Speech-Driven Expressive Talking Lips with Conditional Sequential Generative Adversarial Networks

Najmeh Sadoughi , *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

Abstract—Articulation, emotion, and personality play strong roles in the orofacial movements. To improve the naturalness and expressiveness of *virtual agents* (VAs), it is important that we carefully model the complex interplay between these factors. This paper proposes a conditional generative adversarial network, called *conditional sequential GAN* (CSG), which learns the relationship between emotion, lexical content and lip movements in a principled manner. This model uses a set of spectral and emotional speech features directly extracted from the speech signal as conditioning inputs, generating realistic movements. A key feature of the approach is that it is a speech-driven framework that does not require transcripts. Our experiments show the superiority of this model over three state-of-the-art baselines in terms of objective and subjective evaluations. When the target emotion is known, we propose to create emotionally dependent models by either adapting the base model with the target emotional data (CSG-Emo-Adapted), or adding emotional conditions as the input of the model (CSG-Emo-Aware). Objective evaluations of these models show improvements for the CSG-Emo-Adapted compared with the CSG model, as the trajectory sequences are closer to the original sequences. Subjective evaluations show significantly better results for this model compared with the CSG model when the target emotion is happiness.

Index Terms—Speech-driven model, lip movements, expressive and naturalistic lip movements, generative adversarial network.

1 Introduction

The lower part of the face, which we refer to as the orofacial area (see Fig. 1), plays an important role in conveying lexical, emotional and idiosyncratic information. These factors are integrated in a nontrivial manner, facilitating face to face communications. It is important to generate proper facial movements for *virtual agents* (VAs) to communicate a message more effectively and more naturally. Although emotion is expressed throughout the whole face, there are emotional states such as happiness for which the orofacial area plays a big role (e.g., smile). For these emotions, in particular, careful modeling of the relationship between emotion and articulation is required to have more natural and expressive VAs.

Several factors contribute to the variation in the orofacial area. The orofacial muscles are activated by the articulatory movements imposed through the vocal region. The relation between lip motion and phonetic content is colored by the emotional cues expressed in the message. This coupling between lexical and emotional contents is also affected by idiosyncratic characteristics across people. The integration between these factors in the orofacial area is complex [1, 2]. Most of the previous studies on lip movement synthesis have relied on the recordings from one subject in order to avoid speaker variations [3, 4, 5]. Since multimodal emotional corpora usually include multiple speakers with limited data per subject [6], it is important that the models can effectively capture speaker variability. If these variations are not carefully considered, the model may predict trajectories that average these variations, creating over-smoothed movements. Furthermore, most of the previous models for lip movements rely on transcriptions (e.g., phonemes or triphonemes) [7, 8, 9], or transcriptions plus the target emotional categories [5, 10, 11]. The need for transcriptions limits the domain of applications. We envision a data-driven lip generation framework that does not require transcription, and can effectively capture the temporal relations between speech, lip movement and emotion.

Speech conveys verbal and nonverbal cues, having a direct influence in the visual appearance of the orofacial area. For example, speech is one of the primary channels to convey emotions [12]. Therefore, relying on speech for modeling the nonverbal behaviors in the orofacial area can help the model to capture the fine expressive movements shown during natural interactions. Our envisioned framework relies on speech features to generate lip motion. From an application perspective, having a system that synthesizes lip motion only from speech is very appealing. It brings flexibility to the system that otherwise is not possible without adding additional modules. From a theoretical perspective, the problem is also appealing, exploring and learning directly the temporal relationship between speech features and lip motion (e.g., lip appearance).

This paper proposes to use a *conditional generative adversarial network* (cGAN), composed of *long short-term memory* (LSTM) for generating realistic and expressive lip movements. The approach is called *conditional sequential generative adversarial networks* (CSG). The model learns the distribution of the orofacial movements conditioned on speech features (lips are represented in terms of the X, Y and Z positions of

• *N. Sadoughi and C. Busso are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, Richardson TX 75080.*
E-mail: najme.sadoughi@gmail.com, busso@utdallas.edu

Manuscript received May 15, 2018; revised December 16, 2018; accepted May 7, 2019.

motion capture markers around the mouth – see Fig. 1). The training of the models consists of an adversarial objective that combines a generator and a discriminator, such that it generates more convincing lip movements. A key feature of the adversarial training is to teach the model to capture the temporal relationship between acoustic features and lip motion. This objective is achieved by creating fake sequences with mismatched speech and lip motion trajectories that the discriminator has to recognize. The resulting lip motion sequences capture the temporal coupling between speech and lip movements, creating realistic sequences, which convey the underlying lexical content. We compare the CSG model with three baselines proposed in previous studies [3, 13, 14], which use conventional non-adversarial methods. The experimental evaluations with objective and subjective metrics demonstrate that the proposed CSG model achieves better performance than these methods.

Another appealing property of the CSG framework is that it can be easily extended to consider emotions. Experiencing emotions includes *felt emotion*, *expressed emotion* and *perceived emotion*, which are not necessarily the same. Felt emotion corresponds to the true emotion experience by an individual. The expressed emotion corresponds to the emotions externalized by the individual. The perceived emotion corresponds to the emotion perceived by others. Our goal is to generate emotional behaviors that convey the intended emotion. Therefore, we want to create sequences that are successful in affecting the perceived emotion of the animation. To explicitly incorporate emotions, it is more practical for a VA to represent emotion with discrete categories such as anger, happiness, fear, surprise, sadness, and disgust (e.g., basic emotions [15]). It is easier to specify that a CA is "happy" than to specify that its "arousal level is 0.4" (core emotion theory) or that its "AU6 has intensity 4" (*facial action coding system* (FACS)). Therefore, this study describes emotion using categorical emotions. We build two expression-dependent models: (1) by adapting the CSG model to different emotions, called CSG-Emo-Adapted model, and (2) by conditioning the CSG model on categorical emotion of the speaker, called CSG-Emo-Aware model. Objective and subjective evaluations show that the CSG-Emo-Adapted model generates better expressions compared to the CSG model, when the target emotion is happiness. The results validate our proposed method, which can generate more convincing and expressive orofacial movements.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the resources used in this study including the corpus, the extracted audiovisual features, and the rendering toolkit. Section 4 describes the *conditional sequential GAN* (CSG) method, and its two expression-dependent extensions: the CSG-Emo-Adapted and CSG-Emo-Aware models. Section 5 describes the experimental evaluation. Section 6 presents the results of the models, comparing the CSG methods with the baselines. Section 7 summarizes the contributions of this work highlighting the advantages and disadvantages of the approach, and possible future directions.

2 Related Work

This section summarizes previous studies on generating lip motion. We group these studies into three categories: canonical shape selection, *hidden Markov model* (HMM)-based approaches, and *deep neural network* (DNN)-based approaches.

2.1 Canonical Shape Selection

These methods consist of selecting canonical shapes, which are blended for each predefined unit of articulation. For example, canonical shapes can be blended with appropriate weights to represent different phonetic units. Xu et al. [8] considered several canonical shapes for lips, where the weights for phonemes and bigrams were carefully defined by artists. While blending the shapes, the approaches need to model coarticulation between phonemes, which is an important task for realistic lip motion generation. Coarticulation is the phenomenon of adjusting the articulation according to the adjacent phonemes, capturing their dependencies. Deng et al. [9] modeled coarticulation between phonemes, relying on real recordings of human data. They found the weights for the linear combinations of the canonical shapes for diphones and triphones, minimizing the error between the predicted and the original movements. Cao et al. [10] developed a framework to generate expressive facial movements. Their framework used tuples containing phoneme, emotion, prosody, and lip trajectories. During testing, the input was parsed with the phonetic content, and the target or predicted emotion. The database was searched with the sequence of tuples derived from the input, while imposing correct co-articulation and smooth constraints. The selected segments were aligned with the input using time-warping. Finally, the motion segments were blended and smoothed to create facial movements.

Unit selection methods require emotion-dependent speech units to account for expressive lip motions, where the weights have to be redefined for each of the target emotion.

2.2 HMM-based Modeling

HMM-based models learn to synthesize lip movements from text or speech by implicitly modeling the underlying co-articulations. Choi et al. [16] proposed to use HMM inversion for audio to visual conversion. They used a three-state HMM to model each phoneme. During testing, they relied on the Baum-Welch algorithm to find the maximum likelihood estimates of the visual features. Xie and Liu [17] proposed to use *coupled HMMs* (CHMMs) to model the dependencies as well as the differences between the audio and visual modalities (e.g., their asynchrony and different number of phonetic units). Anderson et al. [5] designed a system to create emotional facial movements using *cluster adaptive training* (CAT), which was built upon HMMs for text-to-speech systems. Their HMM modeled quinphones created with five states, where a decision tree was used to handle the sparseness of the quinphones in the data. The decision tree is also used to find the mean and variances of the Gaussian distributions for the quinphone. The proposed CAT framework captured emotion dependent quinphones by finding emotion-dependent linear combinations between clusters.

