|  |  |
| --- | --- |
| **000** | **054** |
| **001** | **055** |
| **002** | **056** |

**003**

**004**

**005**

**006**

**007**

**008**

**009**

**010**

**011**

A Commercially Viable Speech Query to Video Response Model

Anonymous WACV submission Paper ID 900

**057**

**058**

**059**

**060**

**061**

**062**

**063**

**064**

**065**

**012**

**013**

**014**

**015**

**016**

**017**

**018**

**019**

**020**

**021**

**022**

**023**

**024**

**025**

**026**

**027**

**028**

**029**

**030**

**031**

**032**

**033**

**034**

**035**

**036**

**037**

**038**

**039**

**040**

**041**

**042**

**043**

**044**

**045**

**046**

**047**

**048**

**049**

**050**

**051**

**052**

**053**

## Abstract

*We present* ***speech query to video response model with emotions and creativity****, the unique architecture that gener- ates 3D Holographic video responses of the query asked by the user through voice. Our approach is composed of tra- ditional machine learning method as well as fully trainable neural modules in different sections. More precisely, we use six main modules: a speech to text[*[*16*](#_bookmark18)*] network, rule based chat bot model[*[*15*](#_bookmark17)*], a text-to-speech network[*[*22*](#_bookmark24)*] that is emotional is based on the PALASH framework[*[*1*](#_bookmark3)*], a time- delayed LSTM to generate mouth-key points synced to the audio[*[*14*](#_bookmark16)*], a network based on Pix2Pix GAN to generate the video frames conditioned on the key points[*[*10*](#_bookmark12)*] and a hologram based model to produce 3D video[*[*13*](#_bookmark15)*]. This com- mercially viable 3D video bot has been deployed as part of companies Customer Relationship Management system.*

## Introduction

A conscious and emotional 3D video bot is the need of the hour to make interactions with customers and prospects more human like. Data driven approaches for speech recognition[[3](#_bookmark5)], chat bot development and generating images have recently surpassed traditional way of computing these sections. In this work, we show that we can combine some of the independent AI structures to generate artificial 3D holographic video of a person reading aloud the response of the query asked by the user. Our video generation model can be trained on any set of close shot videos of an ac- tual person speaking. The result is a system that generates speech from the text generated as a response from chat bot, is trained on a data lake based on a framework that makes the bot conscious and emotional and modifies accordingly the mouth area of one of the videos so that it looks natural and realistic. Although we showcase the method on a male person whose videos were available as per our requirement, our approach can be used to generate videos of anyone in- cluding an anime provided the data availability is there.

# RELATED WORK

Recently, there have been important advances in the gen- eration of photo-realistic videos[[24](#_bookmark26)][[5](#_bookmark7)]. Here, broadly we have merged the speech synthesis[[19](#_bookmark21)][[22](#_bookmark24)], rule based chat bot[[15](#_bookmark17)] and 3D holographic video generation models. We have used the Gradient boosting model for chat bot section and neural network for video generation.

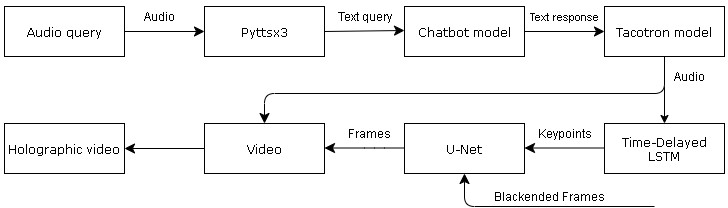


Figure 1: Flow Diagram of the Working Model

# ENHANCEMENT & MODEL DESCRIP- TION

### SPEECH TO TEXT AND TEXT TO SPEECH SYSTEM[[4](#_bookmark6)]

We have used the pyttsx3 model to process the audio as per the text. Audio is recognized using the Google audio recognizer. While for generating the emotional audio[[12](#_bookmark14)] we have used CNN model that is trained on data and inputs as per the PALASH framework[[1](#_bookmark3)].

