)). The multimodal BLSTM baseline fuses speech and text information (BLSTM Speech-Text). It relies on the pretrained BLSTM unimodal baselines. We extract feature representations for those pretrained models, which are concatenated, and further processed by two dense layers implemented with 256 and 128 nodes, respectively (Fig. 4(b)). The output layer creates the 40D vector with the trajectories.

We also implement two unimodal baselines and one multimodal baseline using the CGAN structure. The unimodal CGAN baselines are implemented by constraining the model with either speech (CGAN Speech) or text (CGAN Text). These models demonstrate the benefits of using both modalities. The multimodal method implements the generator by concatenating the feature representations of the CGANs (CGAN Speech-Text). The attention mechanisms are not used, replacing them by concatenating the feature representation of the modalities. This multimodal baseline demonstrates the benefits of using the proposed attention-based fusion approach.

4.3. Implementation Details

The baselines that only consist of BLSTM layers are trained for 200 epochs and use Adam (Kingma and Ba, 2015) as the optimizer with a learning rate of $1e^{-5}$. We pretrain the generator and discriminator of the CGAN models for 10 epochs. Then, the CGANs for each modality are trained for 100 epochs and the fusion models for 120 epochs. The optimizer for these models is Adam with a learning rate of $1e^{-5}$. Finally, the loss function for all the baselines models and the generators is the addition of the *concordance correlation coefficient* (CCC) (Trigeorgis et al., 2016) loss $(1-CCC)$ and the *mean squared error* (MSE) of the six *inner mouth landmarks* (Fig. 5). The variables y and $\hat{y}$ are the true and predicted landmarks, respectively. For the CCC loss (Eq. 8), σ is the standard deviation, μ is the mean, and the ρ is the Pearson correlation coefficient between the true and the predicted landmarks.

$$CCC = \frac{2\rho\sigma_{\hat{y}}\sigma_{y}}{\sigma_{\hat{y}}^{2} + \sigma_{y}^{2} + (\mu_{\hat{y}} - \mu_{y})^{2}} \tag{8}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{9}$$

$$\mathcal{L} = (1 - CCC) + MSE \tag{10}$$

The MSE loss (Eq. 9) helps the model to capture when the mouth is closed or opened, which is particularly useful for phonemes such as /m/, /p/, /b/, /æ/ and /oʊ/. Meanwhile, the *CCC* loss helps to increase the correlation between the predicted and true values, while minimizing the distance between their trajectories (squared difference of the means in the denominator of Eq. 8). While the CCC loss is bounded, the MSE is not and can take big values dominating the combined loss. Therefore, we prevent this problem by dividing the MSE by the highest MSE value of the initial batch. This step is only implemented in the initial batch. All the models are trained with a batch size of 128 sequences and a sequence of 15 frames as an input to the model.

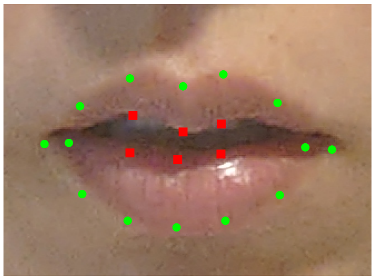


Figure 5: Illustration with the 20 facial landmarks used in this study to generate lip movements. The landmarks highlighted in red (squares) correspond to *inner mouth landmarks* used to estimate the MSE in Equation 9.

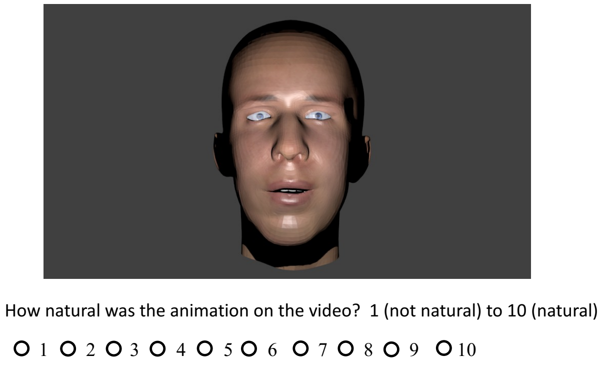
5. Experimental Results

This section presents the results obtained by our proposed multimodal models based on CGANs and the attention mechanisms. We design subjective and objective evaluations to assess the naturalness of the animations (Sec. 5.1), and a subjective evaluation to directly compare preference between alternative methods (Sec. 5.2). Section 5.3 presents an analysis of the attention weights estimated while fusing the text and speech modalities.

5.1. Perceptual Evaluation of Naturalness

We evaluate our models with subjective and objective evaluations. In total, we have nine alternative animations per video created with different methods. We have two proposed models implemented with alternative attention-based fusion methods (Sec. 3.3). We have three BLSTM baselines and three CGANs baselines (Sec. 4.2). As a reference, we also have a model animated with the original facial landmarks in the recordings. We create 20 videos for each of the nine methods, where 10 videos are rendered with the female character, and 10 with the male character. These animations are generated with the methodology described in Section 3.4.