2.3 DNN-based Modeling

DNN-based models directly learn how to predict the movements from speech features. Taylor et al. [3] proposed a fully connected feedforward neural network for audio to visual conversion. Their network gets the speech features over a specified contextual window, predicting current and future orofacial movements. The approach used sliding windows with step size of one frame where the average of the predictions for each window is considered as the target value for the center of the window. Their model outperformed the HMM inversion approach proposed by Choi et al. [16]. Fan et al. [14] explored the use of deep learning structures built with *bidirectional long short-term memory* (BLSTM) to synthesize head and face movements driven by transcriptions, speech, and transcriptions plus speech. The inputs of the system correspond to triphoneme labels from transcriptions, and/or *mel frequency cepstral coefficients* (MFCCs) and their first and second order derivatives from speech. The study compared the results achieved with their model with a HMM-based approach, showing improvements in terms of objective and subjective metrics. Previous studies have also modeled the relationship between speech and facial movements with LSTMs, optimizing the L2 norm of the error between the predictions and ground truth. To model co-articulation, these studies use future frames ([18] [19]), or context features consisting of more than one frame [19]. Li et al. [6] proposed strategies using BLSTM models to create emotional facial movement by having access to a small emotional dataset. They proposed several approaches to leverage recordings from a neutral corpus and a small emotional corpus, aiming to improve the emotional regression result. Their best result was achieved with a cascade framework, where the predictions obtained with the neutral corpus were concatenated with audio features and used as features of a second system. The second system is trained with the emotional corpus.

Karras et al. [4] proposed a framework with *convolutional neural networks* (CNNs) to predict facial movements from raw speech signal. Their framework disentangled the facial configurations explained by audio features and emotional states. This goal is achieved by considering a dedicated emotional state learned for each training sentence. Their framework predicted the facial pose one frame at a time, utilizing a contextual window of 260ms with previous and future frames. They trained separate models per speaker with three to five minutes of synchronized audiovisual data. They compared their method with the results achieved by the faceFX software [20] using subjective evaluations showing higher preferences for their models. Parker et al. [11] proposed an approach for generating emotional audiovisual content from transcriptions and target emotion. They proposed to share the layers of the network across all the emotions, with the exception of the last layer, which was adapted for each emotion using regularized least squares. They compared their results with a HMM system, showing improvements when using their method.

2.4 Contributions

This paper proposes to use a conditional GAN structure, called CSG, composed of BLSTM units to learn the distribution of orofacial movements. The proposed CSG framework relies on adversarial training by jointly training a generator and a discriminator. During the adversarial training, a discriminator learns to recognize two sets of fake samples: the samples generated by the generator, and samples from uncoupled recordings from the original database where the audio does not match the lip motion sequence. While the first type of fake samples guides the generator to create lip motions with the correct distribution, the second type of fake samples provides more diverse training examples for the discriminator to explicitly learn the temporal relation between speech and lip motions. When the discriminator learns these errors, it helps the generator to improve the coupling between speech and lip motion. As the generator learns to create realistic sequences to fool the discriminator, our method generates realistic samples which are timely coupled with the audio. To the best of our knowledge, this is the first time that adversarial training is used to synthesize lip motion. By using conditional GAN, we effectively model the relationship between emotion, speech and orofacial movements. This approach does not require transcriptions as most previous studies, opening opportunities for cases where transcriptions are not available (e.g., real time applications, spontaneous dialogs, virtual meetings, and visual displays for hearing impaired individuals to enhance phone conversations). This framework departs from deep learning approaches used by previous studies, providing a systematic strategy to generate emotional lip sequences.

3 Resources

3.1 The IEMOCAP Corpus

This study uses the IEMOCAP corpus [21]. This database comprises video, audio, and motion capture recordings from 10 actors in improvised and script-based scenarios. The scenarios were designed such that they elicited different emotions from the actors. We use the data from all the subjects, where 60% of the data is used for training, 20% for validation and 20% for testing. The database is annotated with categorical emotions by three annotators at the speaking turn level. They annotated the emotional content using ten classes: neutral state, anger, happiness, sadness, surprise, fear, frustration, excitement, disgust, and other. Similar to previous studies relying on the IEMOCAP corpus, we merge the turns labeled with excitement and happiness [22, 23]. We calculate the consensus labels for each turn by estimating the majority vote across the annotations. This approach creates hard emotional classes for each sentence. While the IEMOCAP corpus was annotated with ten classes, the data recording protocol targeted scripts and improvisation scenarios to elicit only happiness, anger, sadness, and frustration. As a result, there are few turns labeled as surprise, fear, disgust or excited. The frequencies of emotional categories for the consensus labels are 605 (neutral state), 621 (anger), 882 (happiness), 653 (sadness), 1 (disgust), 20 (fear), 998 (frustration), 31 (surprise), and 2 (other). The evaluators do not reach agreement in 1,228 segments. Due to the sparsity of the classes disgust, fear, and surprise, we merge all these segments with the speaking turns without consensus, assigning them to the class other. Consensus labels such as majority vote discard information provided by

(a) Markers - IEMOCAP corpus (b) Markers used by Xface

Fig. 1. The location of the 15 markers from the IEMOCAP corpus considered in this study. These markers are used to render the animations using Xface.

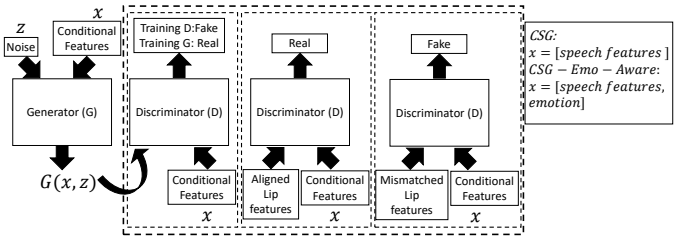


Fig. 2. Proposed CSG model. Figure 4 provides more detailed diagrams.

individual evaluations (see study by Lotfian and Busso [24]). Therefore, we also rely on soft assignments by considering the individual annotations (three annotations per turn). The frequencies of the emotional classes assigned to the turns when we consider individual annotations are: 2,538 (neutral state), 2,108 (anger), 3,795 (happiness), 2,047 (sadness), 55 (disgust), 138 (fear), 3,961 (frustration), 200 (surprise) and 281 (other). For consistency, we restrict the analysis to the six classes neutral state, anger, happiness, sadness, frustration, and other. The soft labels are created by estimating the distribution of the labels assigned to the speaking turn. For example, if there are two annotations for anger and one for frustration, we consider a 6D vector with 0.66 for anger, 0.33 for frustration and 0.0 for the remaining categories.

3.2 Audiovisual Features

We extract two sets of features from the audio. The first set of features are 25 MFCCs extracted with Praat [25] over 25ms windows every 8.33 ms. We choose 25 MFCCs, because Taylor et al. [3] evaluated their models for predicting lip movements with a different number of MFCCs, finding that 25 MFCCs gives the best result. By moving the analysis window in increments of 8.33ms, we create 120 feature vectors per second, matching the sampling rate of the motion capture recordings. We also extract the fundamental frequency and intensity with Praat using the same window and step size. Moreover, we extract 17 additional *low level descriptors* (LLDs) from the *extended Geneva minimalistic acoustic parameter set* (eGeMAPS) [26], which is a feature set carefully selected for paralinguistic tasks. The eGeMAPS features are extracted with OpenSmile [27]. The fundamental frequency, intensity and eGeMAPS set are collectedly referred to as *emotional speech features*, and the 25 MFCCs are referred to as *spectral features*.

From the motion capture recordings, we use the (X, Y, Z) locations of 15 markers around the mouth area (Fig. 1). The sampling rate is 120 fps. Busso et al. [21] describes the steps to derive the motion capture data.

3.3 Xface

We rely on Xface [28] for rendering the VA. Xface uses *facial action parameters* (FAPs) to animate the face. FAPs are directly tied to facial *action units* (AUs) in FACS [29], making them a suitable representation for emotional facial movements. FAPs control *facial points* (FPs) on the face which are based on the MPEG-4 standard. Most of the facial markers in the IEMOCAP corpus follow the locations of FPs in the MPEG-4 standard (Fig. 1(b)), so it is possible to linearly map the markers and FAPs. This mapping is achieved by mapping a neutral pose of the recording of each subject as the reference pose. Then, we map the range of movements for each marker to the range of movements allowed by the FAPs in Xface. More details about this mapping is provided by Mariooryad and Busso [30]. This study uses a female character for all the subjective evaluations.

While there are other sophisticated rendering toolkits, Xface allows us to easily animate the motion capture data in our corpus. As a result, we can directly focus on the modeling part of the lip motion generation, which is the focus of this study.

4 Methodology

The study proposes to generate lip motion driven by speech using adversarial training. The proposed framework corresponds to a conditional GAN for generating orofacial movements from audio features. Figure 2 shows the overall framework for this model, which is called *conditional sequential GAN* (CSG). The figure demonstrates how the generator and the discriminator are trained using the real and fake samples. The discriminator is trained to distinguish between the real and fake samples, where the real samples are the lip sequences aligned with the input audio, and the fake samples are either the lip sequences synthesized by the generator or the real lip samples which are not aligned with the input audio (i.e. mismatched). The generator is trained to fool the discriminator (i.e., the target label is real). Two strengths of the approach are that (1) it does not require any lexical label, since it directly learns the mapping between speech and lip motion, and (2) it can be adapted to synthesize expressive behaviors when the intended emotion is provided as input. This section presents our proposed speech-driven framework for lip synthesis, describing the required building blocks and their roles in solving this problem.

