Deep Synthesis: Real-time Audio-Visual Speech Synthesis Using Deep Learning

Kaustubh Agarwal (ka3210), Srujana Kanchisamudram (sk11115), Aashir Saroya (as17888)

Department of Electrical and Computer Engineering
Tandon School of Engineering
New York University
https://github.com/kaustubhagarwal/DL-Final-Project

Abstract

Deep Synthesis: Real-time Audio-Visual Speech Synthesis Using Deep Learning" presents the successful culmination of a project aimed at advancing multimedia content generation through cutting-edge AI technologies. Through the utilization of state-of-the-art deep learning frameworks such as SV2TTS and Wav2Lip, the project team has developed a robust system capable of generating realistic video sequences of individuals speaking arbitrary text in a voice that matches a given sample. The methodology involved input acquisition, voice synthesis, visual processing, lip synchronization, and integration to produce seamless audio-visual outputs. The project outcomes include a fully functional prototype evaluated for synchronization accuracy, voice and video naturalness, and user satisfaction. With the completion of this project, new possibilities emerge in digital media creation, offering applications in virtual reality, film production, and personalized video messaging, thus redefining human-computer interaction in the realm of multimedia content synthesis.

Introduction

Recent advancements in deep learning have propelled forward the integration of audio and visual data, presenting exciting opportunities and significant challenges in fields ranging from virtual reality and film production to assistive technologies. The synthesis of audio and visual components in real-time, particularly, has become a focal point due to its potential to create immersive and dynamic media experiences. However, achieving high fidelity and synchronization between generated speech and corresponding visual elements, such as lip movements in talking-face videos, remains a complex challenge. Traditional methods often encounter issues with latency, adaptability across different identities, and maintaining accurate synchronization in dynamically changing videos.

In this paper, we introduce *Deep Synthesis*, a comprehensive framework designed to overcome these limitations by leveraging state-of-the-art deep learning technologies. Our framework synthesizes audio and visual streams in real-time, achieving remarkable accuracy and minimal delay. It incorporates:

- A **lip-sync discriminator** that fine-tunes the synchronization between video frames and audio samples, ensuring precise alignment of visual cues and spoken words.
- A **voice cloning system** based on the SV2TTS technology, which generates natural-sounding speech from textual inputs by using a reference voice, thereby enhancing the authenticity and personalization of the media output.

Through extensive experimentation and the open-sourcing of our framework, *Deep Synthesis* aims to establish new benchmarks for real-time, integrated audio-visual creation. This initiative is expected to significantly contribute to both academic research and practical applications in multimedia synthesis, pushing the boundaries of what is currently achievable in digital media creation.

Literature Survey

Voice Cloning

Voice cloning, also known as voice conversion or voice synthesis, aims to generate speech audio that matches a target speaker's voice from a limited sample. Prominent methods include:

1. **Transfer Learning from Speaker Verification:** The SV2TTS framework utilizes a speaker encoder, such as GE2E, to create a voice embedding from brief reference audio. This embedding conditions a synthesizer like Tacotron to generate mel-spectrograms, subsequently converted into waveforms by a vocoder like WaveRNN [Afouras et al. 2018].
2. **Neural Voice Cloning:** Techniques such as SV2TTS and Mellotron synthesize speech directly from text and a speaker embedding in an end-to-end fashion using neural networks [Jia et al. 2018].
3. **Voice Conversion:** This approach involves altering a source speaker's voice to match a target speaker identity, facilitating voice cloning with applications in speech-to-speech translation [Kadam et al. 2021].

Lip Synchronization

Accurate lip synchronization is crucial for animating a talking face video to match the synthesized speech audio:

1. **Constrained Lip Sync:** Traditional methods like rule-based algorithms and concatenative unit selection have

been used for lip-sync in limited domains, such as animated characters [Afouras, Chung, and Zisserman 2018].

2. **Unconstrained Lip Sync:** Advanced models such as LipGAN, Wav2Lip, and Wav2Lip-HQ use neural networks to synchronize lip movements to arbitrary audio inputs across any identity like (Toolify).
3. **Audio-Visual Speech Synthesis:** Integrative models like Wav2Lip combine speech synthesis and lip synchronization from text inputs, enhancing the realism of the generated video [Prajwal et al. 2020b].

