Pixel Polyglots: An Innovative Approach to Pronunciation Enhancement in Language Learning using Computer Vision

Saket Pradhan, Kanishka Gabel, Srushti Hippargi, Shrey Shah

EECS 504: Foundations of Computer Vision, Fall 2023
University of Michigan – Ann Arbor

{saketp, kgabel, shipparg, shreyzz}@umich.edu

***Abstract*—Popular language learning applications often lack effective tools for teaching pronunciation and speech. We propose a service that creates realistic talking AI-generated head videos of users speaking phrases in their target language. This is achieved by leveraging deepfake techniques to map audio clips to facial motions that accurately lip sync and convey expressions. We use SadTalker to achieve this, which predicts 3D face motion coefficients from audio using novel networks conditioned on speaker identity. These coefficients animate a differentiable renderer to output videos fitted to the user's face. By providing immersive visualization of speech, this tool could revolutionize pronunciation practice. Though, this method has one major drawback: the expressions of the AI-avatar are extremely sensitive to the expressions of the subject in the input image. To rectify this error, our proposed methodology uses latent codes of the original input image and corrects them for the aforementioned expressional bias, and recreates them using the StyleGAN's encoder. We demonstrate that through our methodology, we are able to generate better deepfake videos. The proposed service aims to make multifaceted language learning accessible worldwide through user-friendly applications of AI. Our code repository is available at https://github.com/Saketspradhan/EECS-504-F23/**

Keywords—deepfakes, GANs, headpose, facial motion, PoseVAE, Wave2Lip, talking heads, speech visualization

I. Introduction

Language learning applications like Duolingo [1] and Babbel have catalyzed a digital revolution, yet a critical gap persists in effectively teaching pronunciation and speech. As linguists emphasize, conversing with native speakers is optimal for attaining fluency. However, the absence of comprehensive speech visualization tools impedes the immersive experience many enthusiasts seek. This predicament inspires an ingenious solution: creating a service that leverages AI-generated deepfake avatars to provide realistic visualizations of users speaking in their target language, processed directly on their mobile devices with minimal GPU usage. By integrating these cutting-edge innovations, this proposed tool bridges the gap between auditory and visual learning modalities. Visualizing oneself articulating foreign phrases could truly revolutionize the language acquisition process. Our project explores the prospective development and expected results of such a service. It accentuates lifelike speech visualization through AI avatars as pivotal for facilitating personalized, interactive learning experiences that promote engagement and vocabulary retention. Moreover, it analyzes the potential for technological advancements in speech recognition and synthesis to democratize language education on a global scale, empowering learners from diverse backgrounds and abilities. Specifically, this service aims to augment pronunciation and verbal proficiency by leveraging deepfakes to generate accurate visual renderings of users practicing conversations, reinforced through customized feedback. By condensing such complex innovations into user-friendly applications, it strives to make multifaceted language learning accessible worldwide.

II. Related Works

A. Synthesia.io

Synthesia.io is an online video content creation service that uses artificial intelligence to enable users to produce high-quality videos featuring virtual presenters that seamlessly mimic human expressions and gestures. It revolutionizes video creation with AI-driven virtual presenters mimicking human expressions and gestures. Ideal for education, marketing, and entertainment, it transforms content production. For language learners, it's a game-changer, offering personalized practice with virtual hosts in various languages. Improve pronunciation, observe facial expressions, and emulate native speakers for an immersive and effective learning experience.

B. Other related works

Several methods have explored generating talking head videos from audio and a single image. Examples include Wav2Lip , which focuses primarily on accurate lip sync [5] but does not model other facial motions well. PC-AVS (Zhou et al. 2021) disentangles pose and expression in a latent space but struggles with resolution and requires an additional control video. Other works like Audio2Head (Wang et al. 2021) and Audio-Animator (Wang et al. 2022) produce talking heads through latent warping but often have distortions or lack natural motions. In the 3D domain, some methods utilize 3D Morphable Models as an intermediate representation between audio and output video but have faced challenges with inaccurate expressions or rendering artifacts. Compared to these past approaches, SadTalker [4] introduces novel networks to map audio to realistic 3D

facial motion coefficients which then drive a 3D-aware renderer to produce higher quality, stylized, synchronized talking head videos from a single image input.

III. Methodology

The paper proposes a method called SadTalker [4] to generate talking head videos from a single image and an audio clip. SadTalker [4] generates talking head videos from a single image and audio using disentangled 3D face [6] representations. It predicts realistic expression and pose coefficients over time from audio via ExpNet and PoseVAE. These 3D coefficients then drive a novel differentiable renderer based on an unsupervised keypoint mapper and image warping pipeline to produce videos showing desired facial movements and synchronized speech. Losses on rendered outputs give control over expressions and quality. The modular audio-controllable 3D approach allows creating varied, personalized talking heads.