4.1 Bidirectional Long Short-Term Memory (BLSTM)

Incorporating future frames as well as the past frames can help the model to make better predictions. Therefore, our models are built with *bidirectional LSTMs* (BLSTMs). These models consist of forward and backward paths of LSTMs, duplicating the number of hidden states (Fig. 3). While this model can be used in real time using a short delay, we assume that we have the entire sequence of audio features, generating the sequences offline.

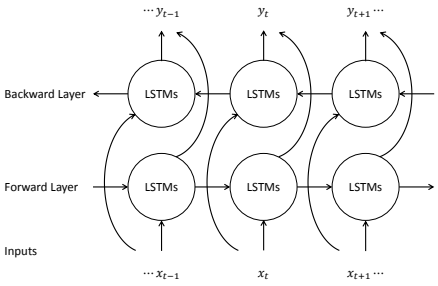


Fig. 3. Illustration of BLSTM composed of forward and backward layers. The layer takes input $\mathbf{x}_t$ creating output $\mathbf{y}_t$

4.2 Generative Adversarial Network (GAN)

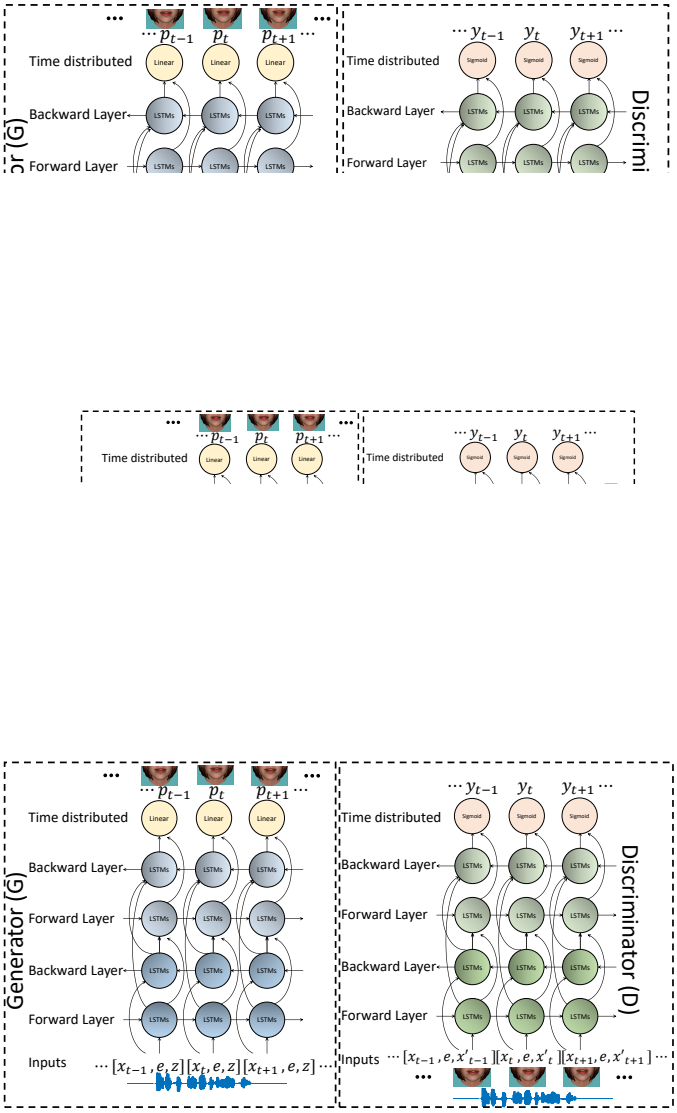
GANs are proposed as a generative model for learning the distribution of the data [31]. The training in GAN is a min-max game between two players, a generator (G) and a discriminator (D). The role of the discriminator is to distinguish between the samples generated by the generator (fake samples labeled as 0), and the samples from the original data (real samples labeled as 1). The role of the generator is to create samples given the input noise ($\mathbf{z}$), which resemble the real data, fooling the discriminator. This game can be achieved by the loss function in Equation 1, where $\mathcal{L}$ is the loss function, E represents the expected value, $\mathbf{x}$ represents the real samples, $\mathbf{z}$ represents the input noise to the generator, $D(.)$ represents the discriminator function, and $G(.)$ represents the generator function.

$$\begin{aligned}\min_G \max_D \mathcal{L}(D,G) =& E_{x\sim p_{data}(x)}\left[\log D(x)\right] \\ &+ E_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right]\end{aligned} \tag{1}$$

4.3 Conditional Sequential GAN (CSG)

Our proposed model is different from a simple GAN [31]. We aim to drive the lip motion with acoustic features. Therefore, we propose to use a conditional GAN model, where the constraints to the discriminator and the generator are acoustic features (see Fig. 2). The input to our model is composed of a window of speech features (i.e., x_i, where i is the frame number) plus a random noise (i.e., z) tied across the frames. The model maps the distribution of noise conditioned on the time-varying speech features to the distribution of original lip movements conditioned on speech features. We call our model *conditional sequential GAN* (CSG), which is shown in Figure 4(a). Previous studies have proposed different sequential GAN models to capture dynamics in videos [32, 33]. However, previous conditional sequential GANs are implemented with static conditions tied across the input sequence [33]. A key feature of our CSG model is that the input variable that conditions the GAN models is a time-varying signal (i.e., speech features).

Since we aim to learn the relationship between time-continuous signals (i.e., speech and lip movements), we build our cGANs with two layers of BLSTMs, where each "forward layer" and "backward layer" constitutes one BLSTM layer in Figure 4(a). We consider a linear output layer tied across all frames for the generator. We consider



(c) CSG-Emo-Aware

Fig. 4. The proposed frameworks to generate expressive lip motion sequences driven from speech, where t represents the time frame index, $\mathbf{x}$ represents speech features, $\mathbf{p}$ represents the output of the generator, $\mathbf{x}'$ represents the orofacial pose, $\mathbf{y}$ represents the output of the discriminator, and $\mathbf{e}$ represents the vector with soft emotional labels.

a sigmoid layer tied across all frames for the discriminator, as the output layer (variables $[\ldots,\mathbf{y}_{t-1},\mathbf{y}_t,\mathbf{y}_{t+1},\ldots]$ in Fig. 4(a)). We condition the generator and discriminator on the input features extracted from speech (variables $[\ldots,\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1},\ldots]$ in Fig. 4(a)).

Note that the errors happening in the generated lip movements can be of two types: 1) the lip movements may not look realistic, 2) the lip movements may not correlate well with the input speech signal. To address this observation and inspired by the matching-aware discriminator training strategy proposed by Reed et al. [34] for text-to-image synthesis, our learning strategy includes two kinds of fake samples during the training of the discriminator: samples generated by the generator, and original samples with lip motion and speech features extracted from different recordings. The first type of fake samples forces the generator to create realistic lip movements by decreasing the

difference between the synthesized and actual lip trajectories. This approach is common in GAN. The second type of fake sequences forces the generator to explicitly capture the temporal relationship between lip motion and speech. It provides more diverse training examples for the discriminator to learn cases where speech is not properly coupled with lip motion. With the adversarial training, the discriminator teaches the generator to avoid these mistakes. Notice that for the second type of fake sequences, it is unlikely that the phonetic content in the speech and lip movements are the same, since we are randomly combining speech segments with lip motion sequences. In the IEMOCAP corpus, the probability for randomly mismatched sequences of having the same lexical content is around 3%. By jointly learning these two types of fake samples, the discriminator helps the generator to create lip motion sequences which are not only realistic, but also strongly coupled with the audio features. Although a conditional GAN model should theoretically learn these two types of fake samples by itself using only samples from the generator (i.e., type one), using fake samples with uncoupled audio and lip motion emphasizes the importance of the temporal relationship between the modalities, which expedites the learning process (i.e., type two).

Equation 2 shows how our loss function is used to train the discriminator and the generator in the proposed CSG model. The vector x is the speech segment, x' is the lip trajectory segment, $p_{data(x,x')}$ represents the distribution of the aligned audio and lip trajectories, $p_{data(\hat{x})}$ represents the distribution of the audio distribution, $\hat{x}$ represents the misaligned audio segment, and z represents the noise distribution. This formulation is different from regular GAN (Equation 1), since (1) this is a conditional GAN where we add the variable x, and (2) the optimization process considers two sets of fake samples (second and third terms in Equation 2) as opposed to one set of fake samples as done in GAN (second component in Equation 1).

$$\begin{aligned}\min_G \max_D \mathcal{L}(D,G) = & E_{(x,x')\sim p_{data}(x,x')}\left[\log D(x,x')\right] + \\ & E_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(x,z\right)\right)\right)\right] + \\ & E_{\hat{x}\sim p_{data}(\hat{x})}\left[\log\left(1-D(\hat{x},x')\right)\right]\end{aligned} \quad (2)$$

The proposed CSG model needs to generate smooth trajectories for lip movements. Therefore, we use the same noise, $\mathbf{z}$, across all the input frames. It represents the global variations of conditional lip movements. To capture the dynamics of the movements, the CSG relies on the time-varying speech features provided across the frames as evidence for the dynamics of the orofacial movements, which is captured by the LSTM units. The success of the sequential generator depends on two factors: each orofacial configuration generated at each frame needs to look realistic with respect to the speech features, and the sequence generated by the generator needs to have realistic dynamics [33]. Therefore, we use fake/real labels not only on the final frame, but also on all the intermediate frames from the discriminator. This approach allows us to minimize the loss function not only on the final frame of the sequence, but also on all the intermediate frames. Figure 4(a) highlights that the discriminator considers the outputs across all the frames of the sequence. Our preliminary experiments demonstrated that this approach expedites learning. Note that we train our model by considering a fixed window length for both the generator and the discriminator.