The perceptual evaluation assesses the naturalness of the animation, annotating one video at a time. For this evaluation, we hired 10 student workers at The University of Texas at Dallas. For the proposed methods and the baselines, each annotator evaluated 10 videos for each method. For the videos with the original facial landmarks, each annotator evaluated all the 20 videos (i.e., each annotator evaluated a total of 100 videos). We ask the annotators to assess “How natural is the animation?” with a 10-point Likert scale



How natural was the animation on the video? 1 (not natural) to 10 (natural)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Figure 6: Graphical interface presented to the annotators to assess the naturalness level of the generated animations.

Table 1: Evaluation of the naturalness perceptual of the animations. Our proposed models are compared with animations generated with the original lip trajectories and with competitive unimodal and multimodal baselines.

Method	Mean	Standard deviation
Original	5.4313	2.0851
BLSTM Speech	3.9875	2.2133
BLSTM Text	4.0750	1.8878
BLSTM Speech-Text	4.1625	2.0466
CGAN Speech	4.7875	2.2764
CGAN Text	5.6750	2.0424
CGAN Speech-Text	5.0000	2.0189
CGAN Speech-Text (self-att)	**5.7875**	2.1447
CGAN Speech-Text (cross-att)	5.7375	2.1093

with extremes values 1 (not natural) and 10 (natural). Figure 6 shows the interface used for the evaluation. As a post-processing step, we estimate the inter-agreement of the annotators. Two of the student workers showed low inter-evaluator agreement and were removed from the annotations, which is a common post-processing approach when using perceptual evaluations (Cao et al., 2014; Mower Provost et al., 2015; Kaur et al., 2018; Gupta et al., 2016).

Table 1 shows the results obtained for this evaluation. We measure the significance of the results by performing multiple t-test. The t-test assumes normal distributions. We verified the homogeneity of the variance by using

Bartlett's test, which returned a p-value = 0.36. This result indicates that we cannot reject the null hypothesis. The groups are expected to have the same variance (i.e., homogeneity of the variance in the data). We compare the scores obtained by our proposed multimodal models and the baselines. We correct the p-values of the multiple t-tests by applying the Bonferroni method, and we assert significance when the p-value is less than 0.05. Table 1 shows that the animations generated by the proposed CGAN speech-text models with the attention mechanisms are perceived as the most natural animations. The CGAN speech-text model with cross-attention achieved an overall score of 5.7375, while CGAN speech-text model with self-attention reached even a higher score equal to 5.7875. The proposed CGAN speech-text models with the attention mechanisms are significantly better than all the baseline models, with the exception of the CGAN text model, which achieved an overall score equal to 5.6750. It is remarkable that the lip trajectories created by the proposed methods were perceived more natural than the animations created with the original facial landmarks, which received an overall score equal to 5.4313.

The results presented in Table 1 for the CGAN models show that text-only method performed better that speech-only method. While speech features help with the synchronization between the lip movements and the audio, and with the emphasis of the sentence, the lexical content emphasizes the right coarticulation which is not directly present in the speech. This emphasis on the right sequence of coarticulation during speech production makes the animations to be perceived as more natural than the animations generated from speech features. Any wrong coarticulation generated with speech will create visible artifacts reducing its naturalness perception.

The results in Table 1 show the contribution of the attention mechanisms to the performance of the methods. The CGAN speech-text approach achieved an overall score of 5.0. This method concatenates the representations from the modalities. Concatenation includes features from both modalities, even when one modality (e.g., text) may be better than the other (e.g., speech). Since the speech-only model has lower performance than the text-only model makes the concatenation approach less effective, leading to a fusion strategy that is not able to outperform the best unimodal system, in this case, the text-only model. For this reason, the integration of speech and text features by using attention is relevant to this problem. The attention mechanism takes the best of each modality, improving the overall performance of the model. When we add the attention mechanisms, we observe overall scores above 5.7, which represents a relative gain of more than 14% in the naturalness perception of the animated videos. Table 1 shows that the performances of the CGAN methods lead to better performance than the BLSTM baselines. The differences in the results between the corresponding CGAN and BLSTM methods are statistically significant.

In addition to the subjective evaluation, we aim to include objective metrics to assess the quality of the video. MSE and correlation have not been good metrics for measuring the generation of human behaviors (Kucherenko et al., 2020; Yoon et al., 2022). The main issue with these objective metrics is that they do not correlate well with subjective metrics, which is the main goal in synthesizing naturalistic behaviors. For these reasons, we present the LipSync metric proposed by Chung and Zisserman (2016) as an objective metric. LipSync measures the synchrony between the lip movements from a video and its audio. The output of this metric is a distance, where lower values indicate higher synchrony between the audio and the video. Notice that while the LipSync original model was designed for real images, we demonstrate in this study that this approach can also be useful as an objective metric to evaluate facial animations.

Figure 7 shows the results that we obtain by using the LipSync metric on the animations generated with (1) our methods, (2) animations with the original landmarks, and (3) the true videos. We observe that the true videos reach the lowest distance, as expected since the audio and the lip movements are perfectly correlated on the video. Our proposed methods CGAN Speech-Text (self-att) and CGAN Speech-Text (cross-att) achieve higher synchrony (lower distance) than the other models, which agrees with the subjective evaluation presented in Table 1. Figure 7 also shows that generative methods based on CGAN offer better synchrony than the BLSTM methods. This result also agrees with the subjective evaluations presented in Table 1.