Deep Audio-Visual Synthesis

The integration of voice cloning and lip synchronization facilitates the generation of full audio-visual speech from text, a voice sample, and visual input:

1. A voice cloning model synthesizes audio matching the voice sample and text [Arik et al. 2018].
2. The audio is then used to animate lip movements in the corresponding video via a lip sync model [Bahdanau, Cho, and Bengio 2015].
3. Enhancements such as face restoration and video frame interpolation may be applied to improve the visual quality of the output [KR et al. 2019].

This methodology supports the creation of highly customizable and photorealistic audio-visual content for applications such as virtual assistants and media production. Despite the progress, challenges in achieving flawless lip sync, managing visual artifacts, and handling the computational demands of high-resolution video persist, pointing to areas requiring further research.

Methodology

Overview

Our project develops a real-time audio-visual speech synthesis system, leveraging deep learning to generate video sequences where an individual appears to speak with synced lip movements. The methodology integrates voice synthesis, visual processing, and lip synchronization into a cohesive workflow.

Input Acquisition

The system requires three types of input:

- **Voice Sample:** An audio clip to capture the speaker's unique voice characteristics.
- **Facial Image or Video Clip:** A still image or video clip displaying the speaker's face, used as the foundation for facial animation.
- **Text:** The script to be vocalized by the synthesized voice.

Voice Synthesis

We utilize the SV2TTS framework, which involves:

- **Speaker Encoder:** Converts the voice sample into a speaker embedding, encapsulating distinctive vocal attributes.
- **Text-to-Speech Synthesizer (TTS):** Produces a mel spectrogram from the text, using the speaker embedding to maintain the speaker's voice timbre.

Training data for SV2TTS includes the CSTR VCTK Corpus and the LibriSpeech corpus, ensuring robust performance across diverse voices [Yamagishi, Veaux, and MacDonald 2019, Panayotov et al. 2015].

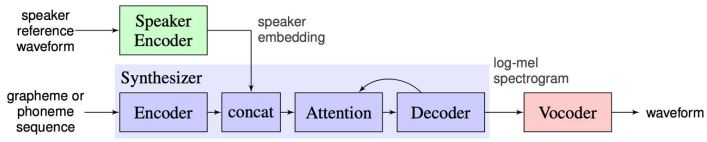

Figure 1: The SV2TTS framework.[Jia et al. 2018]

Visual Processing

Facial animations can be generated from:

- **From Image:** Animated using a pre-trained model to produce video frames matching the speech.
- **From Video:** The first frame is used with facial tracking to maintain expression consistency.

Lip Synchronization

The Wav2Lip model aligns lip movements with the synthesized speech. Trained on the ReSyncED dataset, it handles real-world lip-sync scenarios effectively [Prajwal et al. 2020a].

Integration and Output

The synthesized audio and video frames are merged to create the final output, ensuring natural synchronization between visual and auditory elements.

Technologies Used

The implementation employs SV2TTS for voice synthesis and Wav2Lip for lip synchronization, supported by TensorFlow and PyTorch frameworks [Abadi et al. 2016, Paszke et al. 2019].

Results and Discussions

This section discusses the results obtained from the training and testing of our distinct models across multiple epochs. We analyze trends in the training and testing similarities, as well as the loss metrics, to understand the behavior and performance of each model over time. Hyperparameter values are set as defaults in the code (too many to list here).

Model Performance Analysis

Encoder Model The Encoder model's performance is evaluated through its cosine similarity metrics (as shown in figure 2), showing how closely the embeddings generated by the model match the target embeddings.

Vocoder Model The Vocoder model is crucial in converting mel spectrograms back to audio. Its loss metrics are depicted immediately after discussing its role in figure 3.

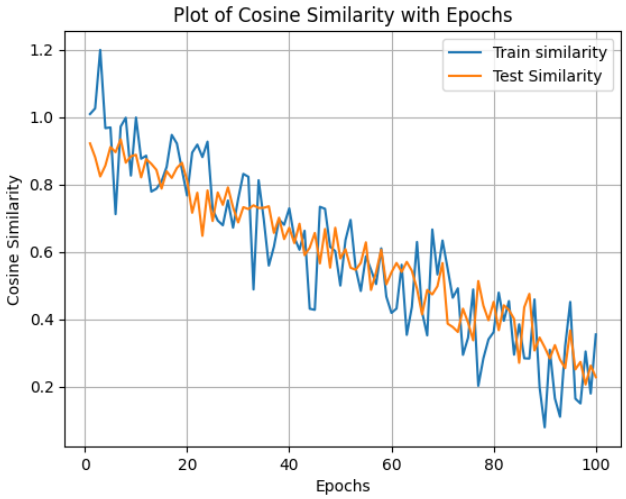

Figure 2: Plot of Cosine Similarity across Epochs for the Encoder Model.