4.4 Expression-Dependent CSG Models

The last building block in our proposed approach is to constrain the models on the target categorical emotion intended for the sentence, which is assumed to be an input for the model. We propose two expression-dependent models as extensions of CSG, which utilize the categorical emotional labels during training and testing the models.

4.4.1 Emotionally Adapted Conditional Sequential GAN (CSG-Emo-Adapted)

Figure 4(b) illustrates the first proposed approach to model emotion. After training the CSG model with all the data, we separately adapt this model using the data associated with four emotions (i.e., the data with consensus labels for anger, happiness, sadness and frustration). Yosinski et al. [35] showed that the lower layers of DNNs are more generalizable than higher layers, which become more specific towards the primary task they are trained on. Therefore, we freeze the weights of the generator in the CSG on the first BLSTM layer, and fine tune the rest of the model, including the discriminator with the data associated with a given emotion. Freezing the weights is important to reduce the number of parameters to be learned given the reduced size of the data belonging to each emotion. We repeat this process for each emotion, creating emotion dependent models. The discriminator is the teacher of the generator. Therefore, it is important that the discriminator is fine-tuned with the adaptation data, so that the errors which correct the generator's mistakes are actually learned from the adaptation data. We hypothesize that this model helps the generator to synthesize more expressive lip motion sequences.

The CSG-EMO-Adapted model starts with the CSG model. In total, we have 2.8M parameters to learn when training the models from scratch (0.61M parameters for the discriminator, and 2.2M parameters for the generator). However, we only fine-tune the discriminator and the last layer of the generator during adaptation (i.e., 0.64M parameters). The number of learnable parameters for adaptation is only 22% of the parameters of the CSG model, which is important since we are using only a portion of the data for each emotion.

4.4.2 Emotion-aware Conditional Sequential GAN (CSG-Emo-Aware)

Figure 4(c) illustrates the second proposed approach. This model conditions both the generator and the discriminator in the CSG model on the soft emotional labels of the speaking turn parametrized with the 6D vector explained in Section 3.1 (see variable $\mathbf{e}$ in Fig. 4(c)). Compared with the CSG-Emo-Adapted model, this model better utilizes the IEMOCAP corpus, since all the segments are used, including the turns without consensus. The relationship between emotion and orofacial movements are assumed to be captured by the discriminator. We use real lip trajectories which are uncoupled with the acoustic and emotional speech features

as fake samples. This approach helps the discriminator to learn this kind of fake instances, forcing the generator to create orofacial sequences that not only are coupled with speech, but also convey the emotional state of the speaker.

4.5 Implementation Details

Our generator and the discriminator have two layers of BLSTMs. We set the number of nodes for the BLSTMs to 256 for the generator, and 128 for the discriminator. We implement our models using Keras with Theano as backend. We use *adaptive moment estimation* (ADAM) [36] as our optimizer. ADAM relies on the history of the gradient, in terms of its first and second moments, scaling the gradient to make the steps invariant to the gradient magnitude. This approach helps adapting the learning rate according to the changes in the loss at each iteration. For ADAM, we tried several learning rates $[0.001, 0.0001, 0.00001]$, selecting 0.0001, which gave the best loss reduction on the validation set. We set our batch size as 128 sequences with a fixed window size. We use the same window size as the one selected by Sadoughi and Busso [37, 38] which is 71 frames (591.7ms). This window size is also close to the the window size selected by Karras et al. [4], which is 520ms.

We noticed that pre-training the generator is very helpful and expedites the training of the GAN. The pre-training process relies on a DNN trained with BLSTM using *concordance correlation coefficient* (CCC). This framework is the BLSTM-CCC baseline used in the experimental evaluation (Sec. 5.1.3). We pre-train the generator for 200 epochs. After the generator is pre-trained, we pre-train the discriminator by freezing the weights of the generator. We train the discriminator for 100 epochs. After pre-training the models, we alternately train the generator and the discriminator on each batch. This scheme freezes the generator's weights, and updates the discriminator's weights on the current batch. Then, it freezes the discriminator's weights, and updates the generator's weights with the goal of fooling the discriminator. This goal is achieved by switching the labels of the fake samples when training the generator to fool the discriminator (i.e., switching the labels from 0 to 1). With this approach, the weights of the generator are updated with the objective of classifying the synthesized samples as real by the discriminator. We train all the CSG models for 50 epochs, alternating at each batch between updating the discriminator and updating the generator. All the adapted CSG models are also fine-tuned using this adversary scheme for 50 epochs.

The CSG models are pre-trained by maximizing the CCC. Equation 3 defines the CCC between two continuous variables y (output) and t (target), where ρ is the Pearson's correlation coefficient between the two variables, μ_y and μ_t are the means of y and t, and σ_y^2 and σ_t^2 represent the variances of y and t. This loss function (ℓ) favors high correlation between the predictions and the true values, while reducing the shift in the predicted values compared with the original ones. Optimizing this loss function not only reduces the *mean squared error* (MSE), but also increases the Pearson correlation. It also increases the variance of the generated movements avoiding over smoothed trajectories.

$$CCC = \frac{2\rho\sigma_y\sigma_t}{\sigma_y^2 + \sigma_t^2 + (\mu_y - \mu_t)^2} \qquad \ell = 1 - CCC \tag{3}$$

5 Experimental Evaluation

5.1 Baselines

We compare our model with three competitive baseline systems. Recent studies have shown that DNN-based approaches are more effective than HMM-based systems for this task [3, 11, 14]. Therefore, we do not consider HMM-based solutions.

5.1.1 Sliding Window Deep Neural Network (SWDNN)

Taylor et al. [3] proposed a model composed of three layers of *rectified linear units* (ReLUs), with 2,000 nodes per layer, and a linear output layer to convert the input audio features to orofacial movements over a smaller window centered at the middle frame of the input window. This model is trained to minimize the MSE between the predictions and real samples. During testing, the average of the predicted output frames are selected as the orofacial pose of the middle frame, moving one frame at a time to generate the entire sequence. We implemented the model following the description of the model, with the same input (340ms $\sim$41 frames) and output (100ms $\sim$13 frames) window sizes. Similarly, we use batch normalization on the layers to speed up the training, and we use a dropout $p = 0.5$ on all the ReLU layers. We refer to this model as *sliding window deep neural network* (SWDNN).

5.1.2 Bidirectional LSTM with MSE Objective (BLSTM-MSE)

Fan et al. [14] proposed to use BLSTMs for learning the relationship between speech and orofacial movements, by minimizing the MSE between the predictions and the original movements. We implemented a model composed of two layers of 256 BLSTM units and a linear output layer, relying on the same objective (i.e., MSE). Fan et al. [14] implemented this model with varying length sequences. In our preliminary evaluation, we followed this approach by varying the length of the sequences, using the entire utterances. However, this approach generated over-smoothed trajectories that were not very appealing. Therefore, we train this framework with fixed window lengths, which generated more realistic sequences. We refer to this model as BLSTM-MSE.

5.1.3 Bidirectional LSTM with Concordance Correlation Objective (BLSTM-CCC)

This model is composed of two layers of 256 BLSTM units and a linear output layer (i.e., same as BLSTM-MSE). We define the loss function of this model based on CCC, inspired by the study of Sadoughi and Busso [13], which investigated facial movement prediction from speech. This model has the same loss function as our CSG models. This model is trained using a fixed window length. We refer to this model as BLSTM-CCC.

5.2 Implementation Details for the Baselines

We implement the baseline models using Keras with Theano as backend. The weights are initialized with the approach proposed by Glorot and Bengio [39] ($W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}}, -\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}}\right]$, where U is the uniform distribution, W is the weight between layers i and $i+1$ and n_i is the number of states for the i^{th} layer). We use ADAM as our optimizer, selecting a learning rate of 0.0001, since it gave the best loss reduction on the validation set for the baseline models. We set our batch size as 128 sequences with a fixed window size of 71 frames (591.7ms). We train all the baseline models for 200 epochs, except for the SWDNN model, which we train for an additional 800 epochs (the results on the validation set indicated that increasing the number of epochs reduced the error).

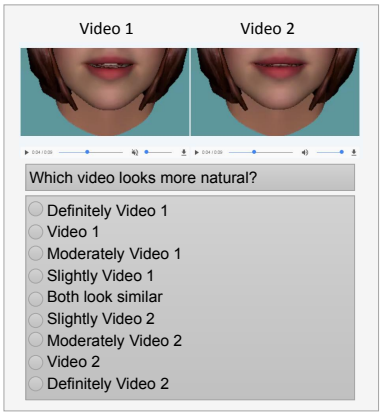


Fig. 5. Interface used for our subjective evaluations using AMT.

