VENTRILOQUIST-NET: LEVERAGING SPEECH CUES FOR EMOTIVE TALKING HEAD GENERATION

Deepan Das, Qadeer Khan, Daniel Cremers

Computer Vision and Artificial Intelligence, Department of Informatics
Technical University of Munich, Garching, Germany

ABSTRACT

In this paper, we propose *Ventriloquist-Net:* A network for Talking Head Generation using only a speech segment and a single source image of a person. It places emphasis on *emotive expressions*. Cues for generating expressions are directly inferred from the speech clip. We formulate our framework to comprise of independently trained modules focusing on each of the aforementioned aspects. This not only expedites convergence but also facilitates handling in-the-wild source images. Quantitative and qualitative evaluations on generated videos demonstrates state-of-the-art performance even on unseen input data.

Index Terms— Talking Head Generation, Speech Emotion

1. INTRODUCTION

Talking Head Generation refers to the process of animating the image of a person's face according to an input speech clip. It has recently attracted a lot of attention because of its wide variety of use-cases [1]. It can be used to produce animated content in short turnaround times or to animate avatars for virtual assistants [2]. Other uses are to increase the compression factor for video-conferencing without loss of quality [3] or to edit target segments of a video in terms of spoken content [4, 5], etc. The ultimate goal is to generate a video of a person that not only matches the spoken words but also includes naturalistic head movements, facial expressions in keeping with the audio clip.

Early works, like [6, 7], trained models tuned to a single subject and hence could not scale to unseen identities. Later works of [8, 9, 10, 11, 12] developed speaker-independent pipelines by disentangling identity and speech features. However, they focused mainly on achieving accurate lip-sync, ignoring all other aspects that make the Talking Head more naturalistic.

More recently [13, 14] utilize a temporal discriminator teaching the model to produce facial motions like eye blinks. Likewise, [15] deployed a cascaded GAN to include rhythmic head movements and eye blinks giving improved realism in the final output. However, it could not capture well the appropriate facial expressions that matches the tone of the audio clip. One of the most significant states-of-art in this area was [16]. Here, a speaker-aware subnetwork modeled speech mannerisms with longer time-dependencies from a small set of speaker identities. Despite photo-realistic outputs containing eye-blinks, the exhibited head movements were barely noticeable. The other such work was [17]. Their 2 major contributions were to use contrastive loss for mapping speech features to lip motion and to have modular control over disentangled identity, posture and speech features. While generating mouth movements from the speech of one video, they could blend in the head postures of a different video.

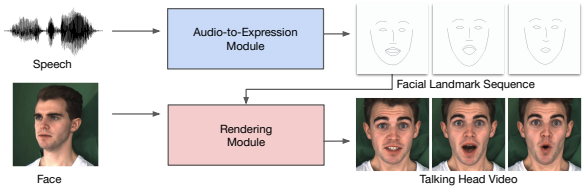


Fig. 1: Schematic of the proposed Ventriloquist-Net, which takes a speech clip and a single source image as inputs and aims to generate a naturalistic, emotive face video.

However, a limitation was that a secondary video of length comparable to the original speech clip was needed to be able to use pose-control. Moreover,the pose sequence might not match the mood, energy etc. of the speech clip. Their fixed-pose version produced very photo-realistic but rigid output videos. All the above one-shot approaches ignored realism in terms of emotional expression. In contrast, our framework is capable of generating emotive expressions in the final video output by implicitly inferring cues from the audio clip. This approach prevents the risk arising from explicitly forcing an emotion that may contradict the tone of the audio clip.

Earlier works such as [18] also focused on emotional expressions. However, their dependence on handcrafted audio features and manually annotated visual features limited scalability. The authors of [2] developed a method producing very realistic outputs with learned subject-specific head movements. However, the model required $\approx$ 20 hours to learn the mannerisms from a 2-3 minute video of every unseen identity. [19] trained a model for generating realistic emotional Talking Heads, but it was person-specific . Moreover, the model needed a target emotion label too as input, instead of inferring the emotion directly from the speech.