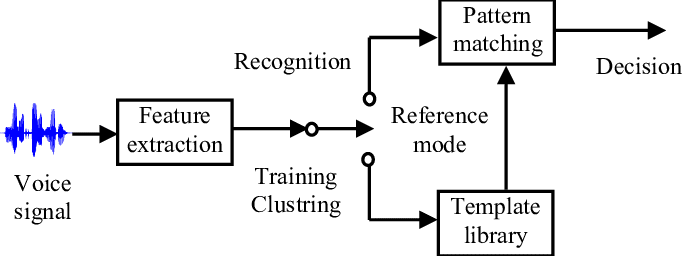


Figure 2: Speech to Text system

**066**

**067**

**068**

**069**

**070**

**071**

**072**

**073**

**074**

**075**

**076**

**077**

**078**

**079**

**080**

**081**

**082**

**083**

**084**

**085**

**086**

**087**

**088**

**089**

**090**

**091**

**092**

**093**

**094**

**095**

**096**

**097**

**098**

**099**

**100**

**101**

**102**

**103**

**104**

**105**

**106**

**107**

1

**108**

**109**

**110**

**111**

**112**

**113**

**114**

**115**

**116**

**117**

**118**

**119**

**120**

**121**

**122**

**123**

**124**

**125**

**126**

**127**

**128**

**129**

**130**

**131**

**132**

**133**

**134**

**135**

**136**

**137**

**138**

**139**

**140**

**141**

**142**

**143**

**144**

**145**

**146**

**147**

**148**

**149**

**150**

**151**

**152**

**153**

**154**

**155**

**156**

**157**

**158**

**159**

**160**

**161**

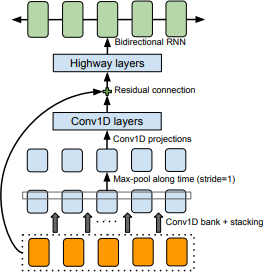


Figure 3: Text to Speech CNN Model

### CHAT BOT MODEL

We have developed rule based text chat bot[[15](#_bookmark17)] with the certain categories and the corresponding outputs. The cat- egories are specified as per the requirement and therefore can be changed with the use case. The text generated from the speech to text system is used as input for this model. The text is processed and tested over the Gradient Boost- ing Classification model to predict the category. Gradi- ent Boosting is an ensemble machine learning algorithm based on the tree structure, i.e. it is the advanced form of the traditional Decision Tree model. In this model ini- tially for the first tree, equal weights are assigned to all the features/parameters and after its evaluation, the second tree weights are assigned i.e. the features which play signif- icant role in the prediction are given more weights while relatively less weights are assigned to the parameters. This flows continues until the final tree predicts the output. Be- cause of this learning feature, this model stands out from the other Machine learning algorithms.

For each category we have several predefined output re- sponses and any one of them randomly gets appear.

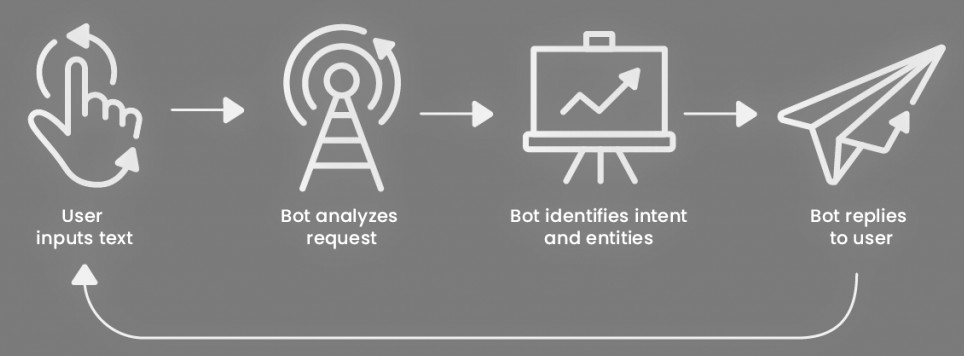


Figure 4: Rule based chat bot working flow

### KEYPOINTS GENERATION[[14](#_bookmark16)]

The text response generated from chat bot is converted to audio using text to speech[[4](#_bookmark6)] system and used as in- put. The mouth shapes are calculated using the normalized

mouth key points[[6](#_bookmark8)] extracted from the face, face location, face rotation and face size. Dlib 68 points Face landmark Detection[[18](#_bookmark20)][[8](#_bookmark10)][[20](#_bookmark22)] is used to outline the mouth section of the face which is 49-68 points.[[14](#_bookmark16)]