5.3 Evaluation Metrics

The models are compared with objective and subjective evaluations. This section describes the metrics and procedure that are consistently used in the experimental evaluation.

5.3.1 Objective Evaluation

Objective evaluations of the results generated by GAN are usually provided by fitting a distribution to the generated samples, and getting the likelihood of the test samples in that distribution [31]. This value shows how well the distribution of the generated samples matches the real samples. We use the Parzen window-based density estimation [31]. Since we use conditional GANs, we provide the input features from the test set, and get the samples from the generator. To estimate the distributions, we treat each frame as one sample. To avoid the curse of dimensionality and increasing the error in the Parzen estimator, we use *principal component analysis* (PCA) on the original samples to reduce the dimension of the samples from 45 to 15. A 15D vector preserves more than 95% of the variance of the original orofacial data. We use cross validation to set the bandwidth of the Parzen estimator on the samples generated by the generator. We estimate the log-likelihood of the test samples (Sec. 3.1) from the estimated distribution, reporting their average values and standard deviations. While we can always draw more samples from the CSG models by sampling different values from the noise distribution, we only generate one trajectory for each speaking turn, since the baseline systems can generate only one value per speech signal.

5.3.2 Subjective Evaluation

The trajectories that we generated not only need to have a similar distribution as the original sequences, but also need to be perceived as realistic. Therefore, we conducted subjective evaluations. People are more consistent in performing relative assessments than absolute ratings [40, 41]. Therefore, we perform subjective evaluations by asking for preference between two sequences generated with competing approaches. We ask "which video looks more natural?" in all the evaluations, except the evaluations in Section 6.3, where we additionally asked for the expressiveness of the videos. Figure 5 shows the interface. We define naturalness as the degree of realism of the facial movements in the animation, which includes closeness to the original trajectories, smoothness and lack of jittery movements. We provide multiple options to allow evaluators to convey their degree of certainty in the annotations ranging from "*definitely video 1*" to "*definitely video 2*". To report the results, we convert the selected options into percentage. For instance, the option "*definitely video 1*" is mapped to 100% for Video 1 and 0% for Video 2, and the option "*moderately video 1*" is mapped to 75% for Video 1 and 25% for Video 2.

We perform two different statistical tests on the comparison results. First, we compare the soft comparison assignments using a two way t-test with the null hypothesis that the two models being compared are perceived as similar (i.e., h_0 : MEAN $= 50\%$). Second, we convert the soft assignment labels into hard assignments, by using Equation 4. The variable i represents the i^{th} sample, n is the total number of samples, and r_i is the i^{th} evaluation. The function $\mathbb{1}()$ is one when the argument of the function is true, otherwise it is zero. For ties (i.e., 50%-50%), we assign one vote to each video. The correction $\sum_{i=1}^{i=n} \mathbb{1}\left(r_i = 50\right)$ in the denominator takes in consideration the ties. This approach allows us to compare the two models using a statistical proportion test on the hard assignments.

$$p = \frac{\sum_{i=1}^{i=n} \mathbb{1}\left(r_i \geqslant 50\right)}{n + \sum_{i=1}^{i=n} \mathbb{1}\left(r_i = 50\right)} \tag{4}$$

All our subjective evaluations are conducted on *Amazon mechanical turk* (AMT). We decided to focus the animations only on the orofacial area so the evaluators could focus on the lip movements. Adding the rest of the face would introduce extra variations not related to the lip motions that would affect the analysis. For subjective evaluations, we randomly select five turns per emotion (i.e., 20 videos in total), and rendered their videos using the trajectory generated by the models. We consider the three baseline models (i.e., SWDNN, BLSTM-MSE and BLSTM-CCC), the three proposed CSG methods (i.e., CSG, CSG-Emo-Adapted, and CSG-Emo-Aware), and the original trajectories from the motion capture data. Therefore, we have 20 videos for each of the seven conditions. We also render 10 videos per emotional class as explained in Section 6.3.

The evaluators compare two videos at a time created for the same sentence. The placement of the videos and the ordering of the pairs are randomized throughout each task. They complete 20 comparisons per *human intelligent task* (HIT). The question is shown after the annotator plays the two videos, reducing the chance of evaluators answering the questions before watching the videos. We use the Cronbach's alpha to quantify the agreement between evaluators. We limit the pool of annotators to people who (1) participated on previous crowd-sourcing perceptual evaluations conducted by our laboratory, and (2) their annotations had high agreement with labels from gold standard sets. These studies included the annotation of emotional labels [42, 43] and dialog acts [44]. In total, we invited 150 evaluators, where the reported evaluations consider scores provided by this pool of raters.

6 Results

6.1 Noise Dimension

An important parameter of our model is the dimension of the noise. We use an m-dimensional Gaussian noise with diagonal covariance matrix and zero mean. To choose the dimension of the noise, we used the CSG model, changing $m \in \{1, 10, 40, 80, 150\}$. We performed subjective evaluations on 10 videos generated by each model from the validation set. Each video is compared with the video rendered with the original lip motion sequences. The results provide indirect comparisons between the models with different noise dimensions. We use the protocol described in Section 5.3.2 for AMT. We recruited 15 evaluators for this evaluation, each comparing 10 pairs of videos, resulting in three evaluations per video. The Cronbach's alpha between the annotators are $\alpha_1 = 0.72$, $\alpha_{10} = 0.65$, $\alpha_{40} = 0.78$, $\alpha_{80} = 0.50$ and $\alpha_{150} = 0.73$ (the subscript of α indicates the noise dimension). We discarded the evaluations of two raters whose average pairwise Cronbach's alpha was less than zero, repeating the HIT with other raters.

Before we explain the results, we describe the charts in Figures 6 to 12, which describe preference between two conditions labeled at the extremes. For example, the first chart in Figure 6 compares the original sequences with the sequences created with the proposed CSG model trained with $m = 150$. The charts give the distributions for the preferences, where the bars show the first and third quartiles of the preferences, the vertical gray lines indicate the median values, the target signs show the mean values, and the dashed lines show extremes values (i.e., $\min$ and $\max$). The horizontal axis represents the preference in terms of percentage, where 50% implies that the models are equally selected by the evaluators. Shifts of these markers toward one extreme indicate higher preference for one option. Depending on the distributions of the preferences, we include red dots representing outliers, which are values above or below thresholds. The upper threshold is set to $q_3+1.5(q_3-q_1)$, and the lower threshold is set to $q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 represent the first and third quartiles, respectively. The points identified as outliers are considered in the calculation of mean, median, q_1 and q_3, but they are excluded from the calculation of minimum and maximum values used for the dashed lines.

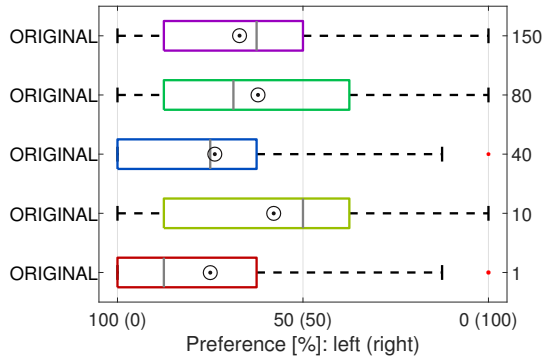


Fig. 6. Comparison of the CSG models for different dimensions of the noise. The bars represent the first and third quartiles. The circle represents the mean values for each condition, the dash lines represent the minimum and maximum values, the vertical gray lines represent the medians, and the red dots represent outliers.

TABLE 1
Comparing the results generated with the CSG and the baseline models in term of log-likelihood of the test samples in the estimated distribution by the Parzen estimator. All the pairwise comparison are statistically significant (p-value < 0.0001)

model	log-likelihood MEAN (STD)
SWDNN	-207.412 (268.452)
BLSTM-MSE	-190.642 (318.110)
BLSTM-CCC	-143.234 (317.674)
CSG	**-125.797 (241.979)**

Figure 6 gives the results for the optimal noise dimension. As expected, the average of the preferences are shifted towards the original sequences. The results suggest that $m = 10$ and $m = 80$ are the most competitive models (i.e., the bars are shifted toward the center). We select $m = 10$ as the dimension for the noise for the rest of the experimental evaluation.

6.2 Comparing the CSG Model with the Baselines

This section compares the CSG model with the three baseline models: the SWDNN, BLSTM-MSE, and BLSTM-CCC approaches. We compare these models with objective and subjective evaluations.

Objective Evaluation: We estimate the distribution of synthesized samples generated by the CSG and baseline models using the Parzen window density estimator. We generate 555K samples per model. Table 1 gives the mean and standard deviations of the log-likelihood of the test samples in the fitted distributions. The proposed CSG model is significantly better than all other alternative baselines. Note that all the pairwise comparisons in this table are statistically different (t-test: p-value < 0.001). The results demonstrate higher log-likelihoods for the CSG model compared with the baselines. Interestingly, BLSTM-CCC outperforms BLSTM-MSE showing the benefit of using CCC as a cost function.

