HOLY-WOOD ACADEMY KOLHAPUR’S

SANJEEVAN ENGINEERING AND TECHNOLOGY INSTITUTE, PANHALA- 416 201

A

Seminar Report

On

“Audio-Driven Facial Animation by Joint End-to-End Learning of Poseand Emotion”

Submitted By

Mr. Vivek Maruti Patil (51)

PRN: 51631520181124210024

Under the Supervision of

Asst. Prof. K. Kari

SANJEEVAN ENGINEERING
& TECHNOLOGY INSTITUTE

Dr. Babasaheb Ambedkar
Technological University
डॉ. बाबासाहेब आंबेडकर तंत्रशास्त्र विद्यापीठ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ACADEMIC YEAR:2019-20

Certificate

This is to certify that the following student of S.Y.B. Tech [Computer Science & Engineering], Sanjeevan Engineering & Technology Institute have completed their Seminar Report on the subject **“Audio-Driven Facial Animation by Joint End-to-End Learning of Pose And Emotion**” within the premises of Institute as per the guidelines laid down by DBATU, Lonere to our satisfaction during the year 2020-2021.

Mr. Vivek Maruti Patil (51)

Date: / /
Place: SETI, Panhala

(Asst.Prof.K. Kari)
Supervisor

Asst. Prof. M.M. Hajare
Head of Department

SIGNATURE OF EXAMINER

Dr. Mohan Vanrotti
Principal

ACKNOWLEDGEMENT

It gives me an immense pleasure to present a report on the successful completion of my seminar work on **“Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion”** We express our deep sense of gratitude to our guide **“Asst.prof.K. Kari.”** for his valuable guidance rendered in all phases of seminar. We are thankful of his wholehearted assistance, advice and expert guidance towards making my seminar success.

My special thanks to respected Principal and Head of the Department for their keen interest, encourage and excellent support.

Roll No.	NAME	PRN No.
51	Mr. Vivek Maruti Patil	51631520181124210024

List of figures

ABSTRACT

This seminar report is about **“Audio-Driven Facial Animation by Joint End-to-End Learning of Pose And Emotion”.** We present a machine learning technique for driving 3D facial animation by audio input in real time and with low latency. Our deep neural network learns a mapping from input waveforms to the 3D vertex coordinates of a face model, and simultaneously discovers a compact, latent code that disambiguates the variations in facial expression that cannot be explained by the audio alone. During inference, the latent code can be used as an intuitive control for the emotional state of the face puppet.

We train our network with 3—5 minutes of high-quality animation data obtained using traditional, vision-based performance capture methods. Even though our primary goal is to model the speaking style of a single actor, our model yields reasonable results even when driven with audio from other speakers with different gender, accent, or language, as we demonstrate with a user study. The results are applicable to in-game dialogue, low-cost localization, virtual reality avatars, and telepresence.

Table of Contents

INTRODUCTION

Expressive facial animation is an essential part of modern computer-Generated movies and digital games. Currently, vision-based performance capture, i.e., driving the animated face with the observedMotion of a human actor, is an integral component of most production pipelines.While audio-based performance capture algorithms are Unlikely To ever match the quality of vision systems, they offer complementary strengths. Most importantly, the tens of hours of dialogue spoken by in-game characters in many modern games is much too expensive to produce using vision-based systems. Consequently,common practice is to produce only key animations, such as cinematics, using vision systems, and rely on systems based on audio and transcript for producing the bulk of in-game material.Our goal is to generate plausible and expressive 3D facial animation based exclusively on a vocal audio track. For the results to look Natural, the animation must account for complex and codependent Phenomena including phoneme coarticulation, lexical stress, and Interaction between facial muscles and skin tissue [Edwards et al.2016]. Hence, we focus on the entire face, not just the mouth and Lips. We adopt a data-driven approach, where we train a deep neural Network in an end-to-end fashion to replicate the relevant effects Observed in the training data.Our method produces expressive 3D facial motion from audio in Real time and with low latency. To retain maximal independence From the details of the downstream animation system, our method Outputs the per-frame positions of the control vertices of a fixed-Topology facial mesh.