We propose Ventriloquist-Net as a one-shot, subject-independent Talking Head Generation model. As shown in Figure 1, it only needs a single speech clip and a single face image of a person as inputs to generate a lip-synced re-animation. In this regard, our primary contributions are:

- Our network uses the emotional content, inferred from the speech without any additional input, to generate expressions and head motions that match the mood, tone, energy etc. of the audio clip.
- We formulate our framework to comprise of independently trained modular components. The two main advantages are that it (1) allows semi-supervised training on datasets which do not provide emotion labels; (2) expedites the model convergence by disentangling gradient flows with potentially competing interests.
- Loss functions are designed not only to reflect our primary objectives but also to stabilize GAN training and prevent mode collapse.

2. PROPOSED METHOD

Our model has 2 major independently trained components, shown in Fig 1. The Audio-to-Expression module converts an input audio into a facial landmark sequence, while incorporating emotive expressions matching the speech. The Rendering module converts the previous sequence into a video, using the source face image. The modules are made independent for increased stability during training. It can also provide flexibility during inference, by giving the user an option to manipulate the intermediate landmarks as required (eg. head re-orientation).

Before discussing the modules in more detail, we first introduce the data modalities recurring throughout the work. These can also be seen in Figures 2 and 3. We primarily use the MEAD dataset [19] designed for emotional Talking Head Generation. It records actors in studio settings uttering phoneme-rich sentences with simulated emotions. We adopt the approach of [20] to extract $k(t)$, denoting the sequence of 2D positions of facial keypoints k varying over time t. To extract speech features, raw input audio is first converted into Mel spectrograms ($s(t)$) and Mel Frequency Cepstral Coefficients ($m(t)$). A spectrogram projects the audio to a low-dimensional space compatible with CNNs, while retaining sufficient information for predicting mouth shapes. This is chosen because CNNs were found to perform better for sequential generation tasks, where RNNs were plagued by gradient issues.

2.1. Joint Pre-training of Emotion-related Subnetworks

Emotion-related subnetworks refer to the CNN-based Speech-to-Emotion Embedder ($Sp2Emo$) and the biLSTM-based Expression-to-Emotion Cross Embedder ($Exp2Emo$) that belong to the Audio-to-Expression module. They are shown in details in Figure 2.

$Sp2Emo$ processes $m(t)$ to produce an embedding vector ($q_{emo}(t)$) and a probability distribution over emotion labels (l_{emo}). Cross Entropy loss is used to pre-train it against ground truth emotion labels. $Exp2Emo$ embeds $k(t)$ to p_{emo}, which lies in the same vector space as l_{emo}. Cosine Embedding Loss between p_{emo} and l_{emo} supervises this subnetwork's pre-training. This loss term results in $Sp2Emo$ and $Exp2Emo$ learning to agree on the predicted probability distribution over the emotion classes, instead of simply agreeing on the most likely class. The latter is insufficient, since any emotion is a composite of the basic emotions [21].

These subnetworks need to be pre-trained prior to training the entire GAN. The reason is that even the slightest ambiguity in the loss functions can lead to the training process collapsing. Otherwise, the generator that is supposed to focus on lip motions/shapes will receive gradients from a loss that was the result of $Sp2Emo$'s prediction error (and vice-versa). This cross-contamination leads to the overall model trying to optimize over multiple loss functions that are at cross-purposes. As an example of such conflict, the expression generator branch is encouraged to generate dynamic, energetic motions and expressions so that the final output looks natural. In contrast, the mouth shape generator branch is discouraged from generating random, energetic motions. It is encouraged to converge to correct, stable lip movements.

