Speech Driven Talking Face Generation from a Single Image and an Emotion Condition

Sefik Emre Eskimez, *Member, IEEE,* You Zhang, *Student Member, IEEE,* and Zhiyao Duan, *Member, IEEE*

***Abstract*—Visual emotion expression plays an important role in audiovisual speech communication. In this work, we propose a novel approach to rendering visual emotion expression in speech-driven talking face generation. Specifically, we design an end-to-end talking face generation system that takes a speech utterance, a single face image, and a categorical emotion label as input to render a talking face video synchronized with the speech and expressing the conditioned emotion. Objective evaluation on image quality, audiovisual synchronization, and visual emotion expression shows that the proposed system outperforms a state-of-the-art baseline system. Subjective evaluation of visual emotion expression and video realness also demonstrates the superiority of the proposed system. Furthermore, we conduct a human emotion recognition pilot study using generated videos with mismatched emotions among the audio and visual modalities. Results show that humans respond to the visual modality more significantly than the audio modality on this task.**

***Index Terms*—Talking face generation, emotion, audiovisual, multimodal**

I. Introduction

SPEECH communication does not solely depend on the acoustic signal. Visual cues, when present, also play a vital role. The presence of visual cues improves speech comprehension [1], [2], [3], [4] in noisy environments and for the hard-of-hearing population. Consequently, researchers developed systems that can automatically generate talking faces from the speech in order to provide visual cues when they are not available [5], [6], [7], [8], [9], [10], [11], [12]. These systems can increase the accessibility of abundantly available audio-only resources for the hearing impaired population and can also increase the quality of human-computer interactions [13], [14]. They also have broad applications in entertainment, education, and healthcare.

During speech communication, emotion has a direct impact on the transmitted message and can change the meaning drastically [15]. Studies have shown that predicting emotions purely from speech audio is quite difficult for untrained people [16] and that we heavily rely on visual cues in emotion interpretation [17]. Therefore, to make the visual rendering more realistic and to improve speech communication, it is important for automatic talking face generation systems to render visual emotion expressions.

One approach to emotional talking face generation is to first estimate the expressed emotions from the speech utterance and then render them in the generated talking faces. This approach, however, is limited by the speech emotion recognition accuracy and does not allow independent control of emotion expression in the visual rendering. In this work, we take a different approach: we ignore emotions expressed in the speech audio and condition the talking face generation on an independent emotion variable. This approach provides direct and more flexible control of visual emotion expression and can enable more personalized applications in entertainment, education, and interactive assistive devices. It also provides a powerful tool for behavioral psychologists to conduct emotion-relevant experiments that were not possible before. For example, one can investigate how humans respond to and interact with their conversational partners' emotional expressions by manipulating these emotions in audio and visual modalities independently.

In this work, we propose the first neural network system that generates emotional talking faces from speech conditioned on categorical emotions. The network takes a speech utterance, a reference face image, and a categorical emotion condition as inputs then generates a talking face that is synchronized with the input speech and contains emotional expressions. Our main contributions are as follows:

- We propose a new talking face generation method that can be conditioned on categorical emotions.
- We propose an emotion discriminative loss that classifies rendered visual emotions.
- We conduct a pilot study on human emotion perception using talking face videos with mismatched emotions among the audio and visual modalities.

The rest of the paper is organized as follows: We first present related work on talking faces in Section II. We then describe the proposed method and objective functions in Section III. Then, we present experimentation details, the objective evaluations, and Amazon Mechanical Turk (AMT) subjective evaluations in Section IV. Finally, we conclude the paper in Section V. Our source code is publicly available[1].

II. Related Work

A. Emotion Models

In affective computing, researchers leverage emotion models in order to develop automatic systems that can detect emotions. The most utilized emotion models for automatic systems are 1) categorical models 2) dimensional models. Readers are referred to [18], [19], [20], [21], [22] for more comprehensive coverage of emotion models.

S. E. Eskimez was previously with the Department of Electrical and Computer Engineering, and Y. Zhang and Z. Duan are currently with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, 14627 USA e-mail: emreeskimez@gmail.com.

[1]https://github.com/eeskimez/emotalkingface

Categorical emotion models assume there are a small number of emotions that are hard-wired to the human brain [23], [24]. Ekman's model suggested six basic emotion categories: anger, disgust, fear, happiness, neutral, and sadness.

Dimensional models argue that emotions are correlated, and each emotion category can be represented with a combination of values from emotional dimensions [25], [26], [27], [28], [21], [22]. The most famous example is the arousal-valance (AV) dimensions, where each emotion is represented with an arousal axis that determines if the emotion is active or passive and with a valance dimension that determines if the emotion is positive or negative. Another example is the hourglass of emotions [21], [22]. These models allow a more precise representation of emotions compared to the categorical models.

It should be noted that the selection of an emotion model for an affective computing task highly depends on the availability of the labels.

B. Multimodal Emotion Analysis

In this section, we cover some of the key trends in multimodal emotion analysis. Cambria et al. [29] stated that affective computing and sentiment analysis is an interdisciplinary effort in combining traditional psychological emotion research with machine learning. There are many works on multimodal sentiment analysis [30], [31], [32], [33], [34], [35], [36], [37], [38].

Chaturvedi et al. [34] proposed a fuzzy sentiment classifier to predict mixed sentiment. Ma et al. [39] summarized empathetic dialogue systems, identifying that most of the systems focus on emotion-expressiveness or emotion-awareness. In emotion-expressive systems, emotion labels are employed for the loss calculation. In emotion-aware systems, emotion labels are taken as additional input, which is more similar to our approach. Another important line of research is sentic blending [30], [36], [38]. The conceptual and emotional information related to natural language can be defined as semantic and sentic, respectively. The key idea is to fuse many single modality systems that have different time scales and output labels.

Our emotion study fits well with this research line but is novel since our generated videos with mismatched emotion provide more insights on which modality humans rely on.

C. Emotional Talking Face Generation

The automatic generation of talking faces from the speech is drawing increasing attention from researchers in recent years. One approach is to first convert the speech input to face landmarks [6], [40], [9], [41], [42], [10], [5], [12] and then estimate video frames using the predicted landmarks. In Suwajanakorn et al.'s two-stage system [40], a long short-term memory (LSTM) network first predicts the principal component analysis (PCA) coefficients of face landmarks from speech features, and then retrieves candidate frames from the dataset according to the PCA coefficients, stitching them together. However, this system works only for a single speaker. Another two-stage system was proposed by Chen et al. [5]. The system first predicts 68 face landmarks from speech using an LSTM-based network [7], and then predicts a few talking face images from the conditioned image and the face landmarks. They also employ a discriminator network to improve image quality. In another work, Egor et al. [43] proposed a style-based landmark-to-image conversion method using generative adversarial networks (GANs) with a few shots of the target face. This method, however, lacks landmark adaptation methods to solve personality mismatch issues.