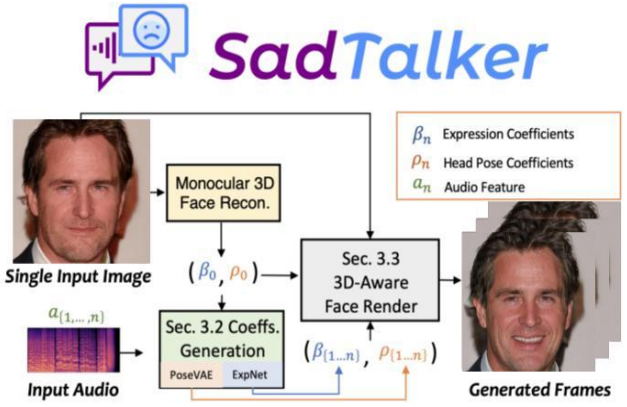

Fig. 1. Basic workflow of SadTalker

A. ExpNet

ExpNet is one of the key components of the proposed SadTalker [4] method for generating talking head videos from audio. Its goal is to predict realistic facial expression coefficients that will animate a 3D facial model. A major challenge in mapping audio to expressions is that the relationship depends on the identity. Different people may raise their eyebrows or smile differently for the same sounds. To address this, ExpNet conditions the predictions on the expression coefficients from the first frame of the reference image. This provides identity-specific conditioning. Another issue is that audio mainly controls some expression motions like lip sync [5], but less so for blinks. ExpNet alleviates this by training to predict lip-only expression coefficients from a state-of-the-art lip sync [5] model called Wav2Lip. Additional losses imposed on rendered 3D faces [6], like landmark loss for blinking eyes, enable learning remaining expressions.

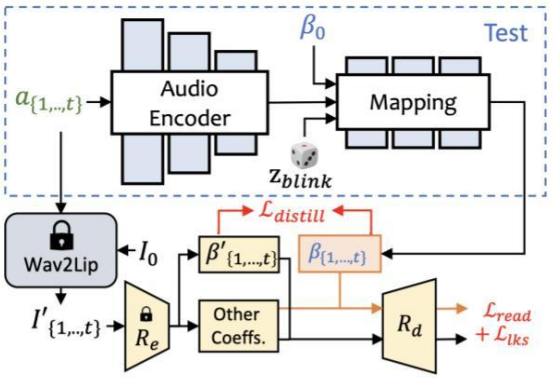

Fig. 2. Brief overview of the structure of ExpNet

The audio encoder embeds mel-spectrogram audio features into a latent space, then a linear mapping network predicts the 64 expression coefficients conditioned on audio, reference coefficients, and a blink control code over time. Objective functions include distillation loss comparing predicted coefficients to lip-sync ground truth, landmark loss, and lip reading loss measuring mouth crop synchronization. The disentangled signals and perceptual losses enable it to accurately generate identity-customized expression coefficients tailored for talking head video generation through implicit 3D face [6] rendering.

B. PoseVAE

PoseVAE is another key component of the proposed SadTalker [4] method for audio-driven talking head generation. While ExpNet focuses on facial expressions, PoseVAE tackles generating realistic and diverse head motions from audio. Since the same speech sounds can be spoken with different poses, PoseVAE adopts a variational autoencoder [7] structure to learn a distribution of plausible head pose coefficients. Instead of directly predicting the 6 coefficient pose sequences from audio, it models the residual motion from the first frame's pose. This enables synthesizing longer and more stable motions. To inject both rhythmic information from audio and personalized motion style, PoseVAE conditions the latent distribution on an audio encoding and a one-hot identity code.

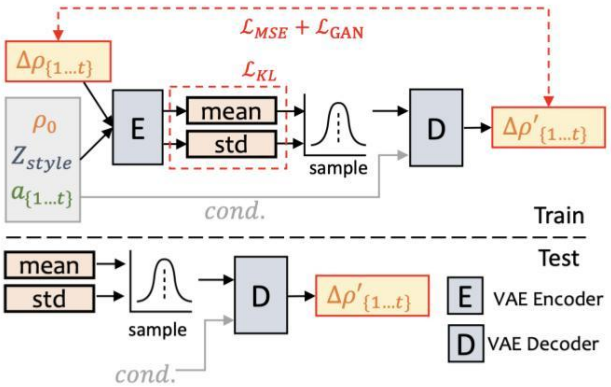

Fig. 3. Pipeline of PoseVAE

Fig. 5. Results showing facial image modification to neutralize the expressions of the subject by revising its latent codes through StyleGAN's encoder.

The encoder outputs the mean and variance vectors modeling the T-frame residual pose distribution, which is sampled and fed to the decoder. The decoder concatenates the sample with audio and identity codes to produce the final residual pose sequence. Adversarial and reconstruction losses train the network. At test time, sampling different latent codes allows outputting varied head movements fitting the input audio. PoseVAE essentially learns a conditional latent space of head motion styles specialized for an identity, which gets modulated over time based on the driving audio track to yield realistic and personalized talking head motions.