5.2. Preference Between Alternative Methods

The second perceptual evaluation directly compares the preference between alternative methods. We present two videos and we ask the annotators which animation is more natural. Figure 8 shows the survey used for this evaluation. Given the results obtained in Section 5.1 we consider four comparisons.

- BLSTM speech-text & CGAN speech-text(self-att)
- CGAN speech-text & CGAN speech-text(self-att)

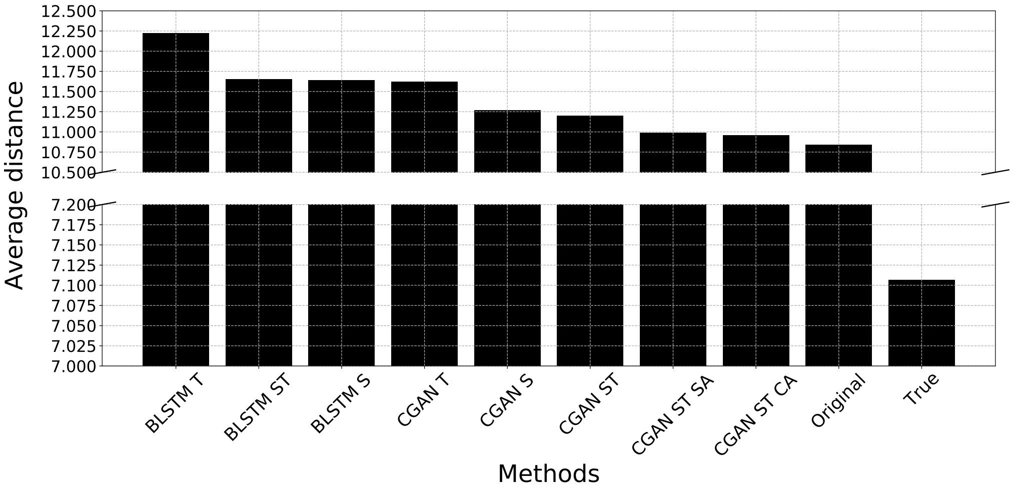
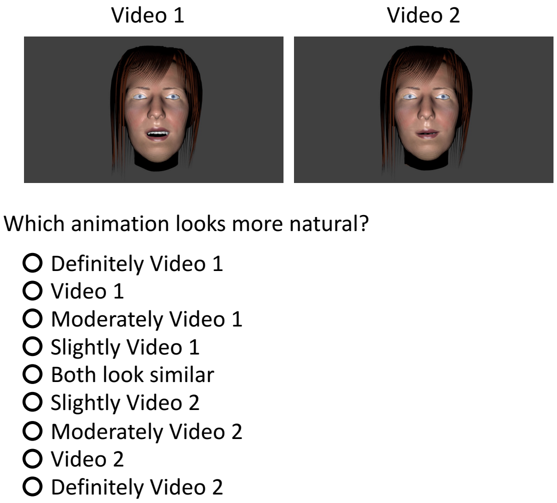


Figure 7: Results using the LipSync objective metric. Lower distance means better synchrony. BLSTM S: BLSTM Speech; BLSTM T: BLSTM Text; BLSTM ST: BLSTM Speech-Text; CGAN S: CGAN Speech; CGAN T: CGAN Text; CGAN ST: CGAN Speech-Text; CGAN ST SA: CGAN Speech-Text (self-att); CGAN ST CA: CGAN Speech-Text (cross-att); Original: Animations using original lip landmarks; True: Original videos.

- CGAN speech-text & CGAN speech-text(cross-att)
- CGAN speech-text(self-att) & CGAN speech-text(cross-att)

These comparisons include the best BLSTM baseline with the proposed model that achieved the best results. It also directly compares the benefits of using the attention mechanics as an alternative to concatenating the feature representations of the individual CGANs. It also directly compares our proposed approaches. For this evaluation, we use the 20 animations (10 females and 10 males) per each method. We recruited six additional student workers who did not participate in the evaluation presented in Section 5.1. Each comparison has six evaluations. The order of the videos was randomized, appearing as either "video 1" or "video 2". We transform the alternatives provided in Figure 8 into percentages to quantify the results using the mapping in Table 2.

Figure 9 presents the results of this evaluation. We perform a statistical analysis (t-test) on each comparison taking 50% as the null hypothesis. We assert significance if the p-value< 0.05. We observe that CGAN speech-text self-attention approach is preferred over the BLSTM speech-text approach, where the preference is statistically significant. The second comparison be-



Which animation looks more natural?

- ○ Definitely Video 1
- ○ Video 1
- ○ Moderately Video 1
- ○ Slightly Video 1
- ○ Both look similar
- ○ Slightly Video 2
- ○ Moderately Video 2
- ○ Video 2
- ○ Definitely Video 2

Figure 8: Graphical interface presented to the annotators to compare animations generated by two different methods.

Table 2: Mapping of preference between alternative methods into percentages. This mapping is used to quantify if one method is preferred over the other.