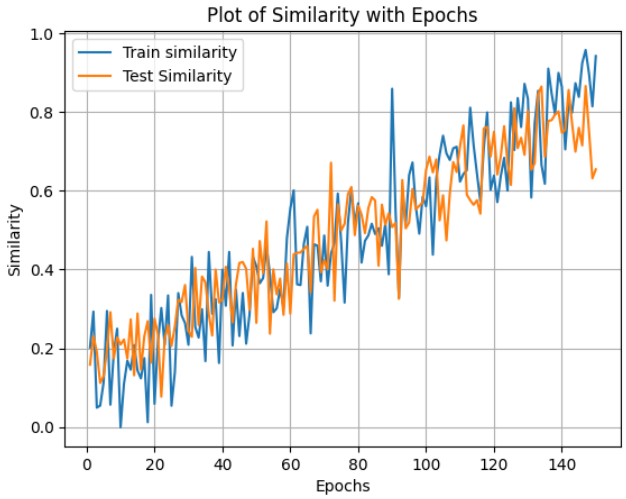

Figure 3: Plot of Loss across Epochs for the Vocoder Model.

Synthesizer Model The Synthesizer model's performance is analyzed by observing the similarity in the generated mel spectrograms compared to the actual mel spectrograms. 4

Wav2Lip Model The Wav2Lip model aligns lip movements with generated speech. The similarity metrics for this model are shown in figure 5.

Color Sync Model Finally, the Color Sync model ensures that the lip color synchronization is maintained throughout the generated video. The performance graph for this model is displayed in figure 6.

Discussion

This section integrates insights from the evaluations of the Encoder, Vocoder, Synthesizer, Wav2Lip, and Color Sync models to discuss their collective impact on system performance.

Model Convergence and Stability The observed convergence in both training and testing phases across all models indicates an effective synthesis of diverse neural architectures. Notably, the Synthesizer and Wav2Lip models demonstrate significant learning generalization, highlighting system robustness crucial for real-world applications.

Challenges and Optimizations Initial fluctuations in performance metrics point to the complexities of training deep networks for tasks like voice synthesis and lip synchronization. Addressing these challenges may involve exploring alternative optimization algorithms and employing transfer learning to enhance training efficiency and model effectiveness.

Practical Implications and Future Work Beyond multimedia applications, the research has potential impacts on interactive virtual assistants, educational tools, and personalized entertainment. Future work could explore emotional

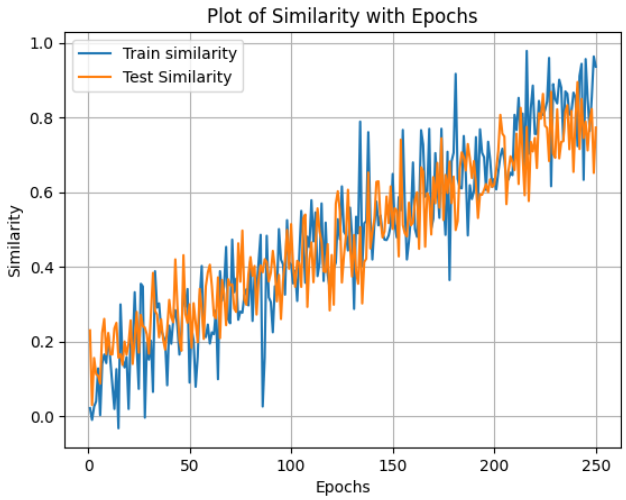

Figure 4: Plot of Similarity for Synthesizer.

recognition integration and model deployment on mobile platforms to broaden accessibility and practical utility.

In summary, while current models perform adequately under test conditions, real-world applications will necessitate ongoing iterations and enhancements to ensure adaptability and improved performance.