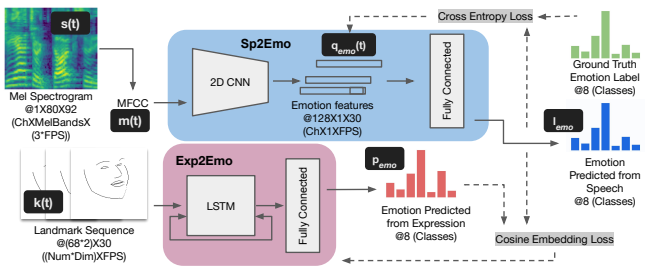


Fig. 2: Detailed view of emotion subnetworks $Sp2Emo$ and $Exp2Emo$. Cross Entropy loss is computed between ground truth emotion labels and l_{emo}, while Cosine Embedding is used between l_{emo} and p_{emo}. Dotted lines link losses to the modules where they back-propagate.

2.2. Emotive Generator and Semi-supervised Training Scheme

The remainder of the Audio-to-Expression module contains the Emotive Generator G, the Sequence Discriminator D and additional losses, as shown in Figure 3. G processes $s(t)$ through CNNs to produce intermediate lingual features ($q_{ling}(t)$). The latter stage of the network then takes both $q_{ling}(t)$ and $q_{emo}(t)$ through their respective 1D CNN branches to generate landmark deformations. These are added through frozen weights for the final output. Frozen weights ensure that the branch processing $q_{ling}(t)$ can only affect the lip and jaw landmarks, while the other branch mostly affects other landmarks like eyes, eyebrows etc (while slightly affecting lips too, eg. for smiling). This also helps separate gradient back-propagations from lip-sync errors and expression errors into their respective branches. Also, inputs varying with time t ensure dynamism in the output.

G generates not the facial landmark sequence itself but rather the divergence of the landmarks ($\delta\hat{k}(t)$) from a default position ($\bar{k}$, pre-computed from the training split of MEAD). The final predicted landmarks are given by $\hat{k}(t) = \bar{k} + \delta\hat{k}(t)$. An advantage of predicting the deformation (instead of the actual landmarks) is that the generator does not need to learn the face shape before it can start predicting landmark motions. This also removes the requirement of a frame-level discriminator. Since the deformation magnitude can be controlled via weight initialization and losses, the predicted output will always be valid face landmarks.

Because of the presence of the pre-trained $Exp2Emo$ subnetwork, an explicit target emotion label is not required for training G. The generator uses emotion features provided by $Sp2Emo$ to generate emotional expressions, while the resulting expressions are then 'checked' by $Exp2Emo$ to make sure they match the speech emotion. The corresponding gradient flows are visualized in Figure 3. Moreover, the pre-trained and frozen $Sp2Emo$ and $Exp2Emo$ do not need to be 'correct' (i.e. high accuracy emotion classifica-

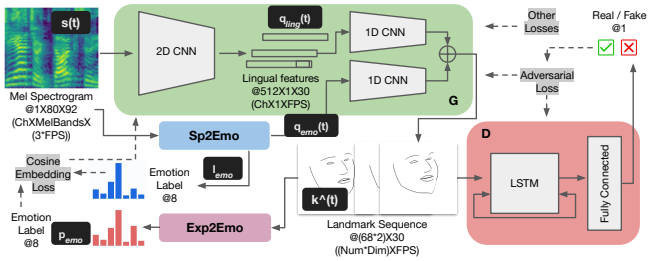


Fig. 3: Detailed view of the complete network showing G and D. G uses both lingual features ($q_{ling}(t)$) from spectrogram and emotional features ($q_{emo}(t)$) from $Sp2Emo$. Dotted lines show that Adversarial loss alone trains D, while that and other losses update G. Any 'disagreement' between emotion predicted from input speech (l_{emo}) and emotion predicted from generated expression (p_{emo}) is captured in Cosine Embedding loss and back-propagated to G.

tion models) as long as they are in agreement. In other words, the emotion class acts like a latent variable in this training scheme and becomes obfuscated. The correctness in classifying speech emotions and expressions individually does not affect the model performance. Rather, a consistency in matching similar 'heard' emotions to similar 'seen' expressions (even if misclassified) makes G's outputs consistent. Note that some supervision (from ground truth emotion labels) is still required during pre-training so that the subnetworks can learn useful features.