Option	Video 1	Video 2
Definitely Video 1	100%	0%
Video 1	90%	10%
Moderately Video 1	75%	25%
Slightly Video 1	60%	40%
Both look similar	50%	50%
Slightly Video 2	40%	60%
Moderately Video 2	25%	75%
Video 1	10%	90%
Definitely Video 2	0%	100%

tween our proposed CGAN speech-text self-attention method and the CGAN speech-text approach shows that using self attention leads to a 60% preference, which is a statistically significant difference. The comparison between the CGAN speech-text and GAN speech-text cross-attention suggest that there is no clear preference (p-value= 0.08). Finally, the figure shows similar preference for our two proposed models, where the difference is not statisti-

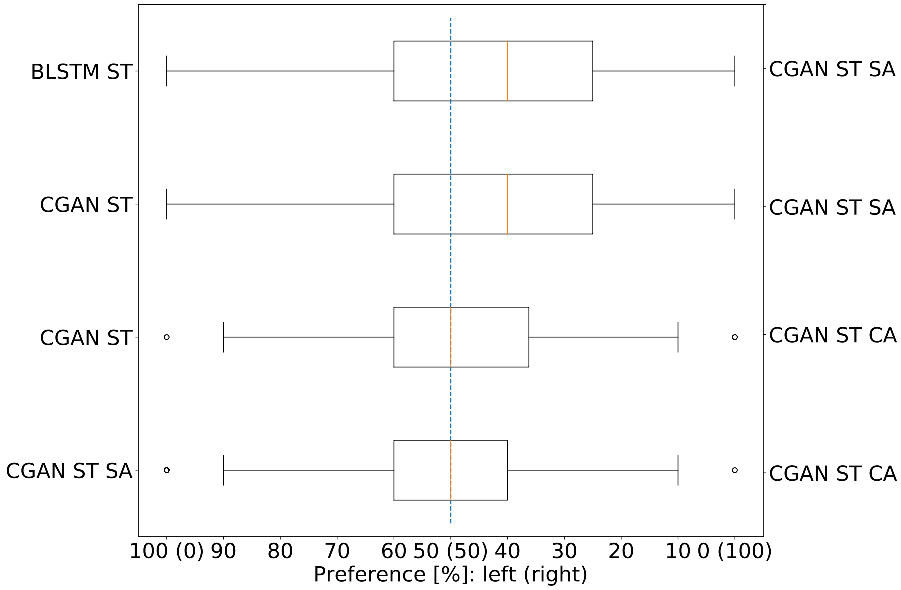
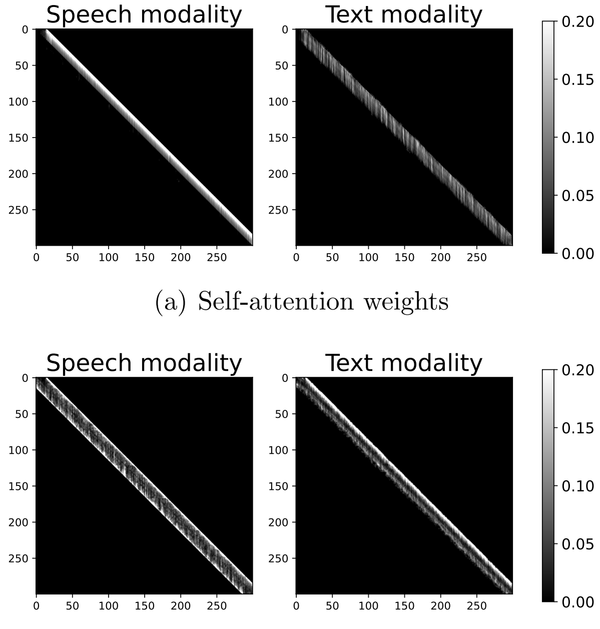


Figure 9: Plot that represents the preferences of the methods evaluated (BLSTM ST: BLSTM Speech-Text; CGAN ST: CGAN Speech-Text; CGAN ST SA: CGAN Speech-Text Self-att, and CGAN ST CA: CGAN Speech-Text Cross-att). The extremes of the box represent the 25th and 75th percentile values, the red line corresponds to the median of the evaluations, the circles represent outliers, and the dashed vertical blue line indicates the 50% of the preferences.

cally significant (p-value$= 0.3$). These results are consistent with the perceptual evaluation presented in Section 5.1 (Table 1), where the performance of the CGAN speech-text self-attention and CGAN speech-text cross-attention models are very similar. Collectively, these results show that our proposed methods with attention mechanisms outperform and improve competitive baselines on the synthesis of lip movements using only facial landmarks.

5.3. Weight Analysis of the Fusion Approaches

The proposed models use the attention mechanism (Figs. 2(a) and 2(b)). The weights estimated by these models provide insights to determine the temporal information considered by each modality in the generation of the lip trajectories. For the analysis, we extract the attention weights for the self-attention and cross-attention and analyze their patterns across different scenarios. Given the segmentation of the sequence into windows (Sec. 4.1), the context contribution for the attention weights is limited to a diagonal



(a) Self-attention weights

(b) Cross-attention weights

Figure 10: Attention weights values for the frameworks based on self-attention and cross-attention mechanisms. The horizontal and vertical axes represent the number of frames in a sentence. The horizontal axis corresponds to the frames from the key modality and the vertical axis correspond to the frames of the query modality.

block structure that includes the main diagonal, 14 past frames, and 14 future frames. Figure 10 shows the average values of the attention weights for the 20 videos in the test set. The colors assigned to a row represent the weight values assigned to the frames at that particular time. The values above the diagonal correspond to the weights assigned to future frames. The values below the diagonal correspond to the attention weights assigned to past frames. We only provide the results for the first 10 seconds for better resolution. In addition to the figures, we quantify the results in Table 3, which lists the average weights of the previous 14 frames, the current frame, and the future 14 frames. A higher value in a region indicates that we pay more attention to that region. Table 3 also shows their corresponding contribution in percentage.

