Animating Face using Disentangled Audio Representations

Gaurav Mittal
Microsoft
gaurav.mittal@microsoft.com

Baoyuan Wang
Microsoft
baoyuanw@microsoft.com

Abstract

Previous methods for audio-driven talking head generation assume the input audio to be clean with a neutral tone. As we show empirically, one can easily break these systems by simply adding certain background noise to the utterance or changing its emotional tone (to for example, sad). To make talking head generation robust to such variations, we propose an explicit audio representation learning framework that disentangles audio sequences into various factors such as phonetic content, emotional tone, background noise and others. We conduct experiments to validate that when conditioned on disentangled content representation, the generated mouth movement by our model is significantly more accurate than previous approaches (without disentangled learning) in the presence of noise and emotional variations. We further demonstrate that our framework is compatible with current state-of-the-art approaches by replacing their original component to learn audio based representation with ours. To the best of our knowledge, this is the first work which improves the performance of talking head generation through a disentangled audio representation perspective, which is important for many real-world applications.

1. Introduction

With recent advances in deep learning, we have witnessed growing interest in automatically animating faces based on audio (speech) sequences, thanks to applications in gaming, multi-lingual dubbing, virtual 3D avatars and so on. Specifically, the talking head generation is formulated as: given an input face image and an audio (speech) sequence, the system needs to output a video where the mouth/lip region movement should be in synchronization with the phonetic content of the utterance while still preserving the original identity.

As we all know, speech is riddled with variations. Different people utter the same word in different contexts with varying duration, amplitude, tone and so on. In addition to linguistic (phonetic) content, speech carries abundant information revealing details about the speaker's emotional state, identity (gender, age, ethnicity) and personality to name a few. Moreover, unconstrained speech recordings (such as from smartphones) inevitably contain a certain amount of background noise.

There already exists a large body of research in the domain of talking head generation[30, 31]. However, inspired by the rapid progress in Generative Adversarial Networks (GAN) [16], most of the recent works focus more on coming up with a better visual generative model to synthesize higher quality video frames. While impressive progress has been made by these prior methods, learning better audio representation specially tailored for talking head generation is being almost ignored without attracting much attention. For example, most of the previous works simply assume the input to be a clean audio sequence without any background noise or strong emotional tone, which is unlikely in practical scenarios as we described above. We highlight in our empirical analysis that the state-of-the-art approaches are clearly unable to generalize to noisy and emotionally rich audio samples. Although recent works, such as [14] and [1], show that visual signals can substantially help improve the audio quality (i.e., remove noise) when the system can see the visual mouth movements, it is however reasonable to assume that in many cases video is not available or there could be misalignment issue between audio and video in practical online applications.

Therefore, to make the system less sensitive to the noise, emotional tone and other potential factors, it is desired to explicitly disentangle the audio representations first, before feeding it into the talking head generation part, rather than simply treating it as black box and expect the network to implicitly handle the factors of variations. We argue that methods which explicitly decouple the various factors should have better chances to scale up the training and generalize well to unseen audio sequences, while implicit methods such as [31, 22, 26, 30] may have high risk of overfitting.

To this end, we present a novel learning based approach to disentangle the phonetic content, emotional tone and other factors into different representations solely from the input audio sequence using the Variational Autoencoder[23]

framework. We encourage the decoupling by adding (1) local segment-level discriminative loss to regularize phonetic content representation, (2) global sequence-level discriminative loss to regularize the emotional tone and (3) margin ranking loss to separate out content from rest of the factors, in addition to the regular VAE loss. We further propose our own talking head generation module conditioned on the learned audio representation, in order to better evaluate the performance. To summarize, there are two major contributions of this work:

- We present a novel disentangled audio representation learning framework for the task of generating talking heads. To the best of our knowledge, this is the first approach of improving the performance from audio representation learning perspective.

- Through various experiments, we show that our approach is not only robust to several naturally-existing audio variations but it is also compatible to be trained end-to-end with any of the existing talking head approaches.

2. Related Work

Speech-based facial animation literature can be broadly divided into two main categories. The first kind uses a blend of deep learning and computer graphics to animate a 3D face model based on audio. [26] uses a data-driven regressor with an improved DNN acoustic model to accurately predict mouth shapes from audio. [22] performs speech-driven 3D facial animation mapping the input waveforms to 3D vertex coordinates of a face model and simultaneously using an emotional state representation to disambiguate the variations in facial pose for a given audio. [32] introduces a deep learning based approach to map the audio features directly to the parameters of the JALI model [12]. [29] uses a sliding window approach to animate a parametric face model from phoneme labels. Recently, [11] introduced a model called Voice Operated Character Animation (VOCA) which takes as input a speech segment in the form of its corresponding DeepSpeech [18] features and a one-hot encoding over training subjects to produce offsets for 3D face mesh for subject template registered using FLAME [25] model. Their approach is for 3D facial animation which allows altering speaking style, pose and shape, but cannot adapt completely to an unseen identity. The paper suggests DeepSpeech features to be robust to noise but we later show that these are not as efficient as our disentangled representations which are modeled to decouple from content not just noise but also other variations including emotion and speaking style.