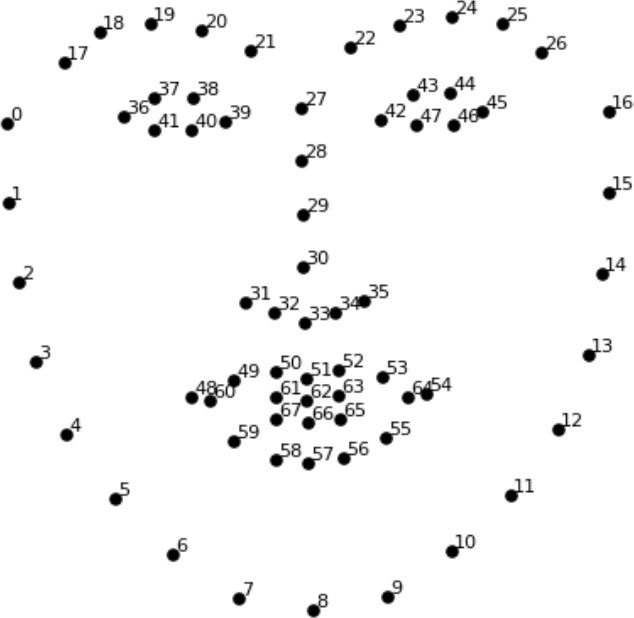


Figure 5: Dlib 68 facial landmarks

PCA is applied over the normalized key points to de- correlate the features and reduce the dimension. For this model we are only considering the most important principal components for mouth shape representation.

We then used the neural model i.e. LSTM network to predict the mouth shape[[2](#_bookmark4)] of each instances with the audio features as input.

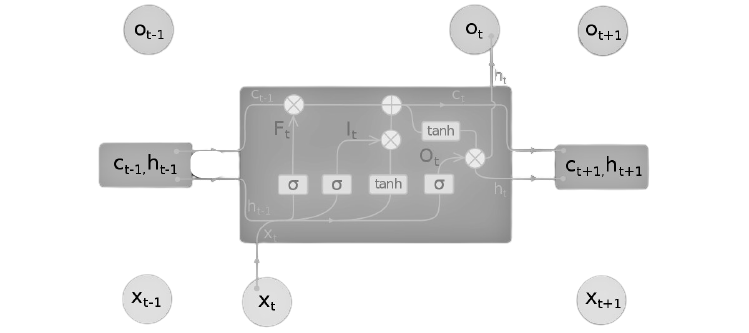


Figure 6: LSTM Model

Long Short Term Memory (LSTM)[[17](#_bookmark19)] is one of the prominent neural network model which is mostly used where we encounter long term dependencies like image processing[[11](#_bookmark13)][[23](#_bookmark25)].