C. 3D Face Render

The last major piece of the SadTalker [4] pipeline is the 3D Face [6] Render, which animates a talking head video given the predicted 3D facial motion coefficients from ExpNet and PoseVAE. Its goal is to map the explicit expression and pose coefficients to implicit representations that can realistically warp and blend facial imagery. The render pipeline builds on state-of-the-art image animator Face-Vid2Vid, which learns to transfer motions between faces using unsupervised keypoints. Face-Vid2Vid requires reference driving videos with tracked motions. To adapt this for driving with 3D coefficients, the proposed renderer introduces a MappingNet module. MappingNet is trained to transform sequences of expression and pose coefficients into the unsupervised keypoint space learned by Face-Vid2Vid's architecture in a reconstructive manner.

At test time, the predicted coefficients are passed through the mapper and renderer to output video frames exhibiting desired facial expressions and head rotations. By factorizing motions through an intermediate 3D disentangled representation and specialized mapping network, the renderer can realistically animate facial motions from audio-derived coefficients. Photorealistic synchronization is ensured via losses imposed on the warped output frames and keypoints. The modular 3D-controllable renderer thus enables flexible audio-driven video portrait animation.

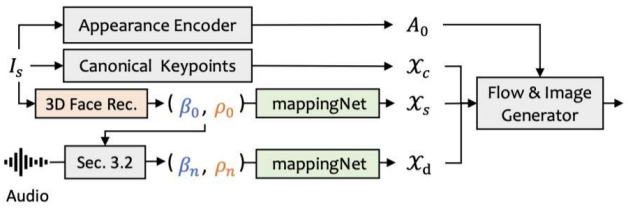

Fig. 4. FaceRender Architecture

D. Expression Correction Through Latent Codes

The SadTalker [4] paper faces challenges in representing eye and teeth variations due to limitations in the 3D Morphable Models (3DMM) used, leading to distorted video generation by failing to capture facial landmarks and treating expressive images as neutral. To address this, the method enhances control over image style and features by manipulating latent codes in a redesigned generator architecture. Disentanglement in the intermediate latent space improves control, potentially correcting specific attributes, including facial expressions.

StyleGAN's latent space controls distinct facial attributes. Encoding the expressive image into StyleGAN's latent space involves mapping methods like optimization algorithms or encoder networks. Manipulating latent codes adjusts emphasis on expressive features, transitioning the image toward neutrality. The modified latent codes are fed into the StyleGAN decoder, generating a new image with the expression transformed. Refinement steps fine-tune the generated neutral image to accurately apply desired

expression changes without compromising other facial attributes or image quality.

E. Datasets

The FaceRender component was trained on the VoxCeleb [8] Dataset, encompassing over 100,000 videos featuring 1251 subjects. Meanwhile, a specific subset of 1890 aligned videos and audios derived from VoxCeleb [8], featuring 46 subjects, was employed for training the PoseVAE and ExpNet components.

For evaluation, the HDTF Dataset was utilized. This dataset comprises 346 in-the-wild talking head videos with high resolution. The initial frame of each video served as the reference image, and the initial 8-second audio segment was employed as the driving signal for video generation in testing scenarios.

StyleGAN is trained on the CelebA dataset. It involves assimilating intricate facial variations, enabling the model to manipulate latent codes effectively for expressive to neutral image conversion. The CelebA dataset's comprehensive representation, especially expressions, supports successful transformation while preserving image fidelity.

IV. Experimentation and Results

We host our project on Modal Labs, a cloud computing service to run the inference, and have tested the model on Python 3.9 and an A10 GPU. On average, it takes ~15 seconds of inference time to generate an AI-avatar video clip, on a 512x512 pixel input image on a 10 second input audio. The environment setup includes essential libraries like Torch, torchvision, and torchaudio. The process involves cloning our GitHub repository, downloading model weights, and executing the inference script. The result is an AI-avatar, demonstrating the fusion of image and speech elements.

Acknowledgment

We gratefully acknowledge the invaluable guidance and support provided by Prof. Jason Corso and the team of Graduate Student Instructors, Sachin Salim, Anurekha Ravikumar, and Shrikant Aaravasu, throughout the Pixel Polyglots project for EECS 504: Foundations of Computer Vision. Their expert mentorship helped fuel our exploration into innovative approaches for pronunciation enhancement in language learning using computer vision. Their dedication to fostering academic growth and their insightful feedback were instrumental in shaping the depth and direction of our research. We extend our sincere gratitude to them for their continuous guidance and support.

References

[1] Learn a language for free (no date) Duolingo. Available at: https://www.duolingo.com/ (Accessed: 06 December 2023).

[2] GmbH, B. (no date) Which language do you want to speak?, Language for Life - Babbel.com. Available at: https://www.babbel.com/ (Accessed: 06 December 2023).

[3] #1 AI Video Generator (no date) Synthesia. Available at: https://www.synthesia.io/ (Accessed: 06 December 2023).

[4] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2022). SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. arXiv preprint arXiv:2211.12194.

[5] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P.Namboodiri,and C.V.Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In ACM MM, 2020.

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In CVPR Workshops, 2019

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2014.

[8] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In INTERSPEECH, 2017

[9] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high resolution audio-visual dataset. In CVPR, 2021.

[10] 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 CVPR, 2021.

[11] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In IJCAI, 2021.

[12] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. Oneshot talking face generation from single-speaker audio-visual correlation learning. In AAAI, 2022.