Figures 10(a) and 10(b) show the attention weights for each modality for the self-attention and cross-attention methods, respectively. We observe that in the self-attention mechanism, the speech modality pays more attention to

Table 3: Average weight values for the self-attention and cross-attention based models. The table provides the results for the past, current, future frames. It also includes the contribution in percentage for each condition. The table shows the impact that noise in the features has on the weights. Higher values indicate higher attentions.

Approach	Modality	Past frames		Current frame		Future frames	
		Value	[%]	Value	[%]	Value	[%]
Self-attention	Speech	0.0006	0.22	0.0657	23.21	0.2168	76.56
	Text	0.0386	22.60	0.0627	36.68	0.0696	40.71
Cross-attention	Speech	0.0925	37.53	0.0514	20.85	0.1026	41.60
	Text	0.0470	24.31	0.0341	17.66	0.1121	58.01

future frames (76.56%), while the text modality distributes the attention to previous (22.6%), current (36.68%) and futures (40.71%) frames, indicating that it takes into account the full context. This behavior changes for the cross-attention mechanism, since the fusion involves a closer interaction between the modalities (i.e., query matrix is shared with the other modality). The text modality forces the speech modality to take into account more context, while the speech modality makes the text modality to consider the current frame less (17.66%), and pay more attention to the past (24.31%) and future (58.01%) frames, which is supported by the results in Table 3.

These results indicate that the difference between the multimodal models using self-attention and cross-attention is how the attention weights are distributed across time. By observing Table 3, we noticed that the multimodal model with self-attention as the fusion method has more contribution from the current and future frames for generating lip movements. Meanwhile, the multimodal model with cross-attention mechanism pays more attention to the past and future frames. Therefore, this distribution of the attention weights plays a key role in the generation of lip movements, creating the difference that we report in the perceptual evaluation of the naturalness of the videos.

6. Conclusions and Future Work

This paper presented two alternative methods based on the attention mechanism for synthesizing lip movements on a 3D shape by using only transcriptions provided by an ASR and speech. These methods follow a CGAN architecture implemented for the two modalities considered in the study: text and speech. The proposed methods relied on separate pretrained CGAN for each modality. The modality dependent representations are fused with the attention mechanism, which provides a powerful framework to quantify the temporal context used by each modality in the generation of lip movements. The first fusion approach relies on self-attention mechanisms separately implemented for each modality. The second fusion approach relies on cross-attention mechanism, where the query matrix is shared across modalities. The proposed methods were evaluated by creating animations using a formulation that requires only the facial landmarks generated by the methods, which are projected into a blendshape model. We used this animation method to evaluate our models using two perceptual evaluation tasks. The first task assessed the naturalness of the animations generated with the predicted facial landmarks produced by the proposed methods and competitive baseline approaches. We observed that the fusion of text and speech with the proposed attention mechanisms was important to achieve good performance. We even observe that the naturalness perception of animations created by our proposed models was higher than the naturalness perception of the animations created with the original facial landmarks. The second task compared two alternative frameworks, establishing preferences between them. The results for this task reaffirmed the results observed from the first task, demonstrating the superior performance of our proposed fusion based approach implemented with either self-attention or cross-attention. The results showed that our models generate better co-articulation movements than the baselines. An advantage of the attention-based fusion approaches is the insights that we obtained from the weights of the attentions. An analysis of the average attention weights in the self-attention model indicated that text and speech modalities emphasize different frames. While speech emphasizes future frames, the attention weights for text are better distributed among past, current, and future frames. For the cross-attention model, the weights for text and speech are more similarly distributed for past, current, and future frames.

As a future work, we are exploring models that are also constrained by the emotional content conveyed in the sentence. A straightforward approach is to introduce emotion as an extra constraint of our CGAN framework. Another important research direction is to identify objective metrics that correlate well with subjective measures using perceptual evaluations. We demonstrate that using the LipSync model (Chung and Zisserman, 2016) as an objective metric is an appealing approach to assess synthesized lip movements. This metric, unlike other approaches, correlates with the results from the perceptual evaluations. Since perceptual evaluations take time, defining good objective metrics can lead to more effective approaches to assess the impact of small changes in the model without having to conduct an additional subjective evaluation.

Acknowledgment

This work was funded by the National Science Foundation, under grant IIS:1718944

References

Ahuja, C., Lee, D., Nakano, Y., Morency, L., 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (Eds.), European Conference on Computer Vision (ECCV 2020). Springer Berlin Heidelberg, Glasgow, UK. volume 12363 of *Lecture Notes in Computer Science*, pp. 248–265. doi:10.1007/978-3-030-58523-5_15.