2 RELATED WORK

We will review prior art in systems whose input is audio or text and output is 2D video or 3D mesh animation. We group the approaches into linguistic and machine learning based models, and also review methods that support apparent emotional states.Models based on linguistics. A large body of literature exists for analyzing and understanding the structure of language, and translating it to anatomically credible facial animation [Lewis 1991; Mattheyses and Verhelst 2015]. Typically, an audio track is accompanied with a transcript that helps to provide explicit knowledge about the phoneme content. The animation is then based on the visual counterpart of phonemes called visemes [Fisher 1968] through complex rules of coarticulation. A well-known example of such a system is the dominance model [Cohen and Massaro 1993; Massaro et al. 2012].

In general, there is a many-to-many mapping between phonemes and visemes, as implemented in the dynamic visemes model of Taylor et al. [2012] and in the recent work dubbed JALI [Edwards et al.2016]. JALI factors the facial animation to lip and jaw movements,based on psycholinguistic considerations, and is able to convincingly reproduce a range of speaking styles and apparent emotional states independently of the actual speech content.A core strength of these methods is the explicit control over the entire process, which makes it possible to, e.g., explicitly guarantee that the mouth closes properly when the puppet is spelling out a bilabial (/m/,/b/,/p/) or that the lower lip touches the upper teeth with labiodentals (/f/,/v/). Both are difficult cases even for vision-based capture systems. The weaknesses include the accumulated complexity of the process, language-specific rules, need for a near perfect transcript for good quality (typically done manually [Ed-wards et al. 2016]), inability to react convincingly to non-phoneme sounds, and lack of a principled way to animate other parts of the face besides the jaw and lips.FaceFX (www.facefx.com) is a widely used commercial package that implements the most widely used linguistic models, including the dominance model.Models based on machine learning. Here we will group the systems primarily based on how they use machine learning.

Given that our goal is to produce 3D animation based on audio,we are not inherently interested in the intermediate representations. Instead, we would like to formulate the entire mapping as an end-to-end optimization task. The early experiments with neural networks used audio to directly drive the control parameters of an animated 3D mesh [Hong et al. 2002; Massaro et al. 1999; Öhman and Salvi 1999], but the networks back then were necessarily of trivial complexity. We revisit this end-to-end formulation with deep convolutional networks, latest training methods, and problem-specific contributions

END-TO-END NETWORK ARCHITECTURE

We will now describe the architecture of our network, alongwithdetails on audio processing and the separation of emotionalstatefrom the speech content.Given a short window of audio, the task of our network is to infer the facial expression at the center of the window. We represent the expression directly as per-vertex difference vectors from a neutral pose in a fixed-topology face mesh. Once the network is trained,we animate the mesh by sliding a window over a vocal audio track and evaluate the network independently at each time step. Even though the network itself has no memory of past animation frames,it produces temporally stable results in practice.

3.1Architecture overview

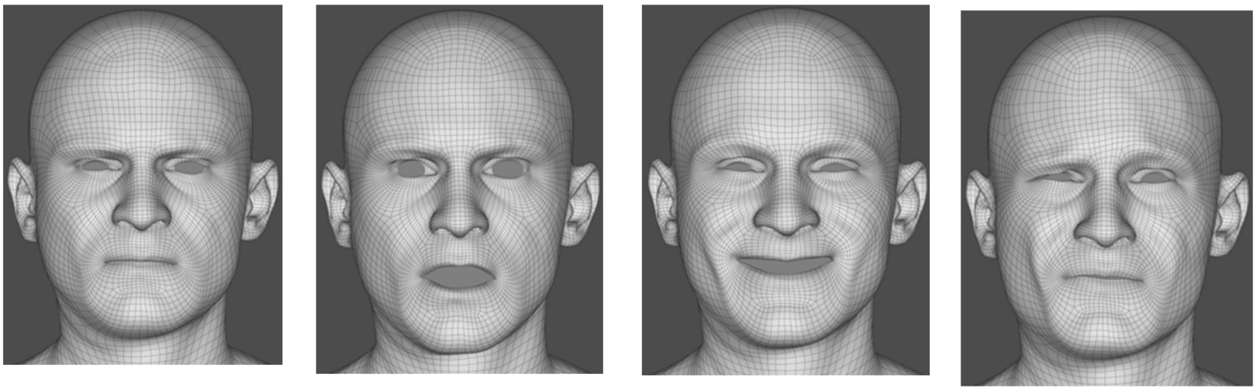
Our deep neural network consists of one special-purpose layer, 10 convolutional layers, and 2 fully-connected layers. We divide it in three conceptual parts, illustrated in Figure 1 and Table 1.We start by feeding the input audio window to a formant analysis network to produce a time-varying sequence of speech features that will subsequently drive articulation. The network first extracts raw formant information using fixed-function autocorrelation analysis (Section 3.2) and then refines it with 5 convolutional layers.Through training, the convolutional layers learn to extract short-term features that are relevant for facial animation, such as intonation, emphasis, and specific phonemes. Their abstract, time-varying representation is the output of the 5th convolutional layer.