The second category includes approaches performing audio-based 2D facial video synthesis, commonly called "talking head/face generation". [7] learns a joint audio-visual embedding using encoder-decoder CNN model and [15] uses Bi-LSTM to generate talking face frames. [28] and [24] both generating talking head for specifically Barack Obama using RNN with compositing techniques and time-delayed LSTM with pix2pix [20] respectively. [21] uses RNN with conditional GAN and [30] uses Temporal GAN to synthesize talking faces. [6] employs optic-flow information between frames to improve photo-realism in talking heads. [31] proposes arbitrary-subject talking face generation using disentangled audio-visual representation with GANs.

Almost all of the previous approaches have been trained to work on clean neutral audio and fail to take into account many of the factors of variations occurring in real-world speech such as noise and emotion. Several recent works have demonstrated the importance of disentangled and factorized representation to learn a more generalized model [19]. To the best of our knowledge, our approach is the first attempt to explicitly learn emotionally and content aware disentangled audio representations for facial animation. Some previous approaches [22, 31, 26] do try to perform some kind of disentanglement but none of them explicitly deals with disentangling the different factors of variation in audio.

3. Method

Our proposed method consists of two main stages,

Learning Disentangled Representations from Audio The input audio sequence is factorized by a VAE into different representations encoding content, emotion and other factors of variations (Figure 1). KL divergence, negative log likelihood along with margin ranking loss ensure the learned representations are indeed disentangled and meaningful.

Generating Talking Head Based on the input audio, a sequence of content representations are sampled from the learned distribution which along with the input face image are fed to a GAN-based video generator to animate the face (Figure 2). We use temporal smoothing along with frame and video discriminator [5, 30] here but as we show later, our audio representations are compatible with any existing talking head approach.

3.1. Learning Disentangled Representations

Speech comprises of several factors which act independently and at different temporal scales. Taking inspiration from [19], we intend to disentangle content and emotion in an interpretable and hierarchical manner. We introduce

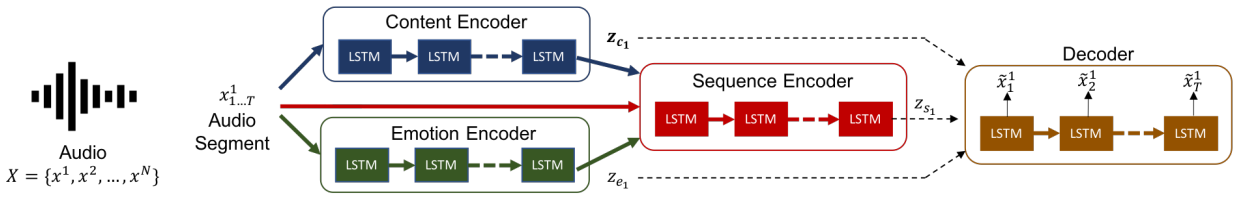

Figure 1. VAE architecture to learn emotionally and content aware disentangled audio representations

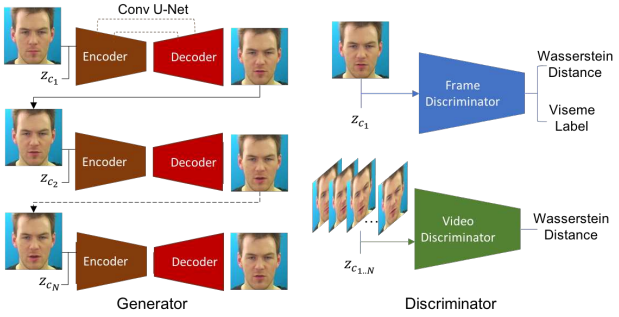

Figure 2. GAN based talking head generation model

several talking head generation specific novelties which include *lateral disentanglement* of content and emotion by explicit decoupling using margin ranking losses, and a mechanism to learn *variation-specific* priors which, unlike [19], may or may not be sequence agnostic. Syllables (linguistic content of an utterance) last only for few hundred milliseconds and do not exhibit significant variation within and between different speech sequences. We call this short duration a *segment* and encode syllables by a set of *latent content variables* regularized by a content-specific (viseme-oriented) prior that is sequence-independent. Since emotion is similar within a subset of utterances, we model emotion-related factors with *latent emotion variables* regularized by a prior shared among sequences with the same emotion annotation. Finally, we need *latent sequence variables* to encode residual variations of an entire utterance (*sequence*) that can't be captured by either content or emotion based variables.

Model Formulation Let $\mathcal{D} = \{\mathbf{X}^i\}_{i=1}^M$ consists of M i.i.d. sequences where every $\mathbf{X^i} = \{\boldsymbol{x}^{i,n}\}_{n=1}^{N^i}$ is a sequence of N^i observed variables with N^i referring to the number of content segments (syllables) in the i^{th} sequence and $\boldsymbol{x}^{i,n}$ referring to the n^{th} content segment in the i^{th} sequence. We omit i in subsequent notations to refer to terms associated with a single sequence without loss in generality.