Some researchers designed systems that directly map speech features to video frames. Features extracted from the speech often include the Mel-frequency cepstral coefficients (MFCCs), energy, and the first- and second-order temporal derivatives of these features. Gutierrez et al. [44] proposed an integral system that employs the k nearest-neighbor (KNN) algorithm to map the speech dynamics to video frames. The KNN procedure requires a lot of memory and has a long search time, which leads to impracticality in real applications. Some approaches that model the conversion from speech features to the movement of articulators [45], [46], [47], but they only focus on specific regions of the face, thus generating less natural or expressive animation. Chung et al. [6] proposed a convolutional neural network (CNN) that takes as input a face image and speech features and generates a talking face video. The generated video is then sharpened by another CNN, which is trained on pairs of artificially blurred images and their clear originals. Chen et al. [48] proposed another method that predicts video frames of the lip region from speech features and a conditioned lip image. They introduced a GAN loss in addition to the reconstruction loss to sharpen the generated overly smooth video frames. However, this method is limited to only generating the lip region instead of the entire talking face. Zhou et al. [12] proposed a GAN-based method that models the whole face and introduced a temporal-GAN loss in addition to the reconstruction loss to improve the temporal dependency across frames. Song et al. [9] proposed another method that generates talking faces by using a conditional recurrent adversarial network to improve the realness. Yu et al. [11] adopted optical flow and a self-attention mechanism to capture adjacent and long-range temporal dependencies across video frames.

In addition to the above-mentioned two-stage or speech-feature-driven approaches, there are also end-to-end systems that generate talking faces directly from a conditioned image and the speech signal. Vougioukas et al. [42] proposed a temporal-GAN method to generate more realistic image sequences. They further improved their methods with three discriminators [10] that focus on improving the realness of video frames, the continuity between generated frames, and the synchronization between audio and visual data. Eskimez et al. [49] proposed an end-to-end talking face generation system that is robust to noisy speech input. The system contains a frame discriminator to improve image quality and a pair discriminator to improve lip-speech synchronization. They proposed a mouth region mask (MRM) to further improve the lip-speech synchronization and showed that it leads to better alignment than the baselines.

Regarding emotional talking face generation, existing work is somewhat limited. Cosatto et al. [50] sample facial details from a database and then project to a 3D head model to allow realistic expressions. However, the generated emotional expressions focus more on the upper part of the face and lack variation for long animations. Karras et al. [41] adopted an end-to-end network to learn a latent representation of emotion states and use the latent code as a control to generate 3D mesh animations. This method effectively discovers emotion variations in the data, but the learned emotion states are difficult to interpret and do not model facial features such as wrinkled eyes and head motion to generate facial expressions. Sadoughi et al. [51] extended the conditional-GAN-based model to take the target emotion as an input, but this method is limited to generating the lip area instead of the whole face.

Recently, Fang et al. [52] proposed a talking face generation system that takes audio and an image as input. This system enforces the generated videos to convey the emotion contained within the speech input. This is different from our method, which allows independent control of the emotion of the generated visual signal from that of the audio input. Also, their generated videos contain a high amount of visual artifacts (e.g., ambiguity, pixel jittering, face deformation) that render them unrealistic.

D. Multimodal Human Emotion Perception

Emotion perception from auditory and visual stimuli has been examined in recent years. Existing work [53], [54], [55], [56], [57] concludes that different modalities complement each other and that there are also intermodal effects. Cowie [56] showed that perception is sensitive to stimuli from multiple modalities in data from both simulated and natural interactions. Jessen et al. [55] suggested that emotional visual content yields a more reliable prediction of auditory information. Schirmer et al. [53] explored modalities in terms of neural responses and showed that each modality provides a distinct insight, and that multimodal perception converges for holistic emotion recognition.

Most of the existing work was focused on emotionally congruent stimuli from these two modalities; little work examined incongruent stimuli. Tsiourti et al. [58] investigated human responses to emotions expressed by the body and voice of humanoid robots, showing that cross-modal incongruency decreased emotion recognition accuracy. Piwek et al. [59] found that subjects weighted visual cues higher in emotion judgments when presented emotionally incongruent audiovisual clips with happy or angry emotion. However, the visual content was conveyed by point-light displays instead of natural images.

III. Method

Instead of inferring emotion from the input speech [41], [51], in this work, we propose to use emotions as an input condition to our system. The motivation is to decouple the speech and emotion conditions. This allows us to manipulate emotions during the generation of face videos. Figure 1 shows an overview of the system, which employs the GAN framework. Our generator network architecture is built based on our previous work [49], with a modification to accept the emotion condition input. For the discriminator networks, we use one discriminator to distinguish between emotions expressed in videos, and another discriminator to distinguish between the real and generated video frames.

A. Generator

The generator network contains the following sub-networks: speech, image, noise, and emotion encoders, and a video decoder.

1) Speech Encoder: The speech encoder processes the input speech waveform and outputs a speech embedding. It follows the original implementation of [49] without any modification. It contains five convolutional layers with 1-D kernels operating in the time domain. The number of filters, filter sizes, and strides for these layers are as follows: (64, 63, 4), (128, 31, 4), (256, 17, 2), (512, 9, 2), (16, 1, 1), respectively. Each convolutional layer is followed by a LeakyReLU activation with a 0.2 slope. Since our network accepts 8 kHz speech signals, our speech encoder outputs 125 feature vectors per 1 second of speech. We add a context layer after these five convolutional layers to concatenate the past and future speech features. The context layer reduces the 125 time-steps to 25 time-steps by passing only every fifth frame to the next layer. Therefore, our generated videos are in 25 frames-per-seconds (FPS). The output of the context layer is fed to a fully connected layer, followed by two LSTM layers, which output the speech embedding sequence.

2) Image Encoder: The image encoder computes an image embedding from the input condition face image. The architecture follows the original implementation without any modification [49]. It contains six layers of 2-D convolutional layers with the following number of filters, kernel sizes, and down-sampling factors: (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), (512, 4, 1), respectively. A LeakyReLU activation with a 0.2 slope follows each convolutional layer. Note that nearest-neighbor interpolation is used for downsampling rather than using strides. This eliminates the artifacts in the generated images. The final image embeddings and intermediate representations are all passed to the video decoder using U-Net style skip connections [60].

3) Emotion Encoder: The emotion label is first encoded as a one-hot vector and fed into the emotion encoder. The emotion encoder uses a two-layer fully connected (FC) neural network to project the one-hot vector to an emotion embedding. This embedding is replicated for each time step. Again, we use a LeakyReLU activation with a 0.2 slope after every FC layer.

4) Noise Encoder: For each frame of the video, we generate a noise vector drawn from the standard Gaussian distribution. A single-layer LSTM processes this sequence of noise vectors and outputs the noise embedding. This module aims to model the head movements that are not correlated with speech, image, and emotion.