**162**

**163**

**164**

**165**

**166**

**167**

**168**

**169**

**170**

**171**

**172**

**173**

**174**

**175**

**176**

**177**

**178**

**179**

**180**

**181**

**182**

**183**

**184**

**185**

**186**

**187**

**188**

**189**

**190**

**191**

**192**

**193**

**194**

**195**

**196**

**197**

**198**

**199**

**200**

**201**

**202**

**203**

**204**

**205**

**206**

**207**

**208**

**209**

**210**

**211**

**212**

**213**

**214**

**215**

**216**

**217**

**218**

**219**

**220**

**221**

**222**

**223**

**224**

**225**

**226**

**227**

**228**

**229**

**230**

**231**

**232**

**233**

**234**

**235**

**236**

**237**

**238**

**239**

**240**

**241**

**242**

**243**

**244**

**245**

**246**

**247**

**248**

**249**

**250**

**251**

**252**

**253**

**254**

**255**

**256**

**257**

**258**

**259**

**260**

**261**

**262**

**263**

**264**

**265**

**266**

**267**

**268**

**269**

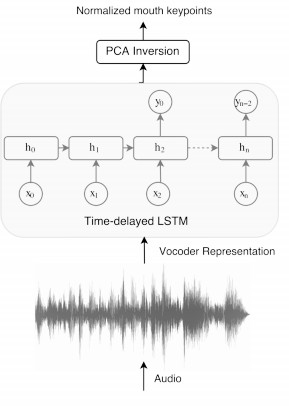


Figure 7: Key point Generation Network

### VIDEO GENERATION

For video generation purpose we have used the one of the prominent models i.e. pix2pix[[10](#_bookmark12)] which is a solution for image to image translation problems. As we have gen- erated the key points and drawn a lip sketch over the patch images as per the audio[[7](#_bookmark9)][[9](#_bookmark11)], we need to convert them and produce images with painted mouth area. The U-Net archi- tecture is what used to draw the lips over the mouth cropped region. This network is trained only using L1- loss in pixel- space and we found that this objective is sufficient to learn the in-painting of the mouth and doesn’t require the ex- tra GAN objective as originally proposed in pix2pix[[10](#_bookmark12)]. Each frame is processed and parallelly merged to form the video[[21](#_bookmark23)]. Then this video is combined with the audio[[2](#_bookmark4)] to get the output. We have used ffmpeg to convert video to images and vice versa[[14](#_bookmark16)]. Further this video is processed through Hologram function to generate the final 3D Holo- graphic video.

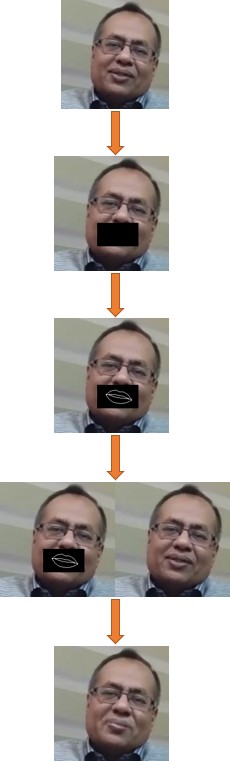


Figure 8: Workflow of the lip-sync model

**270**

**271**

**272**

**273**

**274**

**275**

**276**

**277**

**278**

**279**

**280**

**281**

**282**

**283**

**284**

**285**

**286**

**287**

**288**

**289**

**290**

**291**

**292**

**293**

**294**

**295**

**296**

**297**

**298**

**299**

**300**

**301**

**302**

**303**

**304**

**305**

**306**

**307**

**308**

**309**

**310**

**311**

**312**

**313**

**314**

**315**

**316**

**317**

**318**

**319**

**320**

**321**

**322**

**323**

**324**

**325**

**326**

**327**

**328**

**329**

**330**

**331**

**332**

**333**

**334**

**335**

**336**

**337**

**338**

**339**

**340**

**341**

**342**

**343**

**344**

**345**

**346**

**347**

**348**

**349**

**350**

**351**

**352**

**353**

**354**

**355**

**356**

**357**

**358**

**359**

**360**

**361**

**362**

**363**

**364**

**365**

**366**

**367**

**368**

**369**

**370**

**371**

**372**

**373**

**374**

**375**

**376**

**377**



Figure 9: Holographic video

# RESULTS AND FURTHER IMPROVE- MENTS

The fully functional 3D chatbot was deployed for a cus- tomer of Gazelle Information Technologies by the name of Clean Water Generator (www.cleanwaterproblem.com). The technical and product related queries asked by prospec- tive customers was efficiently handled by this 3D chatbot. The result was increased interaction with the prospects who converted into customers for the futuristic product. This is an improvement over the existing 3D animated models, as these have been actually deployed in the real world. The next step is to generate a 3D holographic human like model of a Yoga instructor who can be positioned as an AI drive Yoga instructor.