Let each audio sequence $\mathbf{X}$ be randomly generated from a content-specific prior $\boldsymbol{\mu_c}$, emotion-specific prior $\boldsymbol{\mu_e}$ and sequence-specific prior $\boldsymbol{\mu_s}$ with N i.i.d latent variables for content $\mathbf{Z_c}$, emotion $\mathbf{Z_e}$ and sequence $\mathbf{Z_s}$ (one for each of the N segments in X). The joint probability for a sequence is therefore given by,

$$p_\theta(\mathbf{X}, \mathbf{Z_c}, \mathbf{Z_e}, \mathbf{Z_s}, \boldsymbol{\mu_c}, \boldsymbol{\mu_e}, \boldsymbol{\mu_s}) = p_\theta(\boldsymbol{\mu_c}) p_\theta(\boldsymbol{\mu_e}) p_\theta(\boldsymbol{\mu_s}) \prod_{n=1}^{N} p_\theta(\boldsymbol{x}^n | \boldsymbol{z}_c^n, \boldsymbol{z}_e^n, \boldsymbol{z}_s^n) p_\theta(\boldsymbol{z}_c^n | \boldsymbol{\mu_c}) p_\theta(\boldsymbol{z}_e^n | \boldsymbol{\mu_e}) p_\theta(\boldsymbol{z}_s^n | \boldsymbol{\mu_s}) \quad (1)$$

where the priors $\boldsymbol{\mu_c}$, $\boldsymbol{\mu_e}$ and $\boldsymbol{\mu_s}$ are drawn from prior distributions $p_\theta(\boldsymbol{\mu_c})$, $p_\theta(\boldsymbol{\mu_e})$ and $p_\theta(\boldsymbol{\mu_s})$ respectively and the latent variables $\boldsymbol{z}_c^n$, $\boldsymbol{z}_e^n$ and $\boldsymbol{z}_s^n$ are drawn from isotropic multivariate Gaussian centred at $\boldsymbol{\mu_c}$, $\boldsymbol{\mu_e}$ and $\boldsymbol{\mu_s}$ respectively. θ represents the parameters of the generative model and the conditional distribution of $\boldsymbol{x}$ (audio segment) is modeled as a multivariate Gaussian with a diagonal covariance matrix.

Since the exact posterior inference is intractable, we use Variational Autoencoder (VAE) to approximate the true posterior p_θ with an inference model q_ϕ given by,

$$q_\phi(\mathbf{Z_c}^i, \mathbf{Z_e}^i, \mathbf{Z_s}^i, \mu_c^i, \mu_e^i, \mu_s^i | \mathbf{X}^i) = q_\phi(\boldsymbol{\mu}_e^i) q_\phi(\boldsymbol{\mu}_s^i) \prod_{n=1}^{N} q_\phi(\boldsymbol{\mu}_c^{i,n}) q_\phi(\boldsymbol{z}_s^{i,n} | \boldsymbol{z}_c^{i,n}, \boldsymbol{z}_e^{i,n}, \boldsymbol{x}^{i,n}) q_\phi(\boldsymbol{z}_c^{i,n} | \boldsymbol{x}^{i,n}) q_\phi(\boldsymbol{z}_e^{i,n} | \boldsymbol{x}^{i,n}) \quad (2)$$

[19] suggests that the mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{s}}^i$ (one for each sequence) of $q_\phi(\boldsymbol{\mu}_s^i)$ be part of a lookup table and learned like other model parameters. We extend this idea to talking head scenario by introducing lookup tables for $q_\phi(\boldsymbol{\mu}_c^{i,n})$ and $q_\phi(\boldsymbol{\mu}_e^i)$ having different values $\tilde{\boldsymbol{\mu}}_{\boldsymbol{c}}^{i,n}$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{e}}^i$ for different viseme and emotion labels respectively. Being at sequence level, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{s}}^{i,n} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{s}}^i$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{e}}^{i,n} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{e}}^i \; \forall \, n$. Based on the annotation, the corresponding value is picked up from the respective tables for optimization. Such *variation-specific* priors allow the latent variables to be modeled effectively with samples with similar viseme/emotion made to lie closer together on the latent manifold. By further aligning $\boldsymbol{z}_s$ with $\tilde{\boldsymbol{\mu}}_{\boldsymbol{s}}^i$, we encourage $\boldsymbol{z}_s$ to encode sequence-specific attributes which have larger variance across sequences but little variance within sequences.

The variational lower bound for this inference model

over the marginal likelihood of $\mathbf{X}$ is given as,

$$\log p_\theta(\boldsymbol{X}) \geq \mathcal{L}(\theta, \phi; \boldsymbol{X}) = \sum_{n=1}^{N} [\mathcal{L}(\theta, \phi; \boldsymbol{x}^n | \tilde{\boldsymbol{\mu}}_c^n, \tilde{\boldsymbol{\mu}}_e, \tilde{\boldsymbol{\mu}}_s) + \log p_\theta(\tilde{\boldsymbol{\mu}}_c^n)] + \log p_\theta(\tilde{\boldsymbol{\mu}}_e) + \log p_\theta(\tilde{\boldsymbol{\mu}}_s) + const \quad (3)$$

Please refer to the supplementary material for proofs and a more detailed explanation of this section.