5) Video Decoder: We modify the video decoder described in [49] to accept the additional emotion embedding. We concatenate the speech, image, noise, and emotion embeddings and feed them into the decoder. For each time step, the decoder

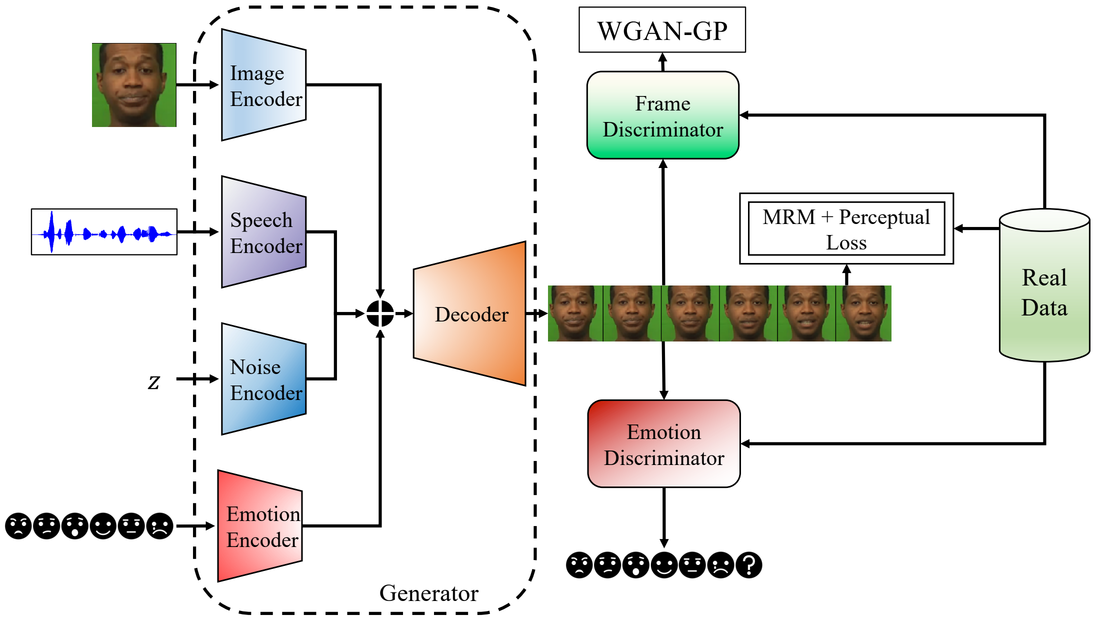


Fig. 1: Overview of the proposed neural network system. It accepts a reference image, a speech waveform, a random vector from the standard normal distribution, and a categorical emotion as input, concatenates their embeddings, and generates a talking face video that is synchronized with the input speech and expresses the input emotion. During training, besides the MRM reconstruction loss and perceptual loss, the network employs two discriminative losses: the frame discriminator for image quality and the emotion discriminator for emotion expression.

uses convolutional layers to project the embeddings into 4×4 images using two FC layers and reshape operations. These 4×4 images are concatenated channel-wise with the skip connections coming from the image encoder in the U-Net fashion for the next layers, except for the last layer. The number of filters in each convolutional layer is the same as for the corresponding layer in the image encoder. A LeakyReLU activation with a 0.2 slope follows each convolutional layer, except for the last layer, where instead, hyperbolic tangent activation is used since the images are normalized to have values between -1 to 1.

B. Frame Discriminator

The frame discriminator aims to improve the image quality of the generated video and to keep the target identity consistent throughout the video. First, we repeat the target image for the number of frames in the input video and concatenate them together. Then, each frame is processed by five layers of 2-D convolutional layers. The number of filters, kernel sizes, and strides of these convolutional layers are as follows: (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), respectively. The output is then flattened and fed into a two-layer FC network, which classifies the frame as fake or real. Each layer is followed by a LeakyReLU activation with a 0.2 slope except for the last layer, where we do not use an activation, since our system employs Wasserstein GAN with a gradient penalty [61].

C. Emotion Discriminator

The emotion discriminator is essentially a video-based emotion classifier, with the inclusion of an additional class for fake videos. It aims to improve the emotional expression generated by our network. The first part of the network follows the same architecture as the frame discriminator: five layers of 2-D convolutional layer followed by two FC layers. We process each frame of the video and feed the resulting sequence into an LSTM layer. The last time step of the output of the LSTM layer is fed into an FC layer that outputs probabilities of the seven classes: six emotions (anger, disgust, fear, happiness, neutral, and sadness) plus the fake class as in [62]. When we take a training step for the discriminator, we calculate the sparse categorical cross-entropy loss using the emotion label for the real video and the fake label for the generated video. When updating the generator, we calculate the sparse categorical cross-entropy loss using the emotion label we used for generating the video.

D. Objective Functions

Our system employs multiple objective functions that focus on different aspects of the generated videos: an MRM loss proposed in [49] to improve mouth-audio synchronization, a perceptual loss to improve image quality, a frame GAN loss for image quality, and an emotion GAN loss for emotion expression.

1) Mouth Region Mask (MRM) Loss: The MRM loss is a weighted L1 reconstruction loss between the generated and ground-truth videos around the mouth region. It uses a 2D Gaussian centered at the mean position of mouth coordinates as the weights. The intuition of MRM is to manually drive the attention of the network to the mouth region to improve the mouth-audio synchronization.

2) Perceptual Loss: We employ a pre-trained VGG-19 network [63] and calculate intermediate features of the following layers from both the generated and ground-truth videos: 4, 9, 18, 27, and 36. Then, a mean-squared loss between these intermediate features is calculated as the perceptual loss to improve image quality.

3) Frame Discriminator Loss: To further improve the image quality, especially the sharpness, we use a frame GAN loss calculated by the frame discriminator. Instead of the vanilla GAN loss, we use Wasserstein GAN for more stable training.

4) Emotion Discriminator Loss: To ensure emotion expression in generated videos, we use an emotion GAN loss calculated by the emotion discriminator, which is a categorical cross-entropy loss using six emotion classes plus a "fake" class, similar to [62]. In a vanilla GAN discriminator, samples are only classified as real or fake, rather than choosing between multiple emotion classes; if the generator only generates samples from a single emotion class all the time, the vanilla discriminator would still classify them as real. The proposed discriminator, on the other hand, incorporates multi-class classification losses and mitigates this issue of mode collapse for a multi-class generation.

The full objective function for the generator step is as follows:

$$J_{GEN} = \alpha L_1^{MRM} + \beta L_2^{Perceptual} + \gamma J_{FD} + \delta J_{ED} \ , \quad (1)$$

where J_{GEN} is the generator loss, L_1^{MRM} is the MRM loss, $L_2^{Perceptual}$ is the perceptual loss, J_{FD} is the frame GAN loss, J_{ED} is the emotion GAN loss, and $\alpha, \beta, \gamma, \delta$ are the respective weights of each component.