Subjective Evaluation: Before we present the subjective evaluation, we evaluated the level of naturalness of the generated videos using Xface. We asked four evaluators to rate the naturalness level of 20 videos using the original lip

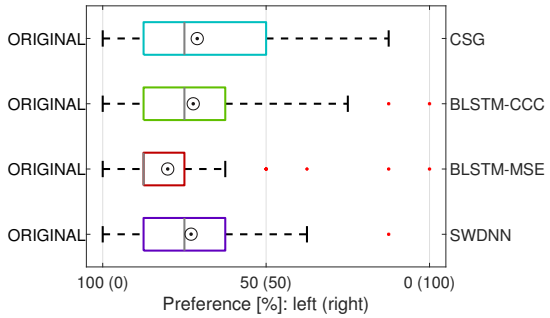


Fig. 7. Comparison of the CSG model and the baseline models with videos generated with the original lip motion sequences. The bars represent the first and third quartiles. The circle represents the mean values for each condition, the dash lines represent the minimum and maximum values, the vertical gray lines represent the medians, and the red dots represent outliers.

movements. They used a 10-point Likert-type scale (1 very unnatural, 10 very natural). The average of these ratings is 6.16, with standard deviation of 1.9. The result shows that the videos are perceived with an adequate level of naturalness.

The first phase of the subjective evaluations compares the animation synthesized by each of the models (CSG, SWDNN, BLSTM-MSE and BLSTM-CCC) with videos generated with the original motion capture recordings. We recruited 16 evaluators who annotated 20 pairs of videos, resulting in four evaluators per comparison. Figure 7 shows the result of these comparisons, where the agreement between evaluators in terms of the Cronbach's alpha are α_{SWDNN} =0.88, $\alpha_{BLSTM-MSE}$ =0.91 and $\alpha_{BLSTM-CCC}$ =0.83, and α_{CSG} =0.82. The t-test shows that the means of all these ratings are not equal to 50% (p-value $< 1e^{-11}$), indicating that the original sequences are preferred. Notice that these approaches are speech-driven models that do not rely on transcriptions, so the synchronization of the lips is not perfect. Therefore, it is expected that videos generated with original trajectories will be preferred by the evaluators. By creating strategies that can better model the lip and speech synchronization, our models aim to reduce this gap. We estimate the proportion of preference for the original motion capture data with each of these models using Equation 4. The original sequences are preferred 87% over the SWDNN model, 92% over the BLSTM-MSE model, 78% over BLSTM-CCC model, and 76% over the CSG model. While the annotators preferred the videos with the original sequences (all these proportions are statistically significant –p-value < 0.01), the CSG and BLSTM-CCC models are the approaches where the preferences are closer to 50%.

Figure 7 provides indirect comparisons between the models. The second phase of the subjective evaluations directly compares our proposed CSG model with the BLSTM-CCC model, which was the most competitive baseline model in the indirect comparisons. We use 20 videos synthesized by the CSG and BLSTM-CCC approaches. We recruit four raters for this task, who evaluated the 20 videos using the approach described in Section 5.3.2 (four evaluations per comparison). Figure 8 shows the results, where the Cronbach's alpha between the annotators is α =0.61. The

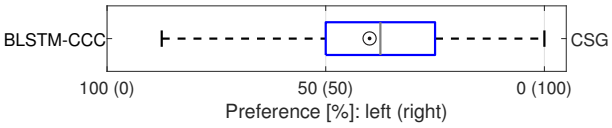


Fig. 8. Comparison of the CSG model with the baseline (BLSTM-CCC). The figure follows the same convention used in Figure 7.

box plot in Figure 8 shows that the mean of the preference for the CSG model is 60% over the best baseline (BLSTM-CCC). This level of preference is statistically higher than 50% (p-value $< 1e-4$). After dichotomizing the labels with equation 4, the evaluation shows that the preference for the CSG model is 68% (also statistically significant with p-value < 0.01). These results show that the CSG model is clearly preferred over the BLSTM-CCC baseline.

Objective and subjective evaluations clearly show better performance for the CSG model over the baseline methods. The next section evaluates the benefits introduced by considering emotion in the expression-dependent CSG models.

6.3 Expression-Dependent CSG Models

This section evaluates the CSG-Emo-Aware and CSG-Emo-Adapted models. The objective evaluations consider the log-likelihood estimations (Sec. 5.3.1), and the accuracy of emotion classifier trained on the original data and tested on the synthesized results. The subjective evaluations consider the preference and expressiveness of the expression-dependent models.

Objective Evaluation: We estimate the distribution of the samples using the Parzen window density estimator. The number of the test samples (i.e., frames) across emotional classes are 62K for anger, 107K for happiness, 92K for sadness and 106K for frustration. We generate the same number of samples using each of the models. Table 2 gives the log-likelihood of the test samples evaluated on the distribution of the generated samples. All the pairwise comparisons between the CSG model and each of the expression-dependent CSG models are statistically significant (t-test: p-value < 0.05), with two exceptions: the comparison between CSG and CSG-Emo-Adapted for sadness and the comparison between the CSG and CSG-Emo-Aware for frustration. These results indicate that adding emotion in the models help in generating samples that are closer to the original sequences. Table 2 shows that the CSG-Emo-Adapted model constantly achieves better results than the CSG-Emo-Aware model. All the pairwise comparisons between the CSG-Emo-Aware and CSG-Emo-Adapted models are statistically significant (p-value < 0.05). Adapting the top layers is an effective method to create expressive-dependent models for lip motion.

We evaluate whether the generated lip movements convey emotional cues by training an emotion classifier. Using the same train, test and validation partitions used for the models, we train a categorical emotion classifier on the original motion capture data. The classification tasks use lip motion sequence to recognize anger, happiness, sadness, and frustration. Since the emotional labels are assigned to each speaking turn, we extract sentence-level features by extracting statistics from the 45D orofacial pose parameters. The statistical features include mean, median, first

TABLE 2
Comparing the CSG model with the two expression-dependent CSG models for each emotional category. The values correspond to the log-likelihood of the test samples in the estimated distribution by the Parzen estimator. Asterisks indicate when the expression-dependent CSG models are significantly better than the CSG model (*: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001).

model	log-likelihood: MEAN (STD)			
	ang	hap	sad	fru
CSG	-76.410 (130.751)	-77.139 (183.269)	**-162.832 (208.430)**	-148.145 (239.602)
CSG-Emo-Adapted	**-72.495 (119.353)*****	**-75.365 (168.302)***	-163.679 (229.455)	**-136.634 (226.355)*****
CSG-Emo-Aware	-74.201 (116.977)**	-81.103 (189.359)***	-179.041 (245.666)***	-147.291 (232.307)

quartile, third quartile, minimum, maximum and standard deviation, resulting in a 315D feature vector (45 parameters $\times$ 7 functionals). We have 1,898 training samples across anger (359), happiness (525), sadness (390) and frustration (624). The validation set has 624 speaking turns (anger-132, happiness-187, sadness-111, and frustration-194), and the test set has 617 speaking turns (anger-129, happiness-169, sadness-152, and frustration-167). We train a SVM classifier, maximizing the F_1-score (i.e., F_1-score$= 2 \times \frac{precision \times recall}{precision + recall}$) on the validation set to determine the kernel function and the soft margin parameter. The best result on the validation set was obtained with a linear kernel and a soft margin equals to $c = 0.8$. We evaluate this model on the test set, using the original motion capture recordings and the lip trajectories generated by the CSG models. Table 3 shows the results in terms of accuracy, recall, precision, and F_1-score. The classifier tested with the original data achieves an F_1-score of 61.5%. The same classifier tested with the samples generated by the CSG model achieves an F_1-score of 37.7%. These results show that the CSG model does not preserve the emotional cues in the lip motion trajectories. This problem is overcome by the expression-dependent CSG models. The same classifiers tested with the samples generated by the CSG-Emo-Aware and CSG-Emo-Adapted models achieve F_1-scores of 62.5% and 70.8%, respectively. These results are statistically significantly better than the F_1-score achieved when using samples generated by the CSG model (p-value $< 1e^{-8}$). Even though we train with the original data and test with synthesized data, the emotion classifiers are able to recognize emotions with similar or better accuracy than when we test with the actual lip motion trajectories. These results demonstrate that the proposed expression-dependent CSG models generate lip motion trajectories conveying expressive cues similar to the original recordings.

Subjective Evaluation: We also perform subjective evaluations on the results, starting with indirect comparisons where the original sequences are used as reference. We recruited eight evaluators for the expression-dependent models, who were asked to evaluate 20 pairs of videos (four evaluators per task). Figure 9 shows the results, where we repeat the results obtained for the CSG model presented in Figure 7. The agreements between the evaluators in terms of Cronbach's alpha are α_{CSG} =0.82, $\alpha_{CSG-Emo-Adapted}$ =0.79, and $\alpha_{CSG-Emo-Aware}$ =0.85. The t-test shows that the evaluators prefer the original models, as expected (p-value $< 10^{-9}$). According to Equation 4, the proportions of preferences for the original motion capture data with each of the models are 76% over the CSG model, 80% over the CSG-Emo-Adapted model and 70% over the CSG-Emo-Aware model. All these proportions are statistically greater than 50% (p-value < 0.01), which is not surprising.