Discriminative Objective It is possible for the priors to learn trivial values for all sequences ($\tilde{\boldsymbol{\mu}}_{\boldsymbol{c}}^i = 0$, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{e}}^i = 0$, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{s}}^i = 0 \; \forall i$) and still maximize the variational lower bound described above. To ensure that the priors are indeed discriminative and characteristic of the variations they encode, we introduce a discriminative objective function that infers *variation-specific* annotation from the corresponding representation. For instance, we enforce the content latent variable $\boldsymbol{z}_c^{i,n}$ for audio segment $\boldsymbol{x}^{i,n}$ to correctly infer its viseme annotation $v^{i,n}$ through classification loss given by,

$$\log p(v^{i,n} | \boldsymbol{z}_c^{i,n}) = \log p(\boldsymbol{z}_c^{i,n} | v^{i,n}) - \log \sum_{j=1}^{V} p(\boldsymbol{z}_c^{i,n} | v^{i,j}) \quad (4)$$

where V is the set of all viseme labels [13]. We similarly enforce discriminative objective over emotion and sequence latent variables to correctly predict the emotion and sequence id associated with the audio sequence.

Margin Ranking Loss To enable effective mapping of the audio content with the facial features and minimize ambiguity, we need to separate out the content from the rest of the factors of variations as much as possible. So we need to ensure that $\boldsymbol{z}_c$, $\boldsymbol{z}_e$ and $\boldsymbol{z}_s$ are as decoupled as possible. The discriminative objectives over the different latent variables ensure that they capture well their respective factors of variations (content, emotion and global sequence variations respectively) but to really disentangle them, we want to make them agnostic to other variations by having them perform badly on other classification tasks (that is, content variable $\boldsymbol{z}_c$ perform poorly in predicting the correct emotion associated with the audio sequence). To this end, we introduce margin ranking losses $\mathcal{T}$ with margin γ on the softmax probability scores of the viseme label for $\boldsymbol{z}_c$ with $\boldsymbol{z}_s$ and $\boldsymbol{z}_e$ given by,

$$\mathcal{T}(v^{i,n}, \boldsymbol{z}_c^{i,n} n, \boldsymbol{z}_e^{i,n}, \boldsymbol{z}_s^{i,n}) = \max\left(0, \gamma + \mathcal{P}(v^{i,n} | \boldsymbol{z}_s^{i,n}) - \mathcal{P}(v^{i,n} | \boldsymbol{z}_c^{i,n})\right) + \max\left(0, \gamma + \mathcal{P}(v^{i,n} | \boldsymbol{z}_e^{i,n}) - \mathcal{P}(v^{i,n} | \boldsymbol{z}_c^{i,n})\right) \quad (5)$$

where $\mathcal{P}(v^{i,n}|.)$ denotes the probability of $v^{i,n}$ given some latent variable. Margin ranking loss widens the inference gap, effectively making only $\boldsymbol{z}_c$ learn the content relevant features. We similarly introduce margin ranking loss on probability scores for emotion label to allow only $\boldsymbol{z}_e$ learn emotion relevant features.

Equation 3 suggests that the variational lower bound of an audio sequence can be decomposed into the sum of variational lower bound of constituent segments. This provides scalability by allowing the model to train over audio segments instead. As shown in Figure 1, the input to the VAE is audio segments each having T time points. Based on the inference model in Equation 2, these segments are first processed by LSTM-based content and emotion encoders, and later by sequence encoder (along with other latent variables). All the latent variables are then fed to the decoder to reconstruct the input. The final segment based objective function to maximize is as follows,

$$\mathcal{L}^F(\theta, \phi; \boldsymbol{x}^{i,n}) = \mathcal{L}(\theta, \phi; \boldsymbol{x}^{i,n}) - \beta[\mathcal{T}(e^i, \boldsymbol{z}_c^{i,n}, \boldsymbol{z}_e^{i,n}, \boldsymbol{z}_s^{i,n}) + \mathcal{T}(v^{i,n}, \boldsymbol{z}_c^{i,n}, \boldsymbol{z}_e^{i,n}, \boldsymbol{z}_s^{i,n})] + \alpha[\log p(i | \boldsymbol{z}_s^{i,n}) + \log p(v^{i,n} | \boldsymbol{z}_c^{i,n}) + \log p(e^i | \boldsymbol{z}_e^{i,n})] \quad (6)$$

where α and β are hyper-parameter weights.

3.2. Talking Head Generation

We use adversarial training to produce temporally coherent frames animating a given face image conditioned on the content representation $\boldsymbol{z}_c$ as shown in Figure 2.

3.3. Generator

Let G denote the generator function which takes as input a face image I_f and sequence of audio-based content representations $\{\boldsymbol{z}_c^n\}_{n=1}^N$ sampled from $\boldsymbol{Z}_c$ given an audio sequence $\boldsymbol{X} = \{\boldsymbol{x}^n\}_{n=1}^N$ having N audio segments. G generates a frame O_f^n for each audio segment $\boldsymbol{x}^n$. Each $\boldsymbol{z}_c^n$ is combined with the input image by channel-wise concatenating the representation after broadcasting over the height and width of the image. The combined input is first encoded and then decoded by G which has a U-Net [27] based architecture to output a video frame with the face modified in correspondence to the speech content. For temporal coherency between consecutive generated video frames, we introduce temporal smoothing similar to [5] by making G generate frames in an auto-regressive manner. We employ L1 loss along with perceptual similarity loss and L2 landmark distance (mouth region) as regularization.