IV. Experiments

In this section, we describe the data used in experiments, the hyper-parameters of the neural networks, and the objective and subjective evaluation procedure. We choose the temporal GAN approach described in [10] as our baseline since it is the closest to our method. We use the pre-trained model and inference code provided by the authors to generate baseline videos. Although it cannot control the emotions through a conditioned input, it can generate emotional expressions that are inferred from the speech.

A. Dataset

We used the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset [64]. It contains video clips of 91 actors (48 male and 43 female) expressing six categorical emotions: anger, disgust, fear, happiness, neutral, and sadness. The age range of the actors is between 20 to 74. Each video clip shows one actor speaking a sentence from a set of 12 sentences and simulating one of the emotion categories. The image resolution of the provided videos is 480x360, and the sampling rate is 30 frames per second (FPS). The audio is sampled at 44.1 kHz. We downsampled the video to 25 FPS and the audio to 8 kHz. We followed the same train (70%), validation (15%), and test (15%) splits as [10]. We used the same files for these splits to ensure a fair comparison. During testing, the speech utterance, the conditioned emotion, and the conditioned image input to the generator network for each generation are all from the same ground-truth video, where the condition image is the first frame of the video.

As the same actor's face in different videos may be at different spatial locations, for easing the training, we need to align them across videos. For alignment, first, we choose a template image for the actor where the face is symmetrical. We extracted face landmarks from this template image as the template landmarks. Then, for each video of the actor, we estimated the similarity transform parameters between the template landmarks and extracted landmarks of the first frame using three points: the temporal mean points of the left eye, the right eye, and the nose. Note that we only took the first frame of each video to estimate the transformation, and used it to align the remaining frames to the template image. In this way, the faces in the resulting videos start from the same spatial location but can wander off to different parts of the scene. This allows us to model the natural head movements in addition to facial expressions.

During training, we randomly augmented the data using the Albumentations library [65] to improve the generalization capability of our network. The data augmentation includes randomly changing brightness, contrast, gamma, hue, saturation, and value. In addition, our algorithm includes contrast limited adaptive histogram equalization, adding random Gaussian noise to the image, and shuffling the channels, and shifting RGB values for each channel.

B. Implementation Details

To initialize our network, we trained it from scratch using only MRM and perceptual losses for 100k iterations. Then, we trained it for another 100k iteration using the full objective function. We used Adam optimizer for all networks with $\beta_1 = 0.5, \beta_2 = 0.99$. The learning rate for the generator was 1e-4 during the initialization and 1e-5 during the GAN training. Both discriminators' learning rates were 1e-4. The constants $\alpha, \beta, \gamma, and\delta$ mentioned in Section III-D were 100, 1, 0.01, and 0.001, respectively. The weight for the gradient penalty when training the frame discriminator was 10. All images were normalized between the -1 to 1 value range. During initialization, the mini-batch size was set to 8, and during GAN training, it was set to 4. The number of frames per sample was set to 32. The training took approximately one week using a GTX 1080 TI GPU. For the baseline method, we use the pre-trained model (trained with the CREMA-D dataset) provided by the authors.

C. Objective Evaluation

We evaluated the image quality of the generated videos using Peak SNR (PSNR) and Structural Similarity (SSIM) [66] between the generated video frames and the ground-truth video frames. To measure the audiovisual synchronization, we used the normalized landmarks distance (NLMD) [49] between landmarks extracted from the generated and ground-truth video frames.

TABLE I: Objective evaluation results for the baseline and our proposed method. For PSNR and SSIM, higher values are better; for NLMD, lower values are better.

Method	PSNR	SSIM	NLMD
Baseline [10]	29.64	0.82	0.124
Proposed	**30.91**	**0.85**	**0.113**

The baseline method generates 96x128 images, while our method yields 128x128 images. In other words, the foreground/background ratio differs in the generated videos. To ensure a fair comparison, we aligned the ground-truth, baseline, and proposed videos into a template image and cropped them into the same size using similarity transformation. Figure 2 shows the aligned videos, and Figure 3 shows example videos generated from the same condition image and speech, but different emotion conditions.

Fig. 2: Four examples comparing spatially aligned and cropped videos of the ground-truth (GT), baseline (BL) and proposed approach (OURS) for objective evaluation. Every fifth frame is shown for each video.

Table I shows the objective evaluation results of the baseline and our proposed methods. It can be seen that our method outperforms the baseline on all of the three metrics. We believe that perceptual loss is responsible for the improvement in image quality (PSNR and SSIM). For audiovisual synchronization (NLMD), even though our method does not use a discriminator paired with a synchronization loss as in [10], the improvement is as high as 8.9%, showing the effectiveness of the MRM loss.

TABLE II: Video-based emotion classification results for the ground-truth and our generated videos of the test set are shown.

Data	Accuracy	F1-Score
Ground-truth	62.71	62.39
Generated	65.67	66.65

1) Video-based Emotion Classification: In order to validate the emotional expression in the generated videos, we trained a video-based emotion recognition network using the CREMA-D train set. This network uses the same architecture as the emotion discriminator in Figure 1. We then classified the emotions within the ground-truth videos and our generated videos of the test set. The results are shown in Table II. The 6-class emotion classification accuracy on the ground-truth videos is 62.71%, which is comparable with [67], suggesting the validity of the video-based emotion classifier. The accuracy and F1-Score on the generated videos are slightly higher, even though the classifier was not trained on generated videos. This suggests that emotions are well expressed, and slightly exaggerated perhaps, in the generated videos.

We further show the confusion matrices of these two classification results in Figure 4. We observe similar patterns. First, they both have a strong diagonal. In particular, happiness is the easiest emotion to classify. This may be because happiness often contains smiling that is distinctive from other facial expressions, allowing the classifier as well as our generation system to capture it clearly. Second, some emotions are commonly confused with each other, such as fear and sadness. On the other hand, there are also differences in these confusion matrices. In particular, in the ground-truth videos, both fear and sadness are often misclassified as disgust, while in the generated videos, no other emotions are misclassified as disgust. Overall, the similarities outweigh the differences, showing that the emotional expressions in the generated videos resemble those of the ground truth. [2]

D. Subjective Evaluation

1) Research Questions: We design our subjective evaluation to investigate the following research questions: 1) Is our model effective in expressing emotions for video rendering? 2) How real are the generated videos of our model? 3) Which modality do people primarily rely on to perceive emotions? We conduct our evaluation on AMT.