Ahuja, C., Ma, S., Morency, L., Sheikh, Y., 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations, in: ACM International Conference on Multimodal Interaction (ICMI 2019), Suzhou, China. pp. 74–84. doi:10.1145/3340555.3353725.

Bishop, C., 1995. Training with noise is equivalent to Tikhonov regularization. Neural Computation 7, 108–116. doi:10.1162/neco.1995.7.1.108.

Brand, M., 1999. Voice puppetry, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999), Los Angeles, CA, USA. pp. 21–28. doi:10.1145/311535.311537.

Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S., 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech and Language Processing 15, 1075–1086. doi:10.1109/TASL.2006.885910.

Busso, C., Deng, Z., Neumann, U., Narayanan, S., 2005. Natural head motion synthesis driven by acoustic prosodic features. Computer Animation and Virtual Worlds 16, 283–290. doi:10.1002/cav.80.

Busso, C., Narayanan, S., 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Transactions on Audio, Speech and Language Processing 15, 2331–2347. doi:10.1109/TASL.2007.905145.

Cao, C., Weng, Y., Lin, S., Zhou, K., 2013. 3D shape regression for real-time facial animation. ACM Transactions on Graphics 32, 41. doi:10.1145/2461912.2462012.

Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K., 2014. FaceWarehouse: A 3D facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics 20, 413–425. doi:10.1109/TVCG.2013.249.

Chai, Y., Weng, Y., Wang, L., Zhou, K., 2022. Speech-driven facial animation with spectral gathering and temporal attention. Frontiers of Computer Science 16, 1–10.

Chang, J., Scherer, S., 2017. Learning representations of emotional speech with deep convolutional generative adversarial networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA. pp. 2746–2750. doi:10.1109/ICASSP.2017.7952656.

Chen, M., He, X., Yang, J., Zhang, H., 2018a. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters 25, 1440–1444. doi:10.1109/LSP.2018.2860246.

Chen, X., Cao, C., Xue, Z., Chu, W., 2018b. Joint audio-video driven facial animation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada. pp. 3046–3050. doi:10.1109/ICASSP.2018.8461502.

Chung, J., Zisserman, A., 2016. Out of time: automated lip sync in the wild, in: Chen, C., Lu, J., Ma, K. (Eds.), Asian Conference on Computer Vision (ACCV 2016 Workshop). Springer Berlin Heidelberg, Taipei, Taiwan. volume 10117 of *Lecture Notes in Computer Science*, pp. 251–263. doi:10.1007/978-3-319-54427-4_19.

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J., 2019. Capture, learning, and synthesis of 3d speaking styles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10101–10111.

Deng, Z., Chiang, P., Fox, P., Neumann, U., 2006a. Animating blendshape faces by cross-mapping motion capture data, in: ACM symposium on Interactive 3D graphics and games (I3D 2006), Redwood City, CA, USA. pp. 43–48. doi:10.1145/1111411.1111419.

Deng, Z., Neumann, U., Lewis, J., Kim, T., Bulut, M., Narayanan, S., 2006b. Expressive facial animation synthesis by learning speech co-articultion and expression spaces. IEEE Transactions on Visualization and Computer Graphics (TVCG) 12, 1523–1534.

Edwards, P., Landreth, C., Fiume, E., Singh, K., 2016. JALI: An animator-centric viseme model for expressive lip synchronization. ACM Transactions on Graphics 35, 127. doi:10.1145/2897824.2925984.

Eskimez, S.E., Maddox, R., Xu, C., Duan, Z., 2018. Generating talking face landmarks from speech, in: Deville, Y., Gannot, S., Mason, R., Plumbley, M., Ward, D. (Eds.), Latent Variable Analysis and Signal Separation (LVA/ICA 2018). Springer, Cham, Guildford, UK. volume 10891 of *Lecture Notes in Computer Science*, pp. 372–381. doi:10.1007/978-3-319-93764-9_35.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7, 190–202. doi:10.1109/TAFFC.2015.2457417.

Fan, B., Xie, L., Yang, S., Wang, L., Soong, F.K., 2016. A deep bidirectional LSTM approach for video-realistic talking head. Multimedia Tools and Applications 75, 5287–5309. doi:10.1007/s11042-015-2944-3.

Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T., 2022. Faceformer: Speech-driven 3d facial animation with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18770–18780.

Ferstl, Y., McDonnell, R., 2018. Investigating the use of recurrent motion modelling for speech gesture generation, in: Intelligent Virtual Agents (IVA 2018), Sydney, NSW, Australia. pp. 93–98. doi:10.1145/3267851.3267898.

Ferstl, Y., Neff, M., McDonnell, R., 2019. Multi-objective adversarial gesture generation, in: Motion, Interaction and Games (MTG 2019), Newcastle upon Tyne, UK. pp. 1–10. doi:10.1145/3359566.3360053.

Ferstl, Y., Neff, M., McDonnell, R., 2020. Adversarial gesture generation with realistic gesture phasing. Computers & Graphics 89, 117–130. doi:10.1016/j.cag.2020.04.007.