3.4. Discriminator

We incorporate WGAN-GP [17] based discriminators which act as critic to evaluate the quality of the generated frames/videos. We introduce a frame-level discriminator D_{frame} which computes the Wasserstein distance of each

individual generated frame conditioned on the input content representation. The architecture of D_{frame} resembles that of PatchGAN [20]. D_{frame} is designed to behave as a multi-task critic network. It also evaluates the conditioning between the generated frame and content representation through an auxiliary classification network that predicts the correct viseme corresponding to the conditioned audio segment (content representation). The loss for this auxiliary network is given by cross-entropy loss over the set of viseme labels.

We introduce a video-level discriminator D_{video} similar to [30] to enforce temporal coherency in the generated video. The architecture of D_{video} is similar to D_{frame} without the auxiliary viseme classification network and has a 3D convolutional architecture with time representing the third dimension. It takes as input a set of frames (real or generated) along with corresponding content representations (concatenated channel wise) and evaluates the Wasserstein distance estimate over the video distribution. By doing so, D_{video} evaluates the difference in realism and temporal coherence between the distribution of generated sequences and real sequences.

4. Experiments

4.1. Datasets

GRID [10] is an audiovisual sentence corpus with high-quality recordings of 1000 sentences each from 34 talkers (18 male, 16 female) in a neutral tone. The dataset has high phonetic diversity but lacks any emotional diversity.

CRowdsourced Emotional Multimodal Actors Dataset (CREMA-D) [4] consists of 7,442 clips from 91 ethnically-diverse actors (48 male, 43 female). Each speaker utters 12 sentences in 6 different emotions (Anger, Disgust, Fear, Happy, Neutral, Sad).

Lip Reading Sentence 3 (LRS3) Dataset [2] consists of over 100k spoken sentences from TED videos. We use this dataset to test our method in an 'in-the-wild' audio-visual setting. Previous approaches have experimented with LFW [8] which is a precursor to LRS3 dataset.

4.2. Training

We use speech utterances from GRID and CREMA-D for training the VAE to learn disentangled representations. We divide the dataset speaker-wise using train-val-test split of 28-3-3 for GRID and 73-9-9 for CREMA-D. We first pre-train the content pipeline of the VAE using GRID (which provides the phonetic diversity) and then, use the learned weights to initialize the training of the entire VAE using CREMA-D (which provides the emotional diversity). To obtain the viseme annotations, we use Montreal Forced Aligner to extract phoneme annotation for each audio segment and then categorize them into 20 viseme groups (+1 for silence) based on [32]. Emotion labels are readily available from CREMA-D dataset for 6 different emotions. We label each audio sequence from GRID having neutral emotion.

We use a setup similar to [19] for training the VAE. Every input speech sequence to the VAE is represented as a 200-dimensional log-magnitude spectogram computed every 10ms. Since the length of a syllabic segment is of the order of 200ms, we consider $\boldsymbol{x}$ to be a 200ms segment implying $T = 20$ for each $\boldsymbol{x}$. We use 2-layer LSTM for all encoders and decoder with hidden size of 256. Based on hyperparameter tuning, we set the dimensions for $\boldsymbol{z_c}, \boldsymbol{z_e}$ and $\boldsymbol{z_s}$ to 32, and the variance of priors to 1 and latent variables to 0.25. α, β and margin γ are set to 10, 1 and 0.5 respectively. For generating talking head, we use GRID and LRS3 dataset. All faces in the videos are detected/aligned using [3] and cropped to 256×256. Adam optimizer is used for training in both stages, and learning rate is fixed at 10^{-3} for VAE and 10^{-4} for GAN.

4.3. Robustness to Noise

We evaluate the quality of the generated videos using Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM). Higher the value of these metrics indicate better overall video quality. We further use Landmark Distance (LMD) (similar to [6]) to evaluate the accuracy of the mouth movement in the generated videos. LMD calculates the Euclidean Distance between the mouth landmarks as predicted by the landmark detection model [3] of the original video and the generated video.

$$LMD = \frac{1}{F} \times \frac{1}{L} \sum_{f=1}^{F} \sum_{l=1}^{L} ||P_{f,l}^{real} - P_{f,l}^{fake}||_2$$

where F denotes the number of frames in the video, L denotes the number of mouth landmarks, and $P_{f,l}^{real}$ and $P_{f,l}^{fake}$ represents the landmark coordinates of the l^{th} landmark in f^{th} frame in the original and generated video respectively. Lower LMD denotes better talking head generation.

To test the robustness of our approach to noise, we create noisy samples by adding uniformly distributed white noise to audio sequences. We experiment with different noise levels by adjusting the loudness of the added noise compared to the original audio. A noise level of -40dB means that the added noise is 40 decibels lower in volume than the original audio. -10dB refers to high noise (almost imperceptible speech), -30dB refers to moderate (above average background noise) and -60dB refers to low noise (almost inaudible noise).