2) Experimental Setup: Our AMT study consists of two Human Intelligence Tasks (HIT). For the first task, we randomly presented subjects generated and ground-truth videos and asked them to rate the realness and provide suggestions for making the videos more real. We also asked subjects to assign an emotion label to each video. This task aimed to answer the first and second research questions. For the second task, we generated videos that contain mismatched emotions in the audio and visual modalities. We asked subjects to assign

[2]For video samples, please visit the project webpage: http://www2.ece.rochester.edu/projects/air/projects/tfaceemo.html

Fig. 3: Frames of different talking face videos generated (in different rows) using the same face image (the first column) and speech utterance but different emotion conditions (from top to bottom: anger, disgust, fear, happiness, neutral, and sadness). One frame is shown for every 0.2 seconds.

one or two emotion labels to these videos. By doing so, we aimed to answer the third research question.

Task 1 - Emotion Classification and Realness Evaluation. In the first task, we pooled videos taken from ground-truth videos of the test set of the CREMA-D dataset and generated videos from the baseline and our models. For the baseline system, each video was generated from the speech recording and the first frame of the ground-truth video, while for the proposed system, each video was generated from the speech recording and the first frame of the ground-truth video, as well as the ground-truth emotion condition. We downsampled the ground-truth and baseline videos to 25 FPS to make them consistent with our generated ones. As described earlier, our method generates talking faces in 128×128 image size, while the baseline method generates videos in 96×128 image size. If we were to set the ground-truth videos to any of the two sizes, the subjects might be negatively biased toward the generated videos with the other size. To avoid this potential problem, we aligned the ground-truth videos with template faces of both sizes and obtained two sets of ground truth videos.

We released six batches of videos in total. For each batch, we randomly selected five videos from the two sets of ground-truth videos, five from the baseline videos, and five from our generated videos. One video from each category was repeated to check the consistency of the subjects' answers. Therefore, there were in total 18 videos in each batch. The videos in each batch were randomly shuffled. Across all of the six batches, the total number of videos for each emotion category was equal.

We recruited a total of 60 valid subjects (i.e., 10 for each batch) from AMT. The subjects were required to be located in the United States and to have a lifetime HIT approval rate higher than 95%. To encourage the subjects to treat the experiments more seriously, we made a bonus payment based on the subject's performance. Subjects were informed about the bonus payment before they started the experiments.

Subjects were informed that some of the 18 videos were recordings of real people, while other videos were rendered by artificial intelligence (AI) based on a single face image of one person and a speech recording of another person. Before presenting the 18 videos, we also presented two example videos of real recordings for each emotion, each in the two image sizes (128×128 and 96×128), to familiarize the subjects with these emotional expressions. These example emotions were ordered in alphabetical order.

We then asked subjects the following three questions for each video: *1) Which emotion is primarily expressed by the person?* This question is a multiple-choice question, and the subjects were asked to select one from the six emotions. *2) How realistic is the video?* The subjects can choose from *Definitely real*, *Somewhat real*, *Neutral*, *Somewhat unreal*, and *Definitely unreal*. *3) Which aspect(s) can be improved to make the video more real?* This question is a checkbox question, and

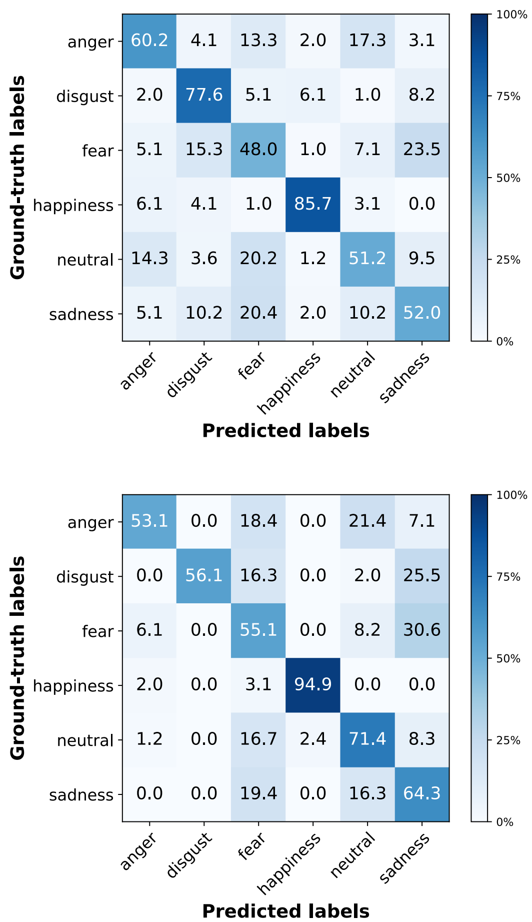


Fig. 4: Confusion matrices of video-based emotion classification on the ground-truth (above) and generated (below) videos using the proposed talking face generation system on the test dataset of CREMA-D. Each row sums to 100%.

the subjects could choose more than one aspect. The choices are *None*, *Image quality*, *Lip synchronization*, *Head movement*, and *Other*.

After receiving a survey, we checked its completeness and the consistency of answers to the nine questions of the three repeated videos. We rejected a total of 11 incomplete surveys and those that did not meet the consistency requirement. We then recruited other subjects until we collected 60 valid surveys. For our answer consistency requirement, an answer was considered inconsistent from the previous answer, if 1) the emotion classification was different for the first question, 2) the realness rating differed more than one level for the second question, or 3) the aspect selection differed for more than one options. Among the nine repeated questions, if more than five answers were inconsistent, then the entire survey was rejected.

Task 2 - Emotion Perception of Videos with Mismatched Emotions. As described in Section II-B, little work on human emotion perception used emotionally incongruent stimuli between the audio and visual modalities, and among these works, none used videos of humans as stimuli. Our emotional talking face generation system makes it possible to investigate human emotion perception from emotionally incongruent stimuli present in human speaking videos.

In the second task, we presented generated videos from our proposed system based on a face image, a speech recording, and an emotion condition. Both the face image and the speech recording were taken from a ground-truth video in the test set of the CREMA-D dataset. As a result, the speech recording conveyed a certain emotion. The emotion condition input, however, was not necessarily the same as the speech emotion, allowing for the possibility of generating videos with mismatched emotions between the audio and visual modalities. As there are six emotions in the dataset, there were 36 emotion pairs and 30 of them were mismatched. We generated 2 videos for each of the 36 pairs, shuffled them, and split them evenly into six batches. We also repeated two videos in each batch to check the answer consistency. Therefore, there were a total of 14 videos in each batch.

We recruited a total of 60 subjects (i.e., 10 for each batch) from AMT, with the same requirements as Task 1. We rejected a total of 4 incomplete surveys and those who had more than two inconsistent answers among the four repeated questions and recruited other subjects until we collected 60 valid surveys. The participants who completed Task 1 could not see this task from the AMT platform. The same bonus mechanism in Task 1 was applied to Task 2. In the survey, subjects were notified that all of the videos were AI rendered. Before presenting the generated videos, the subjects were also presented two example ground-truth videos for each emotion, only in the image size of 128×128, to familiarize them with the emotions. They were asked the following two multiple-choice questions for each video: *1) Which is the primary emotion expressed by the person?* The subjects could select one of the six emotions. *2) Which is the secondary emotion expressed by the person?* The subjects could select one of the six emotions and a *None* option if they only perceived the primary emotion.