By extending this idea, it can be seen that once the emotion-related subnetworks are trained on a smaller emotion dataset, they can then provide supervision for training the full GAN on datasets where emotion labels are not provided. This addresses the issue of training on popular Talking Head Generation datasets, like VoxCeleb2 [22], which do not provide ground truth for emotion. Another problem often encountered is that, when datasets like MEAD do provide emotion labels, the emotions are emulated by actors. This is less accurate than the emotions/expressions exhibited in the in-the-wild settings of VoxCeleb2. However, since our model's performance does not depend on the accuracy of emotion classification, it can train on in-the-wild emotions despite pre-training on acted emotions.

2.3. Discriminator and Additional Losses

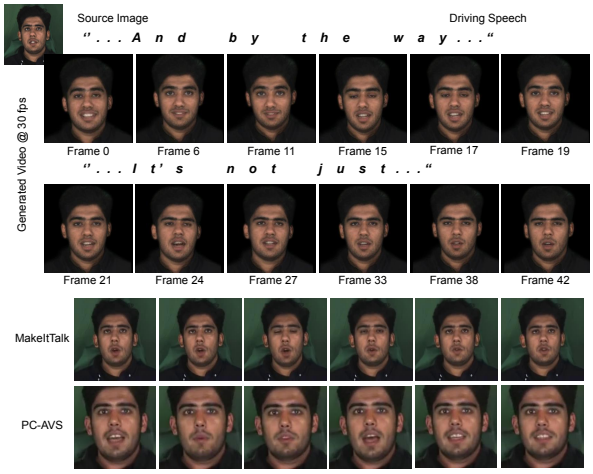
The discriminator D uses biLSTMs to estimate whether the entire sequence $\delta\hat{k}(t)$ is real or not. It trains adversarially with G. By looking at sequences rather than at individual frames, D picks up errors that manifest slowly over time. Examples are absence of eye-blinks, too much or too little motion of facial keypoints, static head, infeasible keypoint locations in a certain frame, etc.

$\mathcal{L}_{emotion}$: $Exp2Emo$ acts as a second discriminator to check $\hat{k}(t)$. Cosine Embedding Loss over expression is formulated as $\mathcal{L}_{emotion} = \mathcal{CE}(Exp2Emo(\hat{k}(t)), l_{emo})$. It encapsulates the mismatch between the emotion 'heard' ($Sp2Emo$ on input speech) and the emotion 'seen' ($Exp2Emo$ on generated expression).

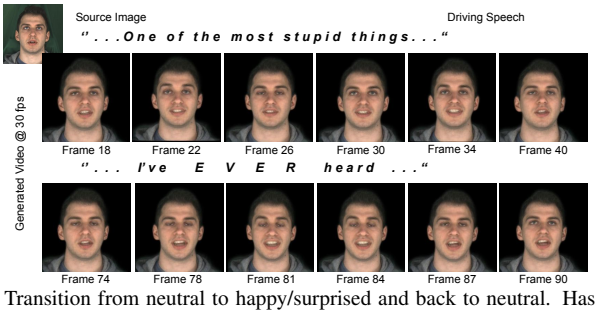
$\mathcal{L}_{mouth}$: L1-loss is used to compare outputs of the lip-shape generating branch with ground truth mouth shapes. An advantage over L-2 loss is that the latter produces smaller gradients the closer a value gets to 0. This implies the mouth may not end up closing all the way (given by $\delta\hat{k}^{mouth}(t) \rightarrow 0$ for some $t = T$, where k^{mouth} includes only lip and jaw landmarks). The loss can be expressed as $\mathcal{L}_{mouth} = ||\delta k^{mouth}(t) - \delta\hat{k}^{mouth}(t)||_1$. However, different speakers uttering the same words will have different mouth shapes and sizes, despite producing similar sounds. An L-norm loss might make the model collapse. This will appear as a 'non-elastic' mouth or rigid mouth in the final output video, showing a limited range of motion. Such an eventuality is countered by weighing the losses appropriately, by using D and with the help of the next loss term.