Table 1 shows the landmark distance estimates for different approaches over different noise levels. We re-

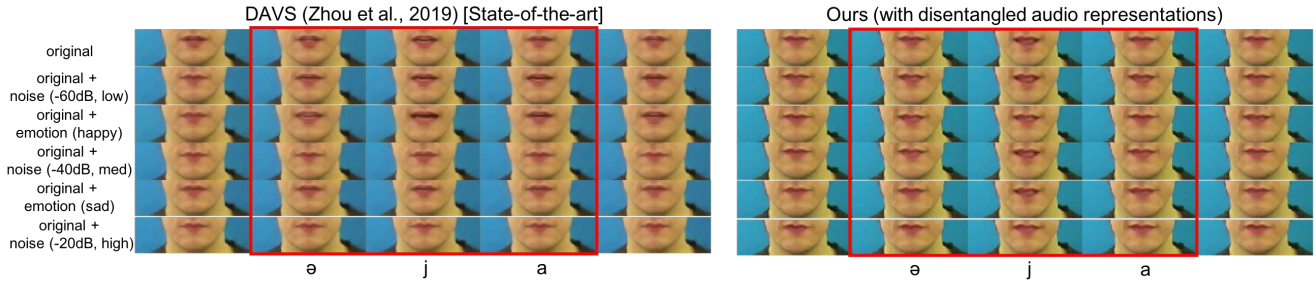

Figure 3. Visual comparison over different methods for different speech variations. If we look at the frames highlighted in the red box, we can observe how the introduction of noise or emotion reduces the performance/consistency of the current state-of-the-art while our approach is robust to such changes. Sentence: Don't forget a jacket. Symbols at the bottom denote syllables.

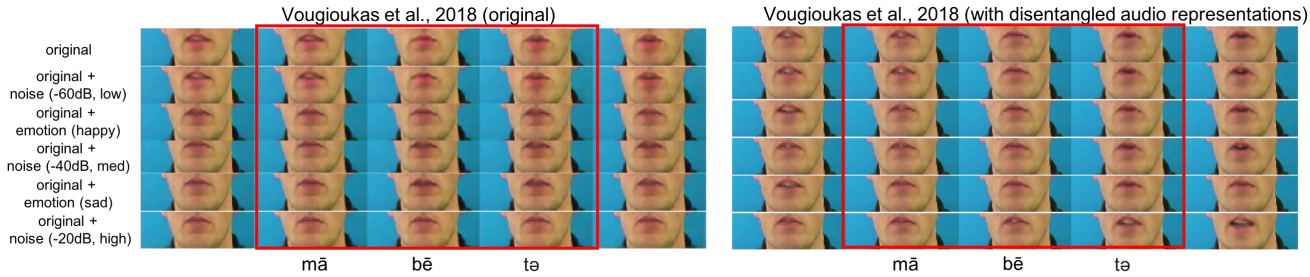

Figure 4. Visual comparison showing the ease of using our disentangled audio representation with existing talking head approaches to improve robustness to speech variations. Sentence: Maybe tomorrow it'll be cold. Symbols at the bottom denote syllables.

Method	GRID				LFW/LRS3			
	Clean	Low -60dB	Med -30dB	High -10dB	Clean	Low -60dB	Med -30dB	High -10dB
[30] (original)	1.32	1.40	1.96	2.87	1.81	1.79	2.56	2.92
[30] (w/ ours)	1.33	1.34	**1.45**	2.71	1.83	1.82	**1.98**	2.73
DAVS [31]	1.21	1.28	1.67	2.56	1.64	1.65	2.1	2.76
Baseline	1.36	1.34	1.73	2.80	1.89	1.85	2.63	2.84
Baseline + Augmentation	1.35	1.38	1.48	2.79	1.87	1.85	1.94	2.81
Baseline (DeepSpeech)	1.31	1.31	1.53	2.84	1.7	1.75	2.05	2.90
Ours (w/o Margin Loss)	1.28	1.26	1.46	2.7	1.65	1.63	1.87	2.81
Ours	1.25	1.27	**1.33**	2.62	1.67	1.66	**1.79**	2.80

Table 1. Comparison of different approaches for audio samples with different noise levels.

implemented [30] and used the public available model for DAVS [31] for obtaining and comparing the results. From the table, we can observe that for low noise levels, the performance of all the approaches is comparable to that for clean audio. But there is a significant rise in the landmark distance for [30] and DAVS as the noise levels become moderately high. While on the other hand, it is in this part of the noise spectrum where our approach excels and significantly outperforms the current state-of-the-art by maintaining a value comparable to clean audio. Clearly, by distentangling content from the rest of the factors of variations, our model is able to filter out most of the ambient noise and allow conditioning the video generation on a virtually cleaner signal. We observe that when the noise levels become exceedingly high, even our approach is unable to maintain its performance. We believe that such high noise levels completely distort the audio sequence leaving nothing meaningful to be captured and since we neither do any noise filtering nor use noisy samples for training explicitly, it is likely for the model to not perform well on almost imperceptible speech. Figure 5 further shows a trend in the landmark distance for increasing noise levels. From the graph in Figure 5, we can observe that the performance of our approach becomes relatively better with increasing amounts of noise up to a reasonable level.

Figure 3 shows a visual comparison of our approach with DAVS for different audio variations. We can notice for -40dB noise level, the mouth movement for DAVS begins to lose continuity with abrupt changes in the mouth movement (quick opening and closing of mouth) unlike for clean audio. By -20dB noise level, the mouth stops opening altogether. On the contrary, our method is much more resilient with mouth movement for -40dB noise level being almost identical to clean audio and for -20dB being only a bit abrupt.

We also show results of our approach on clean audio in Figure 6. Moreover from Table 2, we can observe that for clean neutral spoken utterances, our approach performs at par with other methods on all metrics.