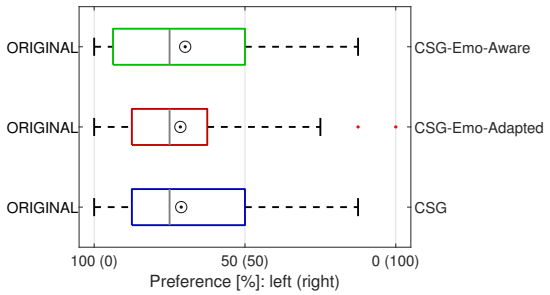


Fig. 9. Comparison of the CSG model and expression-dependent CSG models with videos generated using the original lip motion sequences. The figure follows the same convention used in Figure 7.

We also perform subjective evaluations to directly compare two options where one of the lip motions was generated by the CSG model and the other with the expression-dependent models. We generate 20 videos for the CSG-Emo-Adapted model and 20 videos for the CSG-Emo-Aware model, creating 40 video pairs. We recruit eight evaluators who compare 20 pairs of videos, resulting in four evaluations per comparison. Figure 10 gives the results of this evaluation. The Cronbach's alpha between the annotators are $\alpha_{CSG-Emo-Adapted}$ =0.48 and $\alpha_{CSG-Emo-Aware}$ =0.63. While the mean of the preferences are slightly toward the expressive-dependent models, the differences are not statistically significant (t-test, p-value $= 0.48$ for CSG-Emo-Adapted, and p-value $= 0.29$ for CSG-Emo-Aware). We directly compare the CSG-Emo-Aware and CSG-Emo-Adapted models. We recruited four evaluators to compare the 20 videos generated by these models, resulting in four evaluations per comparison. The Cronbach's alpha between the annotators is α =0.70. Figure 11 shows the results, the preference for the CSG-Emo-Adapted model compared to CSG-Emo-Aware is not statistically significant (t-test, p-value $= 0.27$). We estimate the proportion preference using Equation 4, which shows that 55% of the evaluations prefer the CSG-Emo-Adapted model. The proportion test shows that the preference is not statistically significant (p-value $= 0.25$).

The results on Figures 10 and 11 evaluate preference in terms of the level of naturalness. We conclude the subjective evaluation by assessing the perceived expressions elicited by the different lip motion sequences. Note that the ability to convey emotional cues in the videos is constrained by the expressiveness of the rendering toolkit Xface (see discussion in Section 6.4). We only consider the CSG-Emo-Adapted model, which is the emotionally dependent CSG model

TABLE 3
Emotion recognition results over the synthesized orofacial movements by the models in terms of accuracy (Acc.), precision (Prec.), recall (Rec.) and F_1-score on the four emotions of ang (angry), hap (happy), sad, and frustrated. Accuracy is calculated on the whole data, and for precision, recall and F_1-score we have provided the average of the result across the four classes (ave).

Orofacial Source	Acc. [%]	Prec. [%]					Rec. [%]					F_1-score [%]				
		ang.	hap.	sad.	fru.	ave.	ang.	hap.	sad.	fru.	ave.	ang.	hap.	sad.	fru.	ave.
Original	62.6	70.8	78.7	70.5	46.3	66.6	35.7	76.3	59.9	71.9	60.9	47.4	**77.5**	64.8	56.3	61.5
CSG	39.9	58.6	41.9	50.8	32.0	45.8	13.2	41.4	43.4	55.7	38.4	21.5	41.7	46.8	40.6	37.7
CSG-Emo-Adapted	**70.5**	**78.4**	**85.7**	**83.2**	**51.8**	**74.8**	53.5	**67.5**	**81.6**	**76.6**	**69.8**	**63.6**	75.5	**82.4**	**61.8**	**70.8**
CSG-Emo-Aware	62.2	64.8	75.2	66.7	48.8	63.9	**54.3**	66.3	64.5	62.3	61.8	59.1	70.4	65.6	54.7	62.5

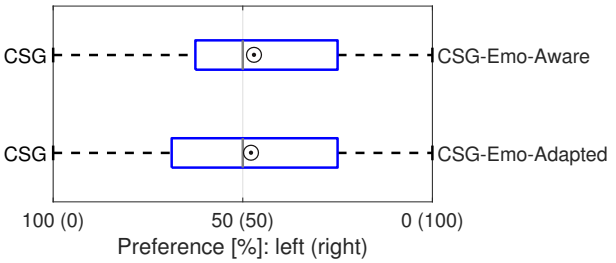


Fig. 10. Comparison of the CSG model with the two expression-dependent models. The figure follows the same convention used in Figure 7.

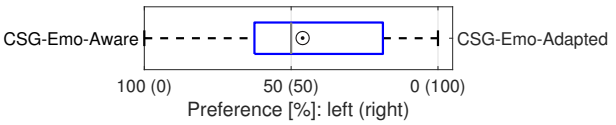


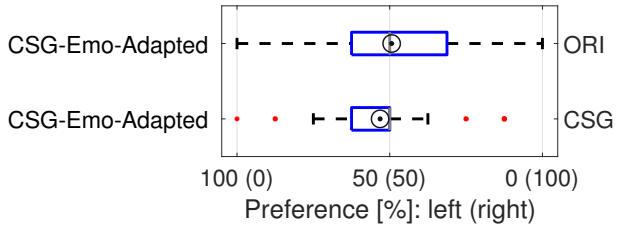
Fig. 11. Comparison of the two expression-dependent CSG model. The figure follows the same convention used in Figure 7.

with better results according to the objective comparisons (Table 2). We compare the expressiveness of the CSG-Emo-Adapted sequences with the original and CSG sequences. We define expressiveness as the degree of effectiveness of the animation in conveying the intended emotion. For this evaluation, we render 10 randomly selected videos per emotional categories (i.e., consensus label). We recruit evaluators, resulting in four evaluations per comparison. For the evaluations, we rely on pairwise comparisons using the same interface shown in Figure 5. The only difference with the previous evaluations is the question, which is rephrased. For example, for happiness, we ask "Which video looks happier?" We ask similar questions for anger, sadness and frustration. Similar to the subjective evaluation in Section 6.2, we only show the orofacial area in the animation. This setting is particularly important for the evaluation of emotions, where the upper face region conveys important emotional information. This paper only focuses on the emotional perception elicited by the lip movements. By adding expressive movements on the entire face, the perception of emotions could not be completely attributed to the lip movements. By keeping a neutral face, the perception of emotion from the lip motion would be masked by the neutral pose of the rest of the face, affecting the analysis. Figure 12 gives the results of this evaluation. The results consistently indicate that the lip motion sequences created by the CSG-Emo-Adapted model are selected as more emotional than the CSG model. The preference is statistically significant for happiness (t-test: p-value $= 0.016$). According to Equation 4, we observe that the sequence of the CSG-Emo-Adapted model are selected over the ones from the CSG model 56% for anger, 65% for happiness, 57% for sadness and 48% for frustration. When we compare the CSG-Emo-Adapted sequences with the original sequences, we observe that the proportion of the preferences for the original sequences estimated with Equation 4 are 54.4% for anger, 40.4% for happiness, 46.7% for sadness, and 72.1% for frustrated. The proportion test only finds the proportion of preferences for frustration to be statistically higher than 50%. This result shows that the CSG-Emo-Adapted model created videos for anger, happiness and sadness where the differences in expressiveness compared to the original videos are not statistically significant.

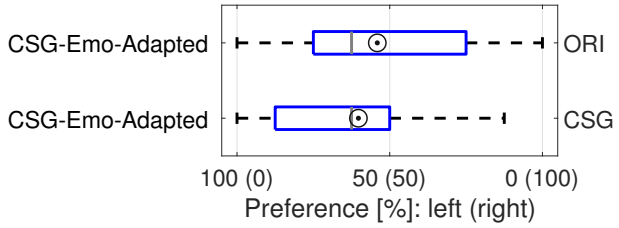
6.4 Discussion

Overall, the experimental evaluations demonstrate that the proposed CSG models perform better than the competitive baselines used in this study. The objective evaluations using log-likelihood of the models reveal the superiority of the expression-dependent CSG models over the CSG model, which show the flexibility of the proposed framework to incorporate expressive lip motions. The results using emotion classifiers also show that the expression-dependent CSG models are able to generate lip motion sequences conveying emotional cues.

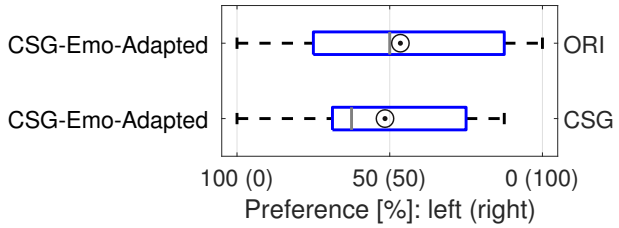
The results also suggest that the model can be improved. While the subjective evaluations show a clear trend across emotional classes, the preference toward the expression-dependent CSG models are statistically significant only for happiness. An important observation is that some emotions may have a stronger effect on the orofacial area. For example, there is a clear relationship between happiness and the lip configuration. For other emotions, the relationship may be more subtle. We hypothesize that the lack of expressiveness in Xface and the lip parametrization used in this study are the main reasons for not obtaining more decisive results in the subjective evaluation. Xface is a simple toolkit that allows us to render an animation by parametrizing the lip shape using motion capture data, which simplifies our modeling setting. The fact that the emotional cues on the expression-dependent CSG models are not clearly perceived by the evaluators suggest that a more sophisticated rendering toolkit is needed. Furthermore, the IEMOCAP database does not have information for the inner part of the lips, so the lip shape is exclusively defined by the outer lip markers. Therefore, our framework may not capture important lip details that are important to convey the target emotion. We are currently working to address these problems. Even



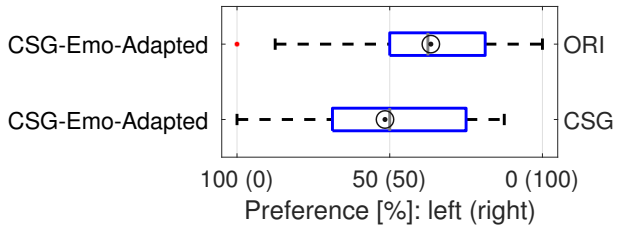
(a) ang: "Which looks angrier?"