$\mathcal{L}_{energy}$: L-2 loss on entropy over all landmark points encourages the network to produce dynamic head motions, dynamic expressions and non-static mouth landmarks. G is penalized for not matching the energy perceived in the ground truth, thus preventing both overly static and overly jittery outputs (static implies $\delta\hat{k}(t) \approx 0$). The loss term is $\mathcal{L}_{energy} = ||\mathbb{E}_t|\delta k(t)| - \mathbb{E}_t|\delta\hat{k}(t)|||_2 + ||\mathbb{E}_t|\delta k'(t)| - \mathbb{E}_t|\delta\hat{k}'(t)|||_2$. Here k' signifies derivative of $k \in \mathbb{R}^{136}$ with respect to time, whereas $\mathbb{E}_t$ is expectation over time t.

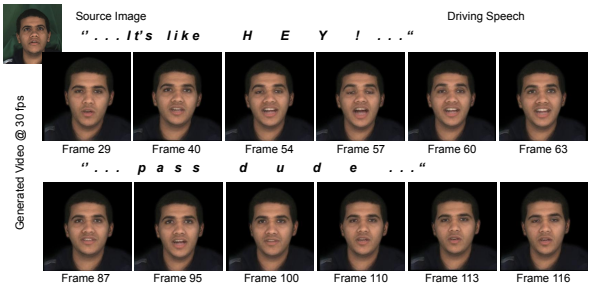
For the Rendering module (Figure 1), we use a publicly available model [23]. It was trained to morph a source face image according to the orientation, expression etc. given by a set of input face landmarks. Its design was made to preserve identity and texture information from the source face in the generated video frames.



(a) Happy expression smoothly transitions into slightly angry or disgusted expression as the speech emotion also shifts. Shows blink, slight changes in the direction in which the subject is looking, head movements (nodding). In comparison, while PC-AVS [17] produces sharper lip motions and high synchronisation confidence, their faces are completely static. In case of MakeItTalk [16], faces are less static but still devoid of changes in expression.



(b) Transition from neutral to happy/surprised and back to neutral. Has an elongated eye blink, accompanying the emphasis on "ever".



(c) Shifts from neutral to surprise or happy, again to neutral and mild disgust. Shows blink. Frame 57 shows a larger mouth opening shaping "Hey" with emphasis. Frame 100 shows the mouth shape corresponding to the sibilant in "pass", while subsequent frames show the lip shape changing to form "dude".

Fig. 4: Video samples generated by our model using a single unseen source image and an unseen driving speech clip, along with transcripts. The generated expressions are in tandem with the perceived emotions in the audio channel. Eye blinks and natural head movements are also generated. For the first example, a comparison with states-of-art is also shown. More video examples are available here.

3. RESULTS

3.1. Training Details

The emotion subnetworks are trained on a 80% subject-level train split of MEAD, augmented by CREMA-D, RAVDESS, SAVEE and TESS datasets [24, 25, 26, 27]. $Sp2Emo$ has an initial learning

rate of 10^{-3}, decaying by 0.5 every 20 epochs till 100 epochs. The weights of $Exp2Emo$ remain frozen for the first 40 of those epochs. After that, it also starts training with a learning rate of 10^{-4}, decaying by 0.1 every 20 epochs. Weights of $Sp2Emo$ and $Exp2Emo$ are then frozen for the rest of the training period. Next, G and D are adversarially trained with learning rates of 10^{-4} and 4×10^{-5} respectively. $\mathcal{L}_{adv}$, $\mathcal{L}_{emotion}$, $\mathcal{L}_{mouth}$ and $\mathcal{L}_{energy}$ are empirically weighted respectively by 0.5, 0.5, 10 and 10. Besides stabilizing GAN training and quickening convergence, this ratio strikes the correct balance between the emphases placed on each aspect of naturalism.