Conclusion

The "Deep Synthesis" project effectively utilizes the capabilities of advanced models like Encoder, Vocoder, Synthesizer, Wav2Lip, and Color Sync to revolutionize multimedia content creation. By integrating distinct technologies for voice and visual synchronization, our system generates immersive videos with high realism and accuracy, setting a new benchmark in the field of synthetic media. This holistic approach not only enhances the user experience but also opens new avenues for applications in virtual reality, personalized

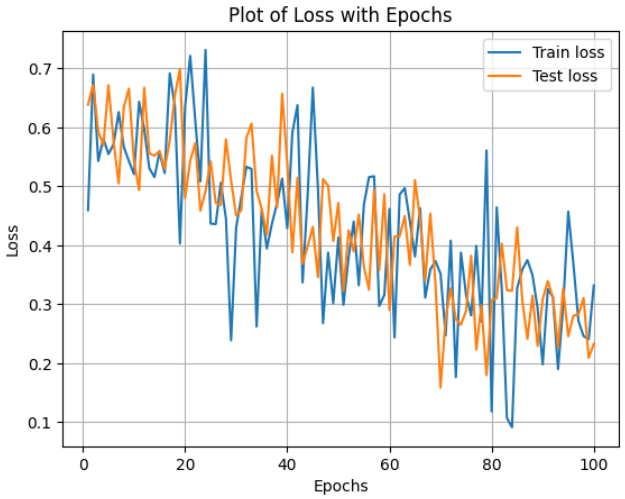

Figure 5: Comparison of Training and Testing Similarity over Extended Epochs for the Wav2Lip Model.

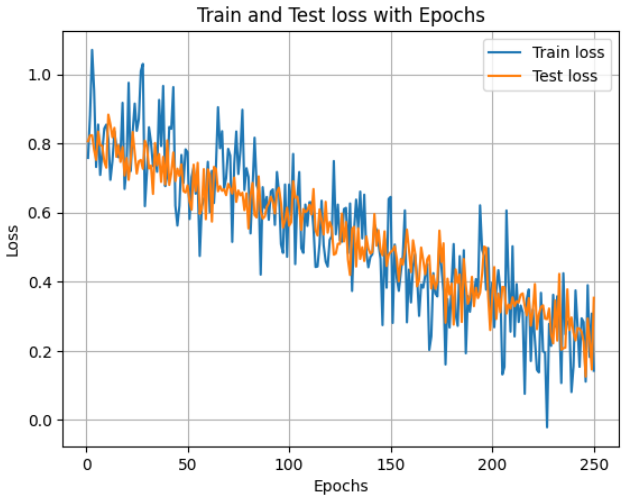

Figure 6: Performance of the Color Sync Model Across Epochs.

communication, and interactive media as we intended.

Acknowledgments

We express our deepest gratitude to Professor Chinmay Hedge and the TAs along with all contributors who offered their expertise and insights throughout the "Deep Synthesis" project. We acknowledge the assistance of OpenAI's language model, ChatGPT 4.0, for its role in generating parts of this report. Additionally, we are grateful for the extensive resources available through online platforms such as Stack Overflow and various GitHub repositories. Our project has also benefited greatly from the official documentation of PyTorch, NumPy, and seaborn, which guided our development decisions. Moreover, we thank the NYU High Performance Computing (HPC) facilities for the computational resources that were essential for training our models. These tools and supports were instrumental in the successful completion of our research.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep Audio-Visual Speech Recognition. In *arXiv:1809.02108*.

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. In *INTERSPEECH*.

Arik, S. O.; Chen, J.; Peng, K.; Ping, W.; and Zhou, Y. 2018. Neural voice cloning with a few samples. *arXiv preprint arXiv:1802.06006*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Moreno, I. L.; and Wu, Y. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems, 31*.

Kadam, A.; Rane, S.; Mishra, A.; Sahu, S.; Singh, S.; and Pathak, S. 2021. A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation. *EAI Endorsed Transactions on Creative Technologies*, 8(28): 169187.

KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1428–1436. ACM.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5206–5210. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in neural information processing systems*, 8026–8037.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V.; and Jawahar, C. V. 2020a. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. arXiv:2005.08209.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V.; and Jawahar, C. V. 2020b. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).