3) Experimental Results: **Task 1 - Emotion Classification.** The confusion matrices of subjective emotion classification for ground-truth videos, baseline generated videos, and our generated videos are shown in Figure 5. Our videos yield a more diagonal confusion matrix compared with the baseline videos and result in patterns similar to those produced from the ground-truth videos. Specifically, subjects are more likely to classify the emotions in the baseline videos as *neutral*, while this happens much less frequently for our generated videos. This shows the power of the emotion condition input that our method utilizes. The overall classification accuracy is 59.2% (ground-truth), 28.9% (baseline), 55.3% (ours), respectively, demonstrating the efficacy in expressing emotions of our proposed emotional talking face generation system. It must be noted that the baseline system infers emotion from the speech input instead of taking the emotion condition as input. As emotion recognition from the speech is itself a challenging task, errors in this stage naturally influence visual emotion expression in the generated videos. Therefore, poor performance from the baseline system is expected. Interestingly, the 59.2% human emotion classification accuracy on ground-truth videos is slightly lower than that of our emotional classifier

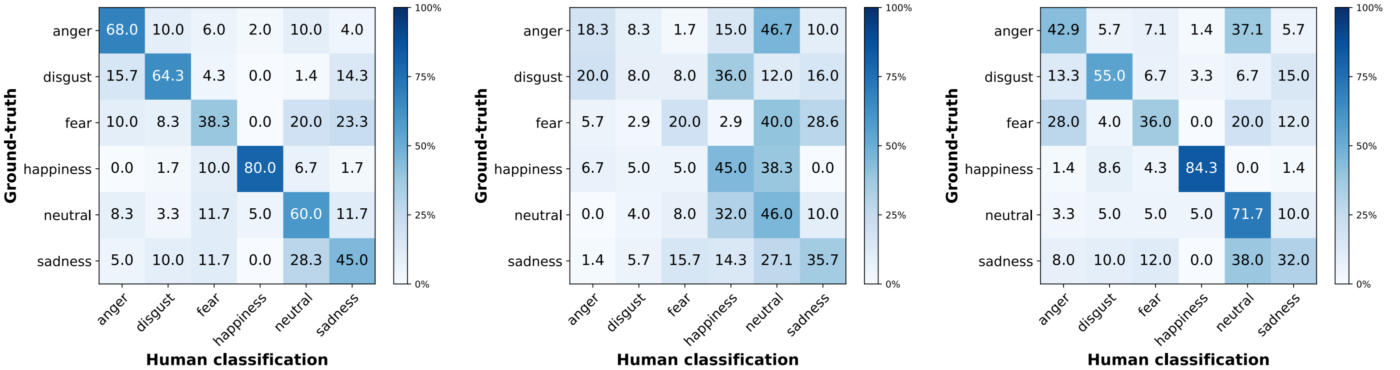


Fig. 5: Confusion matrices of human emotion classification in Task 1 on ground-truth videos (left), baseline generated videos (middle) and our generated videos (right).

in Section IV-C1, showing the challenge of visual speech emotion classification for humans. This observation is similar to a speech emotion classification observation in [16].

Task 1 - Realness Evaluation. For the realness question, the five options are mapped to a scale from 1 to 5, where ”definitely real” corresponds to 5 and ”definitely unreal” corresponds to 1. The result is shown in Figure 6. The average rating across all videos and subjects is 3.94, 3.71, and 3.81 for ground-truth, baseline, and our videos, respectively. This suggests that our generated videos are slightly more realistic than the baseline videos, yet they are still not as realistic as the ground-truth videos. Interestingly, even the ground-truth videos only received an average rating close to 4 (somewhat real). We think that this might be due to the relatively lower image resolution than what the subjects typically see in their daily life. This might also because the generated videos (especially OURS) are quite realistic, lowering the subjects' confidence in rating the ground-truth videos. A Wilcoxon signed-rank test [68] shows that the median difference between our ratings and the baseline ratings is statistically significantly greater than zero, at the significance level of 0.05 ($p = 0.048$).

Figure 7 shows the histograms of aspects suggested by the subjects to improve the realness of the videos. Consistent with the realness question, ground-truth videos received the most “none” votes, while our generated videos received the second most and the baseline received the least . The total count of votes for the four aspects to improve (“image quality”, “lip synchronization”, “head movement”, “other”) is 299 (ground-truth), 337 (baseline) and 325 (ours), respectively. Among the detailed aspects, the baseline videos received the most votes on “image quality” and “lip synchronization”; but it also received the least votes on “head movement” . This might be due to the fact that the baseline method is trained with 30 FPS videos and adopted a sequence discriminator to render head movements. On the other hand, our generated videos performed similarly to ground-truth ones on “lip synchronization” and “head movement”, suggesting the effectiveness of our proposed MRM loss. Nevertheless, the “image quality” of our generated videos is considered to need more improvement than the ground-truth videos.

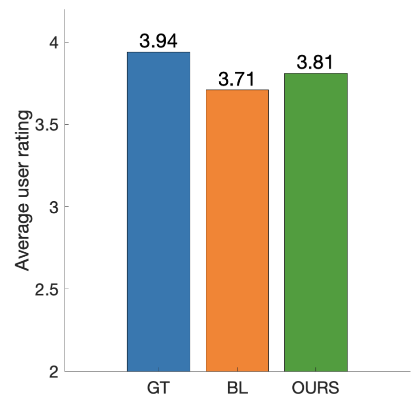


Fig. 6: User ratings on the realness of ground-truth (GT), baseline generated (BL) and our generated (OURS) videos.

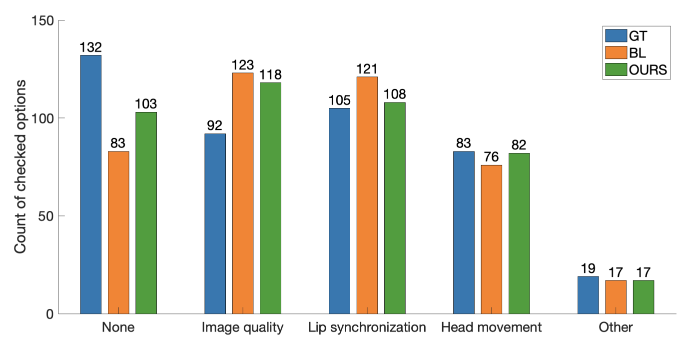


Fig. 7: Total count of chosen aspects for realness improvement of videos.