3.2Audio processing

The main input to our network is the speech audio signal that we convert to 16 kHz mono before feeding it to the network. In our experiments, we normalize the volume of each vocal track to utilize the full [-1,+1] dynamic range, but we do not employ any other kind of processing such as dynamic range compression, noise reduction, or pre-emphasis filter.The autocorrelation layer in Table 1 converts the input audio window to a compact 2D representation for the subsequent convolutional layers. Our approach is inspired by the source–filter model of speech production [Benzeghiba et al. 2007; Lewis 1991], where the audio signal is modeled as a combination of a linear filter (vocal tract) and an excitation signal (vocal cords). The resonance frequencies (formants) of the linear filter are known to carry essential information about the phoneme content of the speech. The excitation signal indicates the pitch, timbre, and other characteristics of the speaker's voice, which we hypothesize to be far less important for facial animation, and thus we focus primarily on the formants to improve the generalization over different speakers.

The standard method for performing source–filter separation is linear predictive coding (LPC). LPC breaks the signal into several short frames, solves the coefficients of the linear filter for each frame based on the first K autocorrelation coefficients, and performs inverse filtering to extract the excitation signal. The resonance frequencies of the filter are entirely determined by the autocorrelation coefficients, so we choose to skip most of the processing steps and simply use the autocorrelation coefficients directly as our representation of the instantaneous formant information. This makes

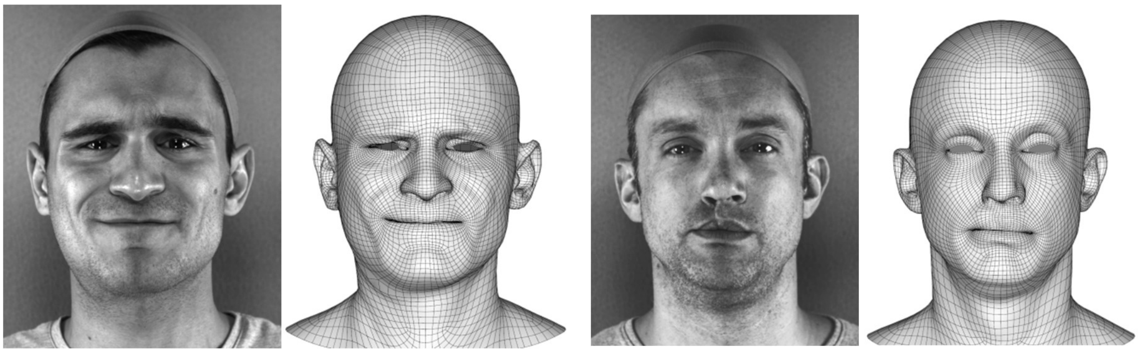
intuitive sense, since the autocorrelation coefficients essentially represent a compressed version of signal whose frequency content approximately matches the power spectrum of the original signal.The representation is a natural fit for convolutional networks, as the layers can easily learn to estimate the instantaneous power of specific frequency bands.



TRAINING

We will now describe the aspects relevant to training our network:

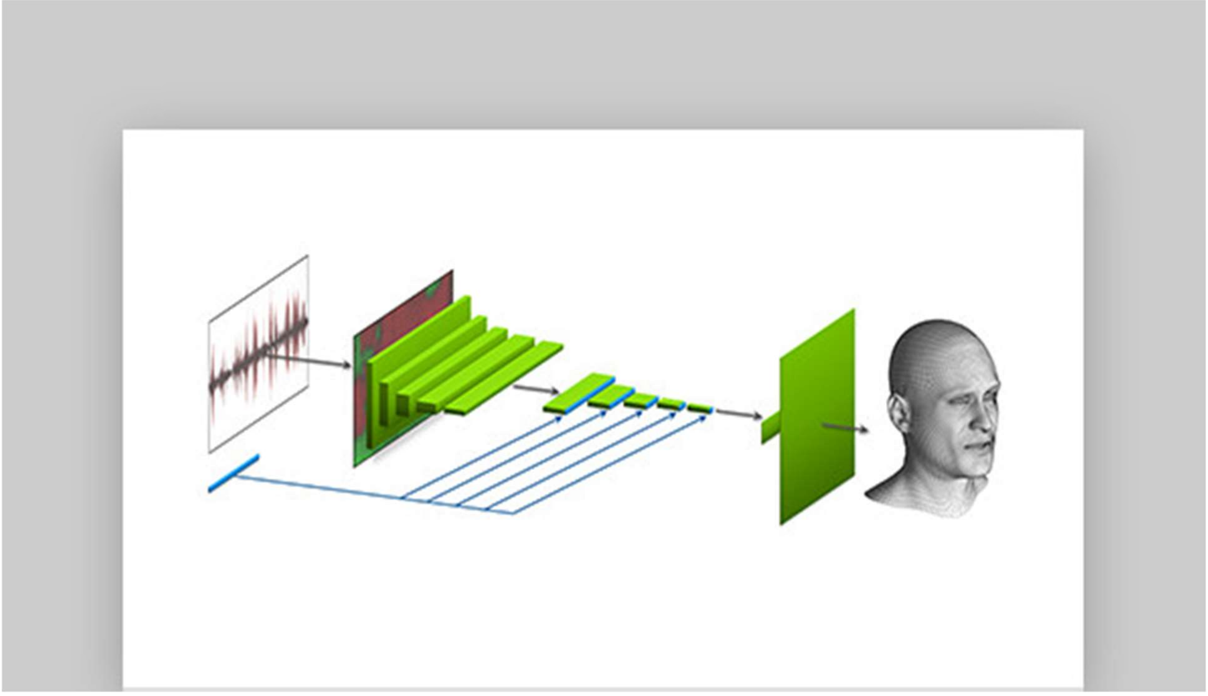
how the training targets were obtained, what our dataset consists of, dataset augmentation, loss function, and training setup



