MercuryNet: Video-to-Prosody Synthesis [Proposal]

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

1. Overview

We propose the MercuryNet system, which aims to reconstruct prosody 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.

1.1. Prosody

Prosody, in a general linguistic sense, refers to all aspects of speech not encoded by literal word meaning—including intonation, stress, and rhythm [7]. 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.

1.2. Prior Work and 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 [4] (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 [6]. 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.

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 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 communication central to sentiment, emotion, and stress compared to raw audio measurements/losses (e.g. spectrograms).

2. Proposed System

2.1. Data

We plan to use the AVSpeech dataset [2], which is publicly available and contains thousands of hours of speech video from YouTube across a variety of languages and contexts. Each video clip can be annotated with its language using the CLD3 language detection model[1] on video title metadata.

2.2. Architecture

We plan to fine-tune existing models such as the encoder portions of Lip2Wav or Vid2Speech (both CNNs) or existing vision transformer-based models. We will use the AVSpeech dataset to get the extracted face region normalized to a uniform size and augment these inputs with the with the video language as well as estimated age, gender, emotion, and race, extracted via the DeepFace model [5]. The final model ultimately outputs two values for each video frame: one that predicts the F_0 in hertz, and another that predicts the RMS loudness in decibels.

We will train the model using MPS locally for smaller tests, and on a Satori cluster for bigger tasks.

2.3. Evaluation

Train and test loss will be measured using either L1 or L2 loss (to be decided empirically) 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. Impacts

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.

[1] https://github.com/google/cld3

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] 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.

[5] 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.

[6] 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 and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

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