We re-iterate that our model is designed such that emotion subnetworks do not need to be trained till their classification accuracies are high. Rather, it is sufficient to train them till they agree on the classification probabilities for corresponding speech emotion and expression pairs. This allows them to be pre-trained on acted emotions or emotion labels with low annotator agreement, and train later on larger in-the-wild datasets without emotion labels.

3.2. Qualitative Results

For testing, high resolution source images are taken from our test split of the MEAD dataset. Low resolution source images and speech samples are taken from the official test split of VoxCeleb2. Figure 4 illustrates the naturalism in the outputs. The provided examples, besides showing the quality of lip-sync, demonstrate the match between speech emotions (judged from the transcripts) and expressions. The dynamic expressions, manifesting as deformations around the eyebrows, eyes, mouth corners, etc., cannot be inferred from the lingual content alone.

We compare primarily with MakeItTalk [16] and the 'fixed pose' version of PC-AVS [17] (since their full version uses a pose source video also as a conditioning input), which are trained on VoxCeleb2. Outputs from our model are compared with their outputs in Fig 4a. Though the state-of-art approaches produce higher fidelity lip motion, their outputs show a static direction in which the speaker is looking, rigid expressions and head positions. Both PC-AVS and our work re-orient the face to face the viewer, for more engaging videos. Our approach produces lip motions while simultaneously making consistent and subtle changes to the head orientation and expression, making the videos look quite naturalistic.

3.3. Quantitative Results and User Preference Studies

SyncNet [28] confidence score is a widely accepted measure of how well the lip motions match the speech. Our model's score on the test split of VoxCeleb2 is 2.08, which is outperformed by MakeItTalk at 2.80 and PC-AVS at 5.90. A reason for the difference in SyncNet scores is that they are estimated from detected landmark motions. Our outputs can be penalized for producing 'unnecessary' motions uncorrelated to the speech. For example, a subject smiling while saying a certain word might register as an erroneous lip-shape, mismatched to the speech.

Metrics to judge image generation quality (SSIM, PSNR etc.) are not compared here since that is a feature of the pre-trained rendering module. It is also difficult to obtain a metric to reflect the quality of emotive expressions. Usual metrics that compare landmark positions with the reference are unusable because of the stochasticity in expressions and head motions produced by our model. Further, there are no publicly available models to classify emotion from speech and/or expression. Hence we cannot take that route to judge the match between input speech emotion and generated expression from our model.

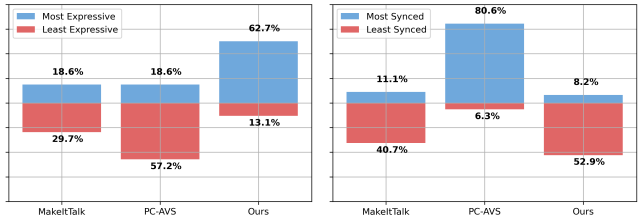


Fig. 5: Comparison of user feedback with state-of-art approaches MakeItTalk [16] and PC-AVS [17]. On the left, blue bars denote percentage of responses which found an approach to be most expressive, while red bars denote the percentage which found it least expressive. The right-hand plot deals similarly with most synced and least synced lip movements.