Goncalves, L., Busso, C., 2022. AuxFormer: Robust approach to audiovisual emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore. pp. 7357–7361. doi:10.1109/ICASSP43922.2022.9747157.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems (NIPS 2014), Montreal, Canada. pp. 2672–2680.

Gupta, A., Jaiswal, R., Adhikari, S., Balasubramanian, V., 2016. DAISEE: dataset for affective states in e-learning environments. CoRR abs/1609.01885. URL: `http://arxiv.org/abs/1609.01885`, `arXiv:1609.01885`.

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK.

Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W., Rätsch, M., Kittler, J., 2016. A multiresolution 3D morphable face model and fitting framework, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (), Rome, Italy. pp. 1–8. doi:10.5220/0005669500790086.

Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J., 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) 36, 94. doi:10.1145/3072959.3073658.

Kaur, A., Mustafa, A., Mehta, L., Dhall, A., 2018. Prediction and localization of student engagement in the wild, in: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. doi:10.1109/DICTA.2018.8615851.

King, D., 2009. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research 10, 1755–1758.

Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations, San Diego, CA, USA. pp. 1–13.

Kipp, M., Neff, M., Kipp, K., Albrecht, I., 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis, in: Pelachaud, C., Martin, J., André, E., Chollet, G., Karpouzis, K., Pelé, D. (Eds.), International Workshop on Intelligent Virtual Agents (IVA2007). Springer Berlin Heidelberg, Paris, France. volume 4722 of *Lecture Notes in Computer Science*, pp. 15–28. doi:10.1007/978-3-540-74997-4_2.

Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G., Kjellström, H., 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. International Journal of Human-Computer Interaction 37, 1300–1316. doi:10.1080/10447318.2021.1883883.

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G., Alexandersson, S., Leite, I., Kjellström, H., 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation, in: ACM International Conference on Multimodal Interaction (ICMI 2020), Utrecht, The Netherlands. pp. 242–250. doi:10.1145/3382507.3418815.

Lee, Y., Terzopoulos, D., Waters, K., 1995. Realistic modeling for facial animation, in: Conference on Computer graphics and interactive techniques (SIGGRAPH 1995), Los Angeles, CA, USA. pp. 55–62. doi:10.1145/218380.218407.

Liu, N., Zhou, T., Ji, Y., Zhao, Z., Wan, L., 2020. Synthesizing talking faces from text and audio: An autoencoder and sequence-to-sequence convolutional neural network. Pattern Recognition 102, 107231. doi:10.1016/j.patcog.2020.107231.

Luo, C., Yu, J., Li, X., Wang, Z., 2014. Realtime speech-driven facial animation using Gaussian mixture models, in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2014), Chengdu, China. pp. 1–6. doi:10.1109/ICMEW.2014.6890554.

Mariooryad, S., Busso, C., 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. IEEE Transactions on Audio, Speech and Language Processing 20, 2329–2340. doi:10.1109/TASL.2012.2201476.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M., 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi, in: Interspeech 2017, Stockholm, Sweden. pp. 498–502. doi:10.21437/Interspeech.2017-1386.

McFee, B., Lostanlen, V., McVicar, M., Metsai, A., Balke, S., Thomé, C., Raffel, C., Lee, D., Lee, K., Nieto, O., et al., 2019. librosa/librosa: 0.7.0. doi:10.5281/zenodo.3270922.

Minnis, S., Breen, A., 2000. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis, in: International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China. pp. 759–762.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. ArXiv e-prints (arXiv:1411.1784) `arXiv:1411.1784`.

Mower Provost, E., Shangguan, Y., Busso, C., 2015. UMEME: University of Michigan emotional McGurk effect data set. IEEE Transactions on Affective Computing 6, 395–409. doi:10.1109/TAFFC.2015.2407898.

Neff, M., Kipp, M., Albrecht, I., Seidel, H., 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) 27, 1–24. doi:10.1145/1330511.1330516.

Ostermann, J., 1998. Animation of synthetic faces in MPEG-4, in: Proceedings Computer Animation, Philadelphia, PA, USA. pp. 49–55. doi:10.1109/CA.1998.681907.

Pelachaud, C., Busso, C., Heylen, D., 2021. Multimodal behavior modeling for socially interactive agents, in: Lugrin, B., Pelachaud, C., Traum, D.

(Eds.), The Handbook of Socially Interactive Agents: 20 Years of Research on Intelligent Virtual Agents, Embodied Conversational Agents, and Social Robotics. Association for Computing Machinery, Human-Centered Computing, New York, NY, USA, pp. 259–310. doi:10.1145/3477322.3477331.

Pham, H., Wang, Y., Pavlovic, V., 2018. End-to-end learning for 3D facial animation from speech, in: ACM International Conference on Multimodal Interaction (ICMI 2018), Boulder, CO, USA. pp. 361–365. doi:10.1145/3242969.3243017.

Pham, H.X., Cheung, S., Pavlovic, V., 2017. Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017), Honolulu, HI, USA. pp. 2328–2336. doi:10.1109/CVPRW.2017.287.

Qiu, Y., Misu, T., Busso, C., 2019. Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data, in: ACM International Conference on Multimodal Interaction (ICMI 2019), Suzhou, Jiangsu, China. pp. 164–173. doi:10.1145/3340555.3353749.