Task 2- Emotion Perception of Videos with Mismatched Emotions. In Task 2, subjects were asked the primary and secondary (if any) emotions they perceived from each video generated by our system, to investigate which modality people primarily rely on for emotion recognition. Overall, 426 of the 840 videos received two emotion labels. We first compared the primary emotion label with the visual emotion (i.e., the condition emotion when generating the video) and the audio emotion, respectively. The confusion matrices are shown in Figure 8. Overall, 35.2% of the primary emotion labels match with the visual emotion, while only 25.1% of them match with audio emotion. If we only consider videos with mismatched emotions, these numbers become 31.4% and 19.6%, respectively. This suggests that the subjects relied on the visual modality much more heavily than the audio modality for emotion perception. Among the six emotions, happiness and disgust seem to be the easiest to perceive from the visual modality, while anger and fear are the most difficult.

We then considered both primary and secondary emotions when comparing them with the audio and visual emotions. In this case, 44.9% of labeled emotions can be matched to the visual emotion, while 33.8% can be matched to the audio emotion. Similarly, if we only consider videos with mismatched emotions, these numbers become 41.1% and 28.2%. Again, this shows that the visual modality has a much greater effect than the audio modality on audiovisual speech emotion recognition for humans.

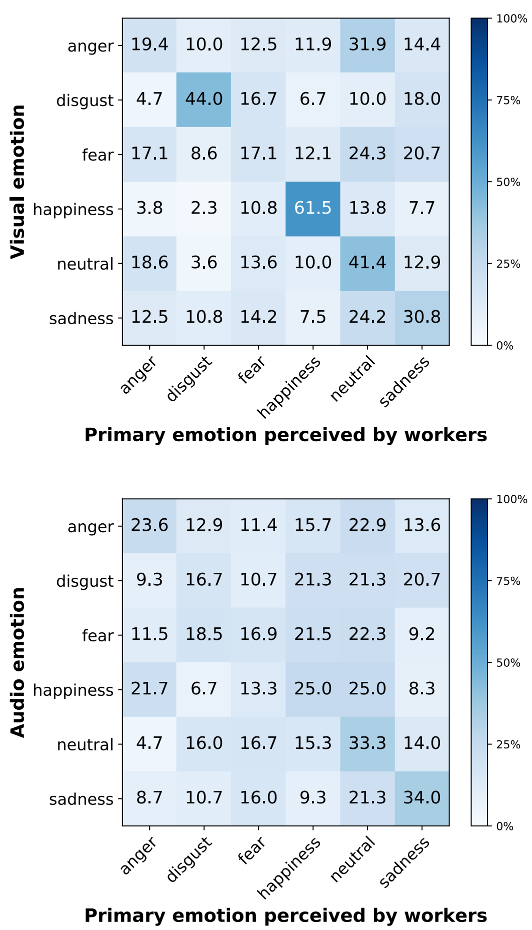


Fig. 8: Confusion matrices of the primary emotion label that AMT subjects give for each video, contrasted with the ground-truth visual emotion (above) and audio emotion (below) in Task 2.

V. Conclusions

In this work, we proposed a novel emotional talking face generation system that is conditioned on speech, reference image, and categorical emotion inputs. We evaluated our network against the ground-truth videos and a baseline system [10] and validated that our method can generate emotional expressions effectively. In addition, we conducted a subjective study on AMT, showing that our method yields close performance to the ground-truth videos in terms of realness and emotion classification. Furthermore, we also conducted a pilot study on human emotion perception from audiovisual speech with mismatched emotions expressed in the audio and visual modalities, showing that visual perception is more dominant than auditory perception. For future work, we plan to improve the image quality of generated videos. We also plan to extend this work to 3D animation and rendering.

Acknowledgment

This work is funded by the National Science Foundation (NSF) grant No. 1741472. You Zhang would like to thank the synergistic activities provided by the NRT program on AR/VR funded by NSF grant No. 1922591.

References

[1] C. A. Binnie, “Bi-sensory articulation functions for normal hearing and sensorineural hearing loss patients,” *Journal of the Academy of Rehabilitative Audiology*, vol. 6, no. 2, pp. 43–53, 1973.

[2] J. G. Bernstein and K. W. Grant, “Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners,” *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3358–3372, 2009.

[3] K. S. Helfer and R. L. Freyman, “The role of visual speech cues in reducing energetic and informational masking,” *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 842–849, 2005.

[4] R. K. Maddox, H. Atilgan, J. K. Bizley, and A. K. Lee, “Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners,” *eLife*, vol. 4, 2015.

[5] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, “Hierarchical cross-modal talking face generation with dynamic pixel-wise loss,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] J. S. Chung, A. Jamaludin, and A. Zisserman, “You said that?” in *British Machine Vision Conference (BMVC)*, 2017.

[7] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “Generating talking face landmarks from speech,” in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 372–381.

[8] A. Jamaludin, J. S. Chung, and A. Zisserman, “You said that?: Synthesising talking faces from audio,” *International Journal of Computer Vision*, vol. 127, no. 11-12, pp. 1767–1779, 2019.

[9] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, “Talking face generation by conditional recurrent adversarial network,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 7 2019, pp. 919–925. [Online]. Available: https://doi.org/10.24963/ijcai.2019/129

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

[11] L. Yu, J. Yu, and Q. Ling, “Mining audio, text and visual information for talking face generation,” in *International Conference on Data Mining (ICDM)*. IEEE, 2019, Conference Proceedings, pp. 787–795.

[12] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, “Talking face generation by adversarially disentangled audio-visual representation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9299–9306.

[13] H. Tang, Y. Fu, J. Tu, M. Hasegawa-Johnson, and T. S. Huang, “Humanoid audio–visual avatar with emotive text-to-speech synthesis,” *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 969–981, 2008.

[14] O. Schreer, R. Englert, P. Eisert, and R. Tanger, “Real-time vision and speech driven avatars for multimedia applications,” *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 352–360, 2008.

[15] M. Alpert, R. L. Kurtzberg, and A. J. Friedhoff, “Transient voice changes associated with emotional stimuli,” *Archives of General Psychiatry*, vol. 8, no. 4, pp. 362–365, 1963.

[16] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, “Emotion classification: how does an automated system compare to naive human coders?” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2274–2278.

[17] A. Esposito, “The perceptual and cognitive role of visual and auditory channels in conveying emotional information,” *Cognitive Computation*, vol. 1, no. 3, pp. 268–278, 2009.

[18] K. R. Scherer *et al.*, “Psychological models of emotion,” *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[19] S. PS and G. Mahalakshmi, “Emotion models: a review,” *International Journal of Control Theory and Applications*, vol. 10, pp. 651–657, 2017.

[20] M. Bourgais, P. Taillandier, L. Vercouter, and C. Adam, “Emotion modeling in social simulation: a survey,” *Journal of Artificial Societies and Social Simulation*, vol. 21, no. 2, 2018.

[21] E. Cambria, A. Livingstone, and A. Hussain, “The hourglass of emotions,” in *Cognitive behavioural systems*. Springer, 2012, pp. 144–157.