## References

1. Goyal et al. creativity in machines: Music composition using artificial intelligence. [1](#_bookmark0)
2. Generation by adversarially disentangled audio-visual repre- sentation. *arXiv preprint arXiv:1807.07860*, 2018. [2](#_bookmark1), [3](#_bookmark2)
3. Triantafyllos Afouras, Joon Son Chung, and Andrew Zisser- man. Lrs3-ted: a large-scale dataset for visual speech recog- nition. *arXiv preprint arXiv:1809.00496*, 2018. [1](#_bookmark0)
4. Sakshi Bhargava, Sravya Tummala, Shravani S, Shivam Koul, and Nilima Kulkarni. *PDF to AudioBook Converter*. [1](#_bookmark0), [2](#_bookmark1)
5. Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966 (2017)*, 2017. [1](#_bookmark0)
6. Joon Son Chung and Andrew Zisserman. Lip reading in the wild. in asian conference on computer vision. *Springer*, pages 87–103. [2](#_bookmark1)
7. Joon Son Chung and Andrew Zisserman. Out of time: au- tomated lip sync in the wild. in workshop on multi-view lip- reading, accv. 2016. [3](#_bookmark2)
8. Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv:2006.07397* *[cs.CV]*, 2020. [2](#_bookmark1)
9. Ohad Fried, Ayush Tewari, Michael Zollho¨fer, Adam Finkel- stein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graph-* *ics (TOG) 38, 4 (2019)*, pages 1–14. [3](#_bookmark2)
10. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adver- sarial networks. *arXiv preprint arXiv:1611.07004*, 2016. [1](#_bookmark0), [3](#_bookmark2)
11. Byeong-Ho KANG. *A Review on Image and Video process-* *ing*. [2](#_bookmark1)
12. Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. *Audio-driven facial animation by joint end- to-end learning of pose and emotion*. ACM Trans. Graph., 36(4):94:1– 94:12. July 2016. [1](#_bookmark0)
13. Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Ab- hishek Jha, Vinay Namboodiri, and CV Jawahar. To- wards automatic face-to-face translation. in proceedings of the 27th acm international conference on multimedia. acm, 1428–1436. 2019. [1](#_bookmark0)
14. Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Bre´bisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017. [1](#_bookmark0), [2](#_bookmark1), [3](#_bookmark2)
15. Maali Mnasri. *Recent advances in conversational NLP : To-* *wards the standardization of Chat bot building*. [1](#_bookmark0), [2](#_bookmark1)
16. Shivangi Nagdewani and Ashika Jain. *A Review on methods* *for speech to text and text to speech conversion*. [1](#_bookmark0)
17. Alex Sherstinsky. *Fundamentals of Recurrent Neural Net- work (RNN) and Long Short-Term Memory (LSTM) Network*. [2](#_bookmark1)
18. Shenghao Shi. *Facial Keypoints Detection*. [2](#_bookmark1)
19. Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio.

Char2wav: End-to-end speech synthesis. 2017. [1](#_bookmark0)

1. Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv:2001.00179 [cs.CV]*, 2020. [2](#_bookmark1)
2. Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *Interna- tional Journal of Computer Vision (2019)*, pages 1–16, 2019.

[3](#_bookmark2)

1. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss†, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyr- giannakis, Rob Clark, and Rif A. Saurous. Tacotron: To- wards end-to-end speech synthesis. [1](#_bookmark0)
2. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Si- moncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* *13, 4 (2004)*, pages 600–612. [2](#_bookmark1)
3. Hang Zhou, Yu Liu, Ziwei Liu Ping Luo, and Xiaogang Wang. Talking face. 2018. [1](#_bookmark0)

**378**

**379**

**380**

**381**

**382**

**383**

**384**

**385**

**386**

**387**

**388**

**389**

**390**

**391**

**392**

**393**

**394**

**395**

**396**

**397**

**398**

**399**

**400**

**401**

**402**

**403**

**404**

**405**

**406**

**407**

**408**

**409**

**410**

**411**

**412**

**413**

**414**

**415**

**416**

**417**

**418**

**419**

**420**

**421**

**422**

**423**

**424**

**425**

**426**

**427**

**428**

**429**

**430**

**431**