Qiu, Y., Misu, T., Busso, C., 2022. Unsupervised scalable multimodal driving anomaly detection. IEEE Transactions on Intelligent Vehicles to appear. doi:10.1109/TIV.2022.3160861.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis, in: International Conference on Machine Learning (ICML 2016), San Juan, Puerto Rico. pp. 1–10.

Richard, A., Lea, C., Ma, S., Gall, J., de la Torre, F., Sheikh, Y., 2021a. Audio- and gaze-driven facial animation of codec avatars, in: IEEE Winter Conference on Applications of Computer Vision (WACV 2021), Waikoloa, HI, USA. pp. 41–50. doi:10.1109/WACV48630.2021.00009.

Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y., 2021b. Meshtalk: 3d face animation from speech using cross-modality disentanglement, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1173–1182.

Rizzo, A., Neumann, U., Enciso, R., Fidaleo, D., Noh, J., 2004. Performance-driven facial animation: Basic research on human judgments of emotional state in facial avatars. CyberPsychology & Behavior 4, 471–487. doi:10.1089/109493101750527033.

Sadoughi, N., Busso, C., 2015. Retrieving target gestures toward speech driven animation with meaningful behaviors, in: International conference on Multimodal interaction (ICMI 2015), Seattle, WA, USA. pp. 115–122. doi:10.1145/2818346.2820750.

Sadoughi, N., Busso, C., 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory, in: Beskow, J., Peters, C., Castellano, G., O'Sullivan, C., Leite, I., Kopp, S. (Eds.), International Conference on Intelligent Virtual Agents (IVA 2017). Springer Berlin Heidelberg, Stockholm, Sweden. volume 10498 of *Lecture Notes in Computer Science*, pp. 389–402. doi:10.1007/978-3-319-67401-8_49.

Sadoughi, N., Busso, C., 2018a. Expressive speech-driven lip movements with multitask learning, in: IEEE Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China. pp. 409–415. doi:10.1109/FG.2018.00066.

Sadoughi, N., Busso, C., 2018b. Novel realizations of speech-driven head movements with generative adversarial networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada. pp. 6169–6173. doi:10.1109/ICASSP.2018.8461967.

Sadoughi, N., Busso, C., 2019. Speech-driven animation with meaningful behaviors. Speech Communication 110, 90–100. doi:10.1016/j.specom.2019.04.005.

Sadoughi, N., Busso, C., 2021. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. IEEE Transactions on Affective Computing 12, 1031–1044. doi:10.1109/TAFFC.2019.2916031.

Sadoughi, N., Liu, Y., Busso, C., 2017. Meaningful head movements driven by emotional synthetic speech. Speech Communication 95, 87–99. doi:10.1016/j.specom.2017.07.004.

Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. HMM-based text-to-audio-visual speech synthesis, in: International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China. pp. 25–28.

Stef, A., Perera, K., Shum, H.P.H., Ho, E., 2018. Synthesizing expressive facial and speech animation by text-to-ipa translation with emotion control, in: International Conference on Software, Knowledge, Information Management & Applications (SKIMA 2018), Phnom Penh, Cambodia. pp. 1–8. doi:10.1109/SKIMA.2018.8631536.

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C., 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics (TOG) 23, 506–513.

Suwajanakorn, S., Seitz, S., Kemelmacher-Shlizerman, I., 2017. Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) 36, 95:1–13. doi:10.1145/3072959.3073640.

Tang, H., Fu, Y., Tu, J., Hasegawa-Johnson, M., Huang, T.S., 2008. Humanoid audio-visual avatar with emotive text-to-speech synthesis. IEEE Transactions on Multimedia 10, 969–981. doi:10.1109/TMM.2008.2001355.

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A., Hodgins, J., Matthews, I., 2017. A deep learning approach for generalized speech animation. ACM Transactions on Graphics (TOG) 36. doi:10.1145/3072959.3073699.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M., Schuller, B., Zafeiriou, S., 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China. pp. 5200–5204. doi:10.1109/ICASSP.2016.7472669.

Tsai, Y.H., Bai, S., Liang, P., Kolter, J., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences, in: Association for Computational Linguistics (ACL 2019), Florence, Italy. pp. 6558–6569. doi:10.18653/v1/p19-1656.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: In Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. pp. 5998–6008.

Vidal, A., Salman, A., Lin, W.C., Busso, C., 2020. MSP-Face corpus: A natural audiovisual emotional database, in: ACM International Conference on Multimodal Interaction (ICMI 2020), Utrecht, The Netherlands. pp. 397–405. doi:10.1145/3382507.3418872.

Williams, L., 1990. Performance-driven facial animation. Computer Graphics 24, 235–242. doi:10.1145/1185657.1185856.

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., Henter, G.E., 2022. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation, in: Proceedings of the 2022 International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA. p. 736–747. URL: https://doi.org/10.1145/3536221.3558058, doi:10.1145/3536221.3558058.

Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X., 2018a. Talking face generation by adversarially disentangled audio-visual representation. ArXiv e-prints (arXiv:1807.07860) , 1–10arXiv:1807.07860.

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K., 2018b. Visemenet: Audio-driven animator-centric speech animation. ACM Transactions on Graphics 37, 161. doi:10.1145/3197517.3201292.