[22] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, “The hourglass model revisited,” *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[23] P. Ekman, “Basic emotions,” *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[24] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[25] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[26] J. A. Russell, “A circumplex model of affect.” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[27] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan, “Neural correlates of processing valence and arousal in affective words,” *Cerebral cortex*, vol. 17, no. 3, pp. 742–748, 2007.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,” *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[29] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Affective computing and sentiment analysis,” in *A practical guide to sentiment analysis*. Springer, 2017, pp. 1–10.

[30] E. Cambria, N. Howard, J. Hsu, and A. Hussain, “Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics,” in *2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI)*. IEEE, 2013, pp. 108–117.

[31] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[32] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[33] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[34] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, “Fuzzy commonsense reasoning for multimodal sentiment analysis,” *Pattern Recognition Letters*, vol. 125, pp. 264–270, 2019.

[35] R. Kaur and S. Kautish, “Multimodal sentiment analysis: A survey and comparison,” *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, vol. 10, no. 2, pp. 38–58, 2019.

[36] Y. Susanto, E. Cambria, B. C. Ng, and A. Hussain, “Ten years of sentic computing,” *Cognitive Computation*, vol. 13, 2021.

[37] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, “End-to-end multimodal affect recognition in real-world environments,” *Information Fusion*, vol. 68, pp. 46–53, 2021.

[38] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, “Sentiment analysis and topic recognition in video transcriptions,” *IEEE Intelligent Systems*, vol. 36, no. 2, 2021.

[39] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, “A survey on empathetic dialogue systems,” *Information Fusion*, vol. 64, pp. 50–70, 2020.

[40] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: learning lip sync from audio,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[41] 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, pp. 1–12, 2017.

[42] K. Vougioukas, S. Petridis, and M. Pantic, “End-to-end speech-driven facial animation with temporal gans,” in *British Machine Vision Conference (BMVC)*, 2018.

[43] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9459–9468.

[44] R. Gutierrez-Osuna, P. K. Kakumanu, A. Esposito, O. N. Garcia, A. Bojórquez, J. L. Castillo, and I. Rudomín, “Speech-driven facial animation with realistic dynamics,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 33–42, 2005.

[45] L. Xie and Z. Liu, “Realistic mouth-synching for speech-driven talking face using articulatory modelling,” *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 500–510, 2007.

[46] J. Park and H. Ko, “Real-time continuous phoneme recognition system using class-dependent tied-mixture hmm with hbt structure for speech-driven lip-sync,” *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1299–1306, 2008.

[47] A. Verma, L. V. Subramaniam, N. Rajput, C. Neti, and T. A. Faruquie, “Animating expressive faces across languages,” *IEEE Transactions on Multimedia*, vol. 6, no. 6, pp. 791–800, 2004.

[48] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu, “Lip movements generation at a glance,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[49] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “End-to-end generation of talking faces from noisy speech,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1948–1952.

[50] E. Cosatto and H. P. Graf, “Photo-realistic talking-heads from image samples,” *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, 2000.

[51] N. Sadoughi and C. Busso, “Speech-driven expressive talking lips with conditional sequential generative adversarial networks,” *IEEE Transactions on Affective Computing*, 2019.

[52] Z. Fang, Z. Liu, T. Liu, C.-C. Hung, J. Xiao, and G. Feng, “Facial expression gan for voice-driven face generation,” *The Visual Computer*, pp. 1–14, 2021.

[53] A. Schirmer and R. Adolphs, “Emotion perception from face, voice, and touch: comparisons and convergence,” *Trends in Cognitive Sciences*, vol. 21, no. 3, pp. 216–228, 2017.

[54] E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, “Multimodal databases of everyday emotion: Facing up to complexity,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[55] S. Jessen and S. A. Kotz, “On the role of crossmodal prediction in audiovisual emotion perception,” *Frontiers in Human Neuroscience*, vol. 7, p. 369, 2013.

[56] R. Cowie, “Perceiving emotion: towards a realistic understanding of the task,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3515–3525, 2009.

[57] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proceedings of the 6th International Conference on Multimodal Interfaces*, 2004, pp. 205–211.

[58] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, “Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots,” *International Journal of Social Robotics*, vol. 11, no. 4, pp. 555–573, 2019.

[59] L. Piwek, F. Pollick, and K. Petrini, “Audiovisual integration of emotional signals from others’ social interactions,” *Frontiers in Psychology*, vol. 6, p. 611, 2015.

[60] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[61] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[62] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[64] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[65] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

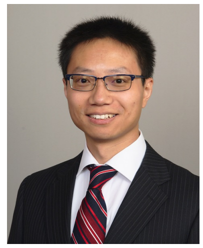
[66] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[67] N.-C. Ristea, L. C. Duţu, and A. Radoi, "Emotion recognition system from speech and visual information based on convolutional neural networks," in *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–6.

[68] E. E. Cureton, "The normal approximation to the signed-rank sampling distribution when zero differences are present," *Journal of the American Statistical Association*, vol. 62, no. 319, pp. 1068–1069, 1967.

S. Emre Eskimez attended Sabanci University and graduated with a Bachelor of Science degree in Mechatronics Engineering in 2011. He began graduate studies in the Department of Mechatronics Engineering at Sabanci University in 2011 and received a Master of Science degree in 2013. He began graduate studies in the Department of Electrical and Computer Engineering at the University of Rochester in 2014, received a Master of Science degree in 2015, and received his Ph.D. in 2019. He joined Microsoft Cognitive Services Research Team (previously Speech and Dialog Research Group (SDRG)) in 2019. His research interests include speech enhancement, generative models, speech processing, natural language processing, multi-modal learning, and deep learning.

You Zhang received a B.E. degree in automation from the University of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, China, in 2019, and an M.S. degree in electrical engineering from the University of Rochester, Rochester, NY, USA, in 2021. He is currently a Ph.D. student in the Audio Information Research lab at the University of Rochester. His research interests lie in machine learning and its applications in speech and audio, such as audio-visual analysis, synthetic voice spoofing detection, spatial audio, etc.



Zhiyao Duan (S'09, M'13) is an associate professor in Electrical and Computer Engineering, Computer Science and Data Science at the University of Rochester. He received his B.S. in Automation and M.S. in Control Science and Engineering from Tsinghua University, China, in 2004 and 2008, respectively, and received his Ph.D. in Computer Science from Northwestern University in 2013. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds. He is also interested in the connections between computer audition and computer vision, natural language processing, and augmented and virtual reality. He received a best paper award at the 2017 Sound and Music Computing (SMC) conference, a best paper nomination at the 2017 International Society for Music Information Retrieval (ISMIR) conference, a BIGDATA award and a CAREER award from the National Science Foundation (NSF). His research is funded by NSF, NIH, and University of Rochester internal awards on AR/VR and health analytics.