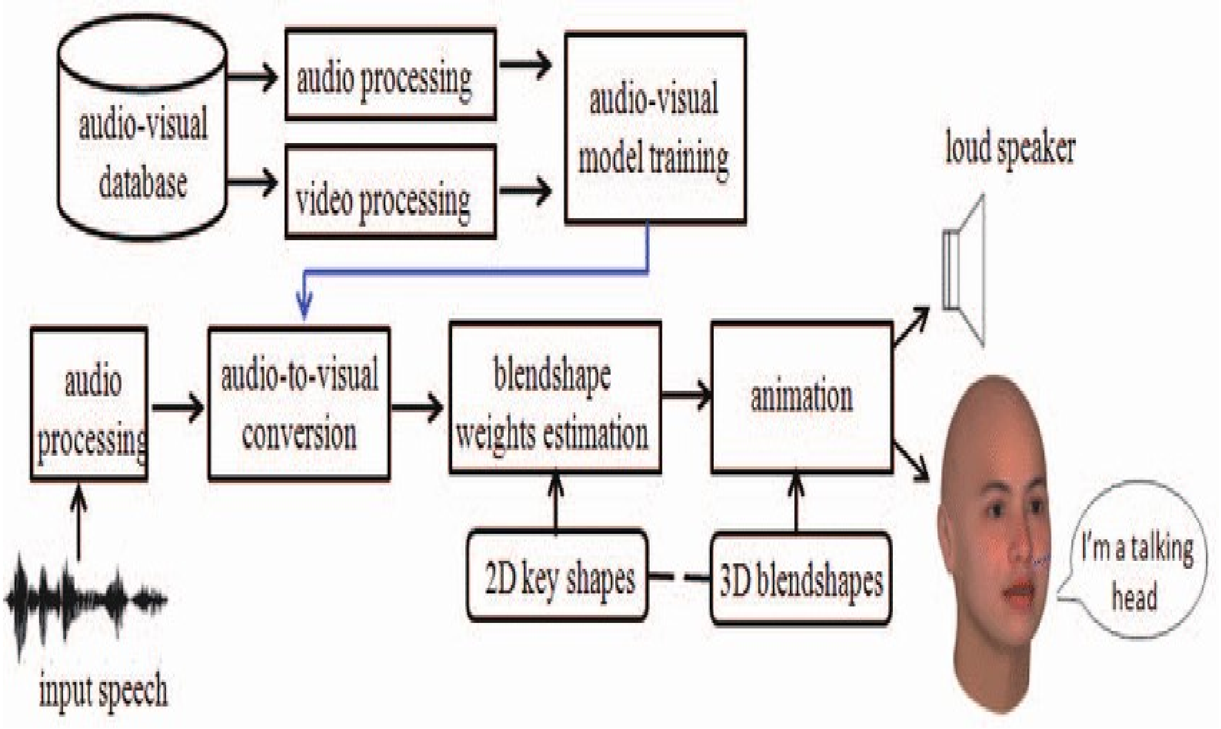
5 .INFERENCE AND RESULTS

Emotional states

When inferring the facial pose for novel audio, we need to supply the network with an emotional state vector as a secondary input.As part of training, the network has learned a latent E-dimensional vector for each training sample, and our strategy is to mine this emotion database for robust emotion vectors that can be used during inference



BLOCK AND PIN DIAGRAM



EVALUATION

To assess the quality of our results, we conducted a blind user study with 20 participants who had no professional experience on animation techniques. In the study, we compared the output of our method ("Ours") against video-based performance capture from the DI4D system ("PC") and against dominance model based animation [Cohen and Massaro 1993; Massaro et al. 2012] produced using FaceFX software ("DM"). The audio clips used as inputs were taken from the held-out validation dataset of the corresponding actor, i.e., they were not used as part of training data when training our network. The study featured two animated characters with a total of 13 audio clips that were 3–8 seconds long. For each clip a video was rendered with each of the three methods. Out of these, three A vs. B pairs were created (PC vs. Ours, PC vs. DM, Ours vs. DM), totaling 39 pairs of videos presented to the participants. These pairs were presented as A vs. B choices in random order, also randomizing which method was A and which was B for each individual question.

The participants, unaware of which videos originated from which method, progressed through the video pairs, choosing the more natural-looking animation of each pair before moving to the next one. The study took approximately 10–15 minutes to complete. All videos were presented at 30 frames per second with audio. For our method, we assigned each audio clip a constant emotion vector that was mined from the emotion database as explained

ADVANTAGES

No predefined emotions (≠ most machine learning methods, e.g. Cao et al, 2005)

● Full face animation (≠ linguistics-based methods)

● Applicable when free-viewpoint, flexible 3D models rendering/animation is needed (≠ methods based on concatenation/blending of video frames e.g. Ezzat et al, 2012)

● Applicable to different face meshes (retargeting) and synthetic audio

● Good temporal stability (thanks to random time-shifting and motion term)

● Good interpolation of emotional states (thanks to manual vector removal)

DISADVANTAGES.

Incapacity to cover residual motion (e.g. eye -saccades, blinks-, head, tongue)

○ Only motion related to articulation

● Emotion vectors have no semantic meaning (e.g. “happy”, “sad”) and potentially not generalizable

○ Not useful for inference if training data not covering all phonemes + coarticulation rules of every emotional state

● Bad performance if testing data hugely differs from training data, mainly:

○ Different (mostly faster) tempo

○ Input volume level of testing data not normalized

CONCLUSION

An automated system for creating an additional channel for communication is presented. From audio anda fewimages ofa person, a facial animation with lip sync and appropriate expressions is generated. The animation looks realistic and individual variability is preserved. It is also possible to generate new lip shapes in expressions previously not seen by the system. For the future it would be worthwhile to consider including other features like correct gaze following, controlled pose variation, eyebrow movement and eye blinking in the animation system

REFERENCES

[1] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal networks hard? arXiv:1905.12681, 2019.

[2] Shih-En Wei, jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Purdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. VR facial animation via multiview image translation. ACM Transaction on Graphics, 38(4), 2019.

[3] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In Advances in Neural Information Processing Systems, pages 5575–5585, 2018.

[4] Hang Zhou, Yu Liu, Liu Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audiovisual representation. In AAAI Conf. on Artificial Intelligence, 2019.

[5] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audiodriven animator-centric speech animation. ACM Transaction on Graphics, 37(4), 2018.