MercuryNet: Video-to-Prosody Synthesis

Matthew Caren Jonas Rajagopal
MIT EECS
Cambridge, MA, USA
{mcaren, jrajagop}@mit.edu

1. Introduction

We propose *MercuryNet*, a system that aims to reconstruct prosodic features from silent videos of human speech. We will specifically target the sequence-to-sequence task of predicting F_0 (pitch) and intensity (volume) from video frames.

2. Background

2.1. Lip-to-Speech Models

Video-to-prosody is closely related to video-to-speech synthesis (a.k.a. "lip-to-speech synthesis"), the task of reconstructing speech audio from silent video of lip/face movement. Lip2Wav [5] (which is open-source and available on GitHub) is the most notable recent system targeting this task, and uses an LSTM/attention-based video encoder paired with a speech-synthesis decoder fine-tuned from Tacotron 2 [7]. Lip2Wav, as well as most other modern lip-to-speech models [1, 3], is trained end-to-end as a self-supervised machine learning problem.

2.2. Prosody

Prosody, in a general linguistic sense, refers to all aspects of speech not encoded by literal word meaning—including intonation, stress, and rhythm [8]. Prosodic features indicate attributes like emotion, focus, and sarcasm, and are essential to spoken communication in every human language.

Although prosody in spoken language is enormously important, it is expressed through a small set of acoustic features: fundamental frequency F_0, phoneme duration, intensity, and spectral qualities. Out of these features, we focus on F_0 and intensity—phoneme duration is closely linked with a specific language's vocabulary, and spectral qualities are highly dependent on individual voices' qualities.

Although a handful of state-of-the-art lip-to-speech models exist, no existing systems specifically target prosody generation from video. Though video-to-prosody is in some senses a narrower problem than video-to-speech, these prosody elements have well-established definitions from linguistics and can be quantifiably measured and analyzed. Furthermore, they are foundational to how humans produce and interpret speech, so they more concretely measure aspects of vocal communication central to sentiment, emotion, and stress compared to features and objective functions canonically used to evaluate raw audio data (e.g. spectrogram-based approaches).

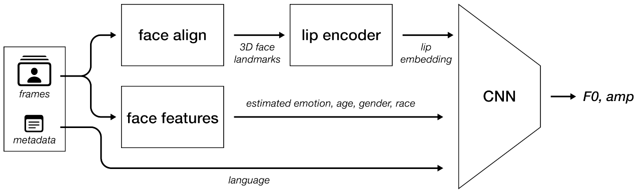

Figure 1. MercuryNet architecture

3. Model Architecture

The MercuryNet model takes a silent video of a person and for each frame, predicts the frequency and intensity of the speech as well as a binary flag for whether the person is speaking.

3.1. Dataset

Data for training and testing was obtained from the AVSpeech dataset [2], which consists of over 100,000 segments of YouTube videos and the timestamp/location of a specific person in the video speaking.

3.2. Preprocessing

Before being used to train the MercuryNet model, the AVSpeech data must undergo a variety of preprocessing steps. First, the appropriate segment of the video is downloaded using *FFmpeg* and *yt-dlp*. Next, the YOLOv8 face detection model[1] is used to acquire a bounding box for each faces in every frame in a video. These faces are cross-referenced with the expected speaker location in the AVSpeech dataset annotations to ensure the correct face is being extracted. A video is only included in the dataset

[1] https://github.com/akanametov/yolov8-face

if a face can be confidently selected for every frame in a video—and if so, the frames are cropped to include just the appropriate face and saved.

The video title is also processed using the LangDetect model[2] to get an embedding for the languange of the speaker.

The inputs could be also be annotated with the estimated age, gender, emotion, and race, extracted via the DeepFace model [6]. However, DeepFace runs extremely slowly so currently this metadata is set to 0.

In addition, the preprocessing pipeline must also create the ground truth data. First, the audio is extracted from the video. Next, the PYIN algorithm [4] is used to extract the fundamental frequency f_0 of the speech. This also returns a voiced flag which indicates if the speaker is talking. The intensity of the signal is calculated via the root mean square.

The throughput for the entire pipeline—including downloading the video and running YOLOv8—is roughly 20 image frames per second on an Apple M1 Max GPU using the PyTorch MPS backend. We have created a dataset of 10,000 video segments and their cropped frames and corresponding ground truth data. There are roughly 1,700,000 total images in the dataset. The AVSpeech dataset contains many more videos so there is room to grow the dataset if necessary.

3.3. The MercuryNet Model

The first part of the MercuryNet model is formed from the encoder of the Lip2Wav model [5][3]. This model is 3D CNN with an LSTM module at the end of it. The model takes a set of `n_frames` frames and returns embeddings of size (`n_frames`, 384). The training code dictates that `n_frames` is 90 with a 30 frame overlap between each window. These embeddings are then passed into a custom-designed CNN which convolves only along the `n_frames` axis. The location of data in the embeddings is not spatially correlated so convolving along that dimension (the 384 one) would not be useful. The model currently consists of 2 linear layers then 3 convolution layers then 2 linear layers then 4 convolution layers then 2 linear layers then a final linear layer to form the (`n_frames`, 3) output as desired. We plan to tune this model architecture more once we are able to train the model. We have a framework and a forward pass is functional however there are some bugs with the loss function.

Train and test loss will be measured using L2 loss on a logarithmic scale to match human perception. Qualitatively, we aim to be able to synthesize a tone that matches the volume and rising and falling pitch of the original speech. Quantitatively, we would measure overall performance by measuring the distribution of F_0 and intensity deviation from ground truth frequency obtained from the AVSpeech dataset.

3.4. Training

[For progress report this is still to do] We plan to train MercuryNet on the AVSpeech dataset [2] as outlined above. [Add training details once we have them]

4. Conclusion

Successful video-to-prosody work can facilitate more accurate artificial speech models that more accurately capture nuances of human speech and tone of voice. One might also consider MercuryNet to be the "missing piece" of lip-to-text transcription systems, which are used to create on-the-fly captioning systems for hard-of-hearing people, as well as in situations where providing audio is infeasible.

5. Contributions

Both authors have contributed to all parts of the system, including the scoping of the project, the preprocessing code, the model design, and the work to train the model.

References

[1] Ariel Ephrat and Shmuel Peleg. Vid2Speech: Speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017.

[2] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[3] Thomas Le Cornu and Ben Milner. Generating intelligible audio speech from visual speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1751–1761, 2017.

[4] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 659–663. IEEE, 2014.

[5] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[7] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech*

[2] https://pypi.org/project/langdetect/

[3] https://github.com/joannahong/Lip2Wav-pytorch

and Signal Processing (ICASSP), pages 4779–4783. IEEE, 2018.

[8] Nancy Helen Woo. *Prosody and phonology*. PhD thesis, Massachusetts Institute of Technology, 1969.