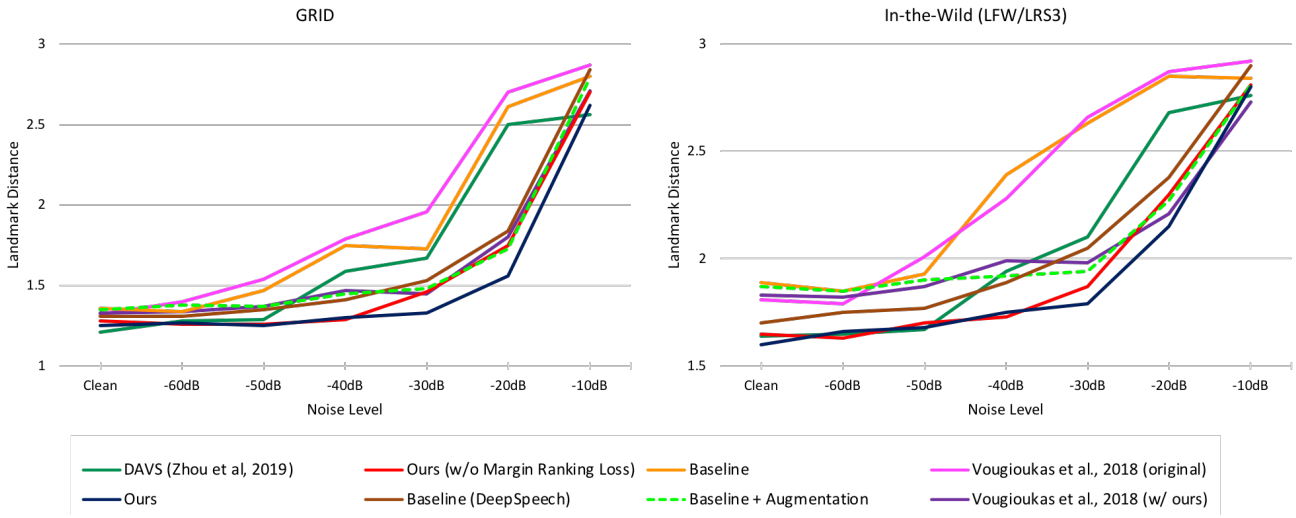

Figure 5. Plot for landmark distance comparison between different methods for different noise levels. Lower means better.

Method	GRID			LRW/LRS3		
	LMD	PSNR	SSIM	LMD	PSNR	SSIM
[30]	1.32	28.88	0.81	1.81	28.49	0.71
[7]	1.35	29.36	0.74	2.25	28.06	0.46
[6]	1.18	29.89	0.73	1.92	28.65	0.53
[31]	1.21	28.75	0.83	1.64	26.80	0.88
Ours	1.25	30.43	0.78	1.67	29.12	0.73

Table 2. Comparison with previous approaches on widely used metrics for original (clean) audio samples.

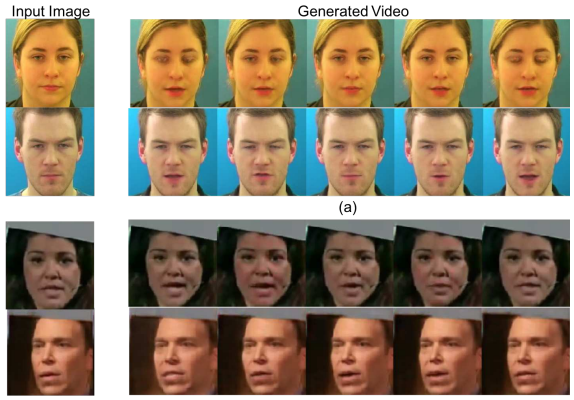

Figure 6. Sample results on (a) GRID and (b) LRS3 dataset for different speakers using clean audio samples.

4.4. Robustness to Emotion

We test the robustness of the disentangled representations to emotional variations by generating talking head for emotionally rich audio sequence from CREMA-D dataset. Due to this cross generation, we can only do a qualitative analysis as shown in Figure 3. We compare the talking head videos generated by our method with DAVS on different emotions. Looking at the frames in the red box, we can observe that although the performance of DAVS for loud emotions like happy is as good as for neutral, the mouth movement becomes abrupt and weak for soft emotions such as sad. On the contrary, our method is able to perform consistently over the entire emotional spectrum as evident from almost similar visual results for different emotions.

4.5. Ease of Compatibility

Our model for learning emotionally and content aware disentangled audio representations is compatible with any of the current state-of-the-art approaches for talking head generation, and can be used in conjunction to improve robustness to factors of audio variations. We demonstrate this by implementing [30] using the content representation from our VAE model in place of that learned by the audio encoder. Table 1 shows a comparison of the landmark distance between the two implementations for different noise levels. Similar to above, we can infer that using a filtered out content representation allows the model to perform significantly better than the original implementation in the presence of moderately high levels of noise. From Figure 5, we can observe that the trend for the 'hybrid' implementation is quite similar to our own implementation. Figure 4 further compares the two implementations qualitatively for both noise and emotional audio. We can observe that [30] using our disentangled representations performs much more consistently than the original implementation. Due to the unavailability of training code/resources, we were unable to test our model with other approaches. But above demonstration proves that our disentangled representations can be easily incorporated with any existing implementation.