(b) hap: "Which looks happier?"



(c) sad: "Which looks sadder?"



(d) fru: "Which looks more frustrated?"

Fig. 12. Comparison of the perceived target emotional category elicited by the CSG-Emo-Adapted models. The comparison includes the original sequences and the CSG model. The figure follows the same convention used in Figure 7.

with these limitations, the study clearly demonstrates the modeling potential of the CSG framework, creating exciting opportunities for lip motion generation with speech-driven methods.

7 Conclusions

This paper proposed the CSG model, a conditional GAN that generates orofacial movements from acoustic features. This model learned the conditional distribution of the data with an adversarial training objective, using a generator and a discriminator. The discriminator has to distinguish between real data and samples created by the generator. This adversary training forces the generator to create lip motion trajectories that are realistic. To capture the complex coupling between lip motion and speech, we also presented samples with real audio and motion capture data from different recordings. This type of fake samples presented to the discriminator imposes special emphasis on the temporal dynamic of the lip motion sequences created by the generator. We compared this model with three competitive baselines. The objective and subjective evaluations of the results demonstrated better performance for our model.

A BLSTM model can be easily implemented with a look-ahead buffer to generate the movements during real-time applications. Therefore, the use of BLSTM does not necessarily prevent real-time use of the algorithm. If latency is an important factor, the CSG models can always be implemented with unidirectional LSTM units. One of the strengths of the CSG model is its flexibility to constrain the trajectories by the underlying emotion content, creating expressive lip motion sequences. We proposed two emotion-dependent extensions of our model, where we know the target categorical emotion during testing: the CSG-Emo-Adapted, and CSG-Emo-Aware models. The CSG-Emo-Adapted model adapts the network by using the partitions associated with each emotion. The CSG-Emo-Aware model explicitly adds the target emotion as an extra input vector. The results demonstrated that the testing data is better represented by the distribution of the samples generated by the expression-dependent CSG models than the ones from the CSG model. The emotion classification evaluation using the generated sequences also indicated that both expression-dependent CSG models can generate emotional cues observed in natural recordings. The subjective evaluation showed that the CSG-Emo-Adapted model is perceived as more emotional across emotional classes, especially for happiness where the preference was statistically significant.

The experimental evaluation demonstrated the benefits of the proposed CSG models, opening new opportunities to improve the models. The current study focuses on the orofacial area, since this area presents a stronger interplay between articulatory and emotional content. A direct extension of the proposed framework is to generate facial expressions for the entire face, where the emotion can be controlled by specifying the target category. A second extension of the approach is to increase the resolution of the parameters describing the lips. The IEMOCAP corpus does not include inner mouth markers. The inner mouth markers contain subtle differences across emotional categories, which we currently ignore. Using a more dense representation for the lip configuration will help us to generate more expressive and naturalistic animations. Likewise, Xface is not a very expressive toolkit. We expect to create better animations by relying on better rendering toolkits. Furthermore, the selection of the emotional classes in this paper was determined by the emotional labels in the IEMOCAP corpus. However, our proposed solution is flexible. If we have data for a given emotion, we can easily create models that replicate the expressive behaviors associated with the emotion using our data-driven framework. Finally, the current version of the framework is exclusively driven by speech, without the need for phonetic information. This is one of the key features of our approach. However, if the target application requires better synchronization between lip motion and the phonetic content, the current framework can be extended by constraining the models with the underlying lexical content. For example, we can incorporate lexical content by adding phonemes as additional constraints in the CSG models. This approach can be implemented with an *automatic speech recognition* (ASR) system that estimates automatic transcriptions. The drawback of using ASR, is the extra computation and

delay in the prediction of lip movements. We will address these research directions in our future work.

ACKNOWLEDGMENTS

This study was funded by the National Science Foundation (NSF) award IIS-1718944.

REFERENCES

[1] S. Mariooryad and C. Busso, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013, pp. 1–6.

[2] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[3] S. Taylor, A. Kato, I. Matthews, and B. Milner, "Audio-to-visual speech conversion using deep neural networks," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 1482–1486.

[4] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 94, July 2017.

[5] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, OR, USA, June 2013, pp. 3382–3389.

[6] X. Li, Z. Wu, H. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 1477–1481.

[7] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 93, 2017.

[8] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro, "A practical and configurable lip sync method for games," in *Motion in Games (MIG 2013)*, Dublin, Ireland, November 2013, pp. 131–140.

[9] Z. Deng, J. Lewis, and U. Neumann, "Synthesizing speech animation by learning compact speech co-articulation models," in *Computer Graphics International (CGI 2005)*, Stony Brook, NY, USA, June 2005, pp. 19–25.

[10] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, October 2005.

[11] J. Parker, R. Maia, Y. Stylianou, and R. Cipolla, "Expressive visual text to speech and expression adaptation using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. New Orleans, LA, USA: IEEE, March 2017, pp. 4920–4924.

[12] C.-C. Lee, J. Kim, A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Speech in affective computing," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. New York, NY, USA: Oxford University press, December 2014, pp. 170–183.

[13] N. Sadoughi and C. Busso, "Joint learning of speech-driven facial motion with bidirectional long-short term memory," in *International Conference on Intelligent Virtual Agents (IVA 2017)*, ser. Lecture Notes in Computer Science, J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, S. Kopp, M. Mancini, and G. Varni, Eds. Stockholm, Sweden: Springer Berlin Heidelberg, August 2017.

[14] B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong, "A deep bidirectional LSTM approach for video-realistic talking head," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5287–5309, May 2016.

[15] P. Ekman, T. S. Huang, T. Sejnowski, and J. C. Hager, "Final report to NSF of the planning workshop on facial expression understanding," National Science Foundation, University of California, San Francisco, CA, USA, Technical Report, July-August 1992.

[16] K. Choi, Y. Luo, and J. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *The Journal of VLSI Signal Processing*, vol. 29, no. 1-2, pp. 51–61, August 2001.

[17] L. Xie and Z.-Q. Liu, "A coupled hmm approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, August 2007.

[18] S. Eskimez, R. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *Latent Variable Analysis and Signal Separation (LVA/ICA 2018)*, ser. Lecture Notes in Computer Science, Y. Deville, S. Gannot, R. Mason, M. Plumbley, and D. Ward, Eds. Guildford, UK: Springer Berlin Heidelberg, July 2018, vol. 10891, pp. 372–381.

[19] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 95:1–13, July 2017.

[20] "FaceFX," https://facefx.com, 2018, accessed Dec., 2018.

[21] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[22] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 2474–2477.

[23] S. Mariooryad and C. Busso, "Facial expression recognition in the presence of speech using blind lexical compensation," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 346–359, October-December 2016.

[24] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017.

[25] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Technical Report 132, 1996, http://www.praat.org.

[26] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[28] K. Balci, "Xface: MPEG-4 based open source toolkit for 3D facial animation," in *Conference on Advanced Visual Interfaces (AVI 2004)*, Gallipoli, Italy, May 2004, pp. 399–402.

[29] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

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

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

[32] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances In Neural Information Processing Systems*, 2016, pp. 613–621.

[33] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," *arXiv preprint arXiv:1707.04993*, 2017.

[34] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016.

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, Montreal, Canada, December 2014, pp. 3320–3328.

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

[37] N. Sadoughi and C. Busso, "Expressive speech-driven lip movements with multitask learning," in *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, May 2018, pp. 409–415.

[38] ——, "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, April 2018, pp. 6169–6173.

[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, Sardinia, Italy, May 2010, pp. 249–256.

[40] G. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017.

[41] ——, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[42] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[43] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2017.

[44] N. Sadoughi, Y. Liu, and C. Busso, "Meaningful head movements driven by emotional synthetic speech," *Speech Communication*, vol. 95, pp. 87–99, December 2017.

Najmeh Sadoughi (S'14) received her BSc and MS in Biomedical Engineering (Bioelectric) from Amirkabir University of Technology (AUT), Tehran, Iran, in 2010 and 2012, respectively. She received her PhD in Electrical Engineering from the University of Texas at Dallas, Richardson, USA, in 2017. She was a member of Multimodal Signal Processing Laboratory (MSP) as a research assistant (2013-2017). She received a Jonsson School Graduate Scholarship (2013-2014). She has been a research scientist in EMR.AI (2018-present). Her research interests are in machine learning, natural language processing, and analysis and synthesis of gestures.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.