In the absence of quantitative metrics, we turn to user preference studies. 18 speech clips from the VoxCeleb2 test set are used to drive the same source face image across all 3 approaches. Viewers are presented with the three videos in a set simultaneously (in a randomized order). 43 university students from various disciplines ranked them with respect to lip-sync and expressiveness. As shown in Figure 5, 62.7% users find videos generated by our approach the most emotionally expressive, while 57.2% agree that videos produced by PC-AVS are the least expressive. On the other hand, an overwhelming 80.6% of the responses find the lip motions of PC-AVS to be the most well-synced to the speech audio. In that regard, both MakeItTalk and our approach are significantly less favoured, with 40.7% and 52.9% responses voting these as the least synced. Thus, our model managed to perform at par with MakeItTalk in terms of lip sync and outperform both in terms of naturalness of expressions and head movements. These results also show that there is a trade-off between expressiveness and lip-sync.

3.4. Ablation Studies

We conduct limited ablation studies, as follows.

$OursnoEmo$: Removing the emotion subnetworks leads to slightly higher SyncNet scores, but the match between speech emotion and expression goes down from 68.5% ($Ours$) to 8.5% (as estimated by our pre-trained subnetworks in the absence of publicly available ones).

$OursnoEntropy$: Removing the loss term $\mathcal{L}_{energy}$ erratically results in excessively jittery or static outputs (depending on other parameters). In one example, ratio between absolute of the mean landmark deformations and the mean of absolutes of the same deformations falls from 0.13 (more stable, $Ours$) to 0.07 (more jitter).

$OursnoPreTrain$: The most important ablation study is where we train all modules simultaneously, instead of pre-training emotion subnetworks. This understandably results in mode collapse. Outputs exhibit lips only opening and closing imperceptibly, while no expressions are produced.

We also verify that convergence is much faster when learning landmark deformations rather than their absolute positions.

4. CONCLUSION

We have developed a one-shot emotional Talking Head Generation model, requiring only a speech clip and a single in-the-wild face image as inputs. We have proposed a semi-supervised training scheme that allows its extension to datasets without explicit emotion labels.

We have shown that, compared to state-of-art approaches, our outputs are visually more realistic and natural, with generated expressions in agreement with the source speech emotion. Ablation studies demonstrate the effectiveness of various design decisions. Future work should focus on improving both lip synchronization and expressiveness without hampering each other. A larger feature space can be adopted for intermediate representations, instead of the limiting face landmark space. Finally, the model will benefit from a voice conversion module to become a more complete solution.

5. REFERENCES

[1] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu, "What comprises a good talking-head video generation?: A survey and benchmark," *arXiv preprint arXiv:2005.03201*, 2020.

[2] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European Conference on Computer Vision*. Springer, 2020, pp. 716–731.

[3] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[4] Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala, "Iterative text-based editing of talking-heads using neural retargeting," 2020.

[5] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[6] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie, "Photo-real talking head with deep bidirectional lstm," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4884–4888.

[7] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[8] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman, "You said that?," *arXiv preprint arXiv:1705.02966*, 2017.

[9] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Lip movements generation at a glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[10] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.

[11] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[12] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi, "Talking face generation by conditional recurrent adversarial network," *arXiv preprint arXiv:1804.04786*, 2018.

[13] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, pp. 1–16, 2019.

[14] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "End-to-end speech-driven realistic facial animation with temporal gans.," in *CVPR Workshops*, 2019, pp. 37–40.

[15] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *European Conference on Computer Vision*. Springer, 2020, pp. 408–424.

[16] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[17] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Taiki Shimba, Ryuhei Sakurai, Hirotake Yamazoe, and Joo-Ho Lee, "Talking heads synthesis from audio with deep neural networks," in *2015 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2015, pp. 100–105.

[19] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, August 2020.

[20] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[21] Robert Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[23] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky, "Fast bi-layer neural synthesis of one-shot realistic head avatars," in *European Conference of Computer vision (ECCV)*, August 2020.

[24] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[26] S. Haq and P.J.B. Jackson, *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pp. 398–423, IGI Global, Hershey PA, Aug. 2010.

[27] Kate Dupuis and M Kathleen Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[28] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.