4.6. Ablation study

We conduct an ablation study to quantify the effect of each module in our approach. We run a baseline experiment where we replace our disentangled audio representation with a generic network which learns directly from MFCC features similar to [9]. As can be seen from Table 1, the baseline performs poorly for noisy samples. This clearly suggests that simple audio representation is not robust to audio variations while generating talking heads.

We introduce a second baseline where we further perform aggressive augmentation of the input audio in the aforementioned baseline of learning directly from MFCC features. Figure 5 and Table 3 show the results of these experiments (labeled Baseline + Augmentation). We observe that the landmark distance estimates are consistently better than the baseline without augmentation. However, these results are still noticeably worse than results of our approach. Data augmentation does make a difference over using normal dataset, however, we believe that simply relying on augmented data for training is not efficient enough as it is very challenging to augment the 'right' noise for the trained model to generalize well for real scenarios.

To further test the effectiveness of the representation, we perform another baseline experiment where we replace the disentangled content features with speech features extracted from robust automatic speech recognition (ASR) model, DeepSpeech [18]. Since [11] shows the noise robustness of DeepSpeech features while generating relative low-dimensional offsets of a 3D face mesh given an audio input, we wish to test their potential in generating in a visual space which is orders of magnitude higher in dimension. As shown in Figure 5 and Table 3, we find these speech features are not as effective as our disentangled audio representation for talking face generation. We believe the difference in performance is because the feature embedding from robust ASR models such as DeepSpeech is essentially a point embedding which, because of being oriented towards solving a discriminative task, loses a lot of key information about the variations in audio and can even be incorrect. Since we use a VAE, our content representation is modeled instead as a distribution which preserves these subtle variations by making it reconstruct the audio while aligning with audio content at the same time. This dual benefit (balance), which ASR models cannot offer, makes our content representation a much more informative and robust input for a high-dimensional generative task of face animation.

In addition to learning a factorized audio representation, we also ensure an increased decoupling of the different representations by enforcing margin ranking loss as part of the training objective. Decoupling is essential to allow different audio variations to be captured exclusively by the designated latent variable which in turn helps in distilling the content information for improved robustness to variations.

Representation	With Margin Ranking Loss		Without Margin Ranking Loss	
	Viseme	Emotion	Viseme	Emotion
Content	**77.1**	24.5	58.7	37.0
Emotion	29.8	**68.4**	35.4	55.3

Table 3. Accuracy (%) over viseme and emotion classification task by disentangled content and emotion representations.

To prove the importance of margin ranking loss, we evaluated the landmark distance metric of the model trained without margin ranking loss. From Figure 5 and Table 3, we can conclude that margin loss makes the approach robust to higher levels of noise, For GRID dataset, although for -40dB noise, the results for with/without margin ranking loss are comparable, there is a noticeable gap for -30dB noise level. Similar trend can also be observed for LRS-3/LFW dataset. We believe that although there is some level of disentanglement without margin ranking loss, when the audio is noisier, we need stronger disentanglement to produce more clear content representation which is possible due to margin ranking loss. To further quantify the effectiveness of margin ranking loss in decoupling, we train auxiliary classifiers over the content and emotion representations for the task of viseme and emotion classification. As shown in Table 3, it is clearly evident that introduction of margin ranking loss makes the latent representation perform badly on tasks other than the designated task. In fact, it not only widens the performance gap between the representations for a particular task, but it also facilitates the designated representation to perform better than without margin ranking loss.

5. Conclusion and Future Work

We introduce a novel approach of learning disentangled audio representations using VAE to make talking head generation robust to audio variations such as background noise and emotion. We validate our model by testing on noisy and emotional audio samples, and show that our approach significantly outperforms the current state-of-the-art in the presence of such audio variations. We further demonstrate that our framework is compatible with any of the existing talking head approaches by replacing the audio learning component in [30] with our module and showing that it is significantly robust than the original implementation. By adding margin ranking loss, we ensure that the factorized representations are indeed decoupled. Our approach to *variation-specific* learnable priors is extensible to other speech factors such as identity and gender which can be explored as part of future work.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. *CoRR*, 2018.

[2] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[3] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[4] 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*, 5(4):377–390, 2014.

[5] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018.

[6] 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)*, pages 520–535, 2018.

[7] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *BMVC*, 2017.

[8] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[9] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[10] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 2006.

[11] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] P. Edwards, C. Landreth, E. Fiume, and K. Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 35(4):127, 2016.

[13] P. Edwards, C. Landreth, E. Fiume, and K. Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 35(4):127, 2016.

[14] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, 2018.

[15] B. Fan, L. Wang, F. K. Soong, and L. Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

[16] 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*, pages 2672–2680, 2014.

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

[18] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[19] W.-N. Hsu, Y. Zhang, and J. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, 2017.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[21] S. A. Jalalifar, H. Hasani, and H. Aghajan. Speech-driven facial reenactment using conditional generative adversarial networks. *arXiv preprint arXiv:1803.07461*, 2018.

[22] 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)*, 36(4):94, 2017.

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[24] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[25] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[26] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo. Video-audio driven real-time facial animation. *ACM Trans. Graph.*, 34(6):182:1–182:10, Oct. 2015.

[27] 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*, pages 234–241. Springer, 2015.

[28] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[29] 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)*, 36(4):93, 2017.

[30] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018.

[31] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[32] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics*, 37(4), 2018.