EAI Endorsed Transactions
on Creative Technologies

Research Article EAI.EU

A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation

Anup Kadam[1], Sagar Rane[1,*], Arpit Kumar Mishra[1], Shailesh Kumar Sahu[1], Shubham Singh[1], Shivam Kumar Pathak[1]

[1]Department of Computer Engineering, Army Institute of Technology, Pune, MH, India

Abstract

The fields like Media, Education and Corporations etc have started focusing on content creation. This has led to the huge demand for synthetic media generation using less data. To synthesize a high-grade artificial video, the lip must be synchronized with the audio. Here we have compared the various methods for voice-cloning and lip synchronization. Voice cloning procedure include state of the art methods like wavenet and other text-to-speech approaches. Lip synchronization methods describe constrained and unconstrained methods. Various recent research like LipGan, Wav2Lip are discussed. The methods are compared and the best method is suggested. Apart from studying and comparing the various methods, their drawbacks, future scopes, and application are also there. Different social and ethical issues are also discussed.

Received on 13 March 2021; accepted on 10 April 2021; published on 14 April 2021

Keywords: Video Synthesis, Voice Cloning, Lip Synchronization, Video Generation Application

Copyright © 2021 Anup Kadam *et al.*, licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution license, which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi:10.4108/eai.14-4-2021.169187

1. Introduction

The project is to make media generation and content more individual-focused, personalized. We aim to make Ads personalized such as by using the targeted user's name. Sending bulk personalized congratulations videos, celebrations videos, etc. The core of our research is to do audio synthesis is the speaker's voice using fewer data. At present media-generated is generic such as Ads we see, video emails we get, etc [1]. Our project purpose is to make generic media content specific to a user. The user will be provided with premium services where each video generated will be mailed to specific clients with custom content of the user. In this kind of service. The user will have to upload a file containing email addresses corresponding to each string. The Ad industry targets many ads on us through videos we see on social networks (Facebook, Instagram, Twitter, etc), video platforms (YouTube).

Deep learning has shown its potential in various fields using machine learning. One of the major usage of Deep learning is Text to Speech(TTS) [2]. While the complete training of a single-speaker TTS model is technically a form of voice cloning, the interest rather lies in creating a fixed model able to incorporate newer voices with little data. The common approach is to condition a TTS model trained to generalize to new speakers on an embedding of the voice to clone [3]. This approach is typically more data-efficient than training a separate TTS model for each speaker, in addition to being orders of magnitude faster and less computationally expensive. Interestingly, there is a large discrepancy between the duration of reference speech needed to clone a voice among the different methods, ranging from half an hour per speaker to only a few seconds. This factor is usually determining the similarity of the generated voice with respect to the true voice of the speaker. Apart from voice cloning, one more important feature is lip synchronization [4]. There have been many useful applications of lip syncing in making a perfect synthetic video. Generally application requires generic and speaker independent models. One more challenge which we face is different sizes of lips [5]. The audio and movement of lips go out of sync which makes an automatic generated video look absurd. Approximately, 1 sec out of sync lip movement is identified by the viewers. To remove this out of sync thing, we use lip-synchronisation technique.

*Corresponding author. Email: sagarrane@aitpune.edu.in

2. Literature Survey

Audio Synthesis is a process to modify frequencies and wavelengths of audio samples so as to produce the desired type of output. There are a lot of methods to synthesize the audio. But the majority of them require a large amount of data. Our aim is to synthesize audio with as little data as possible. A process is well-known as Text-to-Speech(TTS). Progressive work in this field includes Concatenative TTS [6] Concatenative TTS is a technique of synthesizing sound by adding up short sounds of recorded audio clips that are combined together to form a speech. Limitations are Switching to different speakers, altering the emotion of speech, Introducing new words Parametric TTS [7] The text is extracted of the linguistic features like phonemes, and Secondly, the extract vocoder features from it. Limitations are Sounds less natural than Concatenative, more like a robot. Working of WaveNet [8] Over 16,000 samples collected for 1-second of raw-audio. In Generative Model (not word features), Autoregressive Model, output only takes previous time-step into consideration and not future-step. Our objective is to achieve a powerful form of voice cloning. The resulting framework must be able to operate in a zero-shot setting, that is, for speakers unseen during training. It should incorporate a speaker's voice with only a few seconds of reference speech. These desired results are shown to be fulfilled by. Their results are impressive, but not backed by any public implementation. In addition, the methodology integrates a model based on the framework to make it run in real-time, i.e. to generate speech in a time shorter or equal to the duration of the produced speech.

3. Audio Synthesis

Audio synthesis is the technique of generating sound, using electronic hardware or software. The sound spoken by humans is regarded as subtractive synthesis. Output integration is a method of over-output audio with the concept audio, which is characterized by the use of an audio filter in the audio signal. The vocal cords act as the sound source and the mouth and throat modify the sound. Vocal cords are generating a harmonic sound. The two are different in a way by filtering applied on the mouth and throat.

TTS(Text to speech) technology converts a given text to a human-like sound. It is trained on a very large dataset of a speaker and the audio generated is robotic and somewhat similar to that of the speaker. The options available for converting text to speech are very few since one model is created from the voice dataset of only one or very few speakers. Today, these 'tts' are used in various fields like robots, digital assistants, etc. But they sound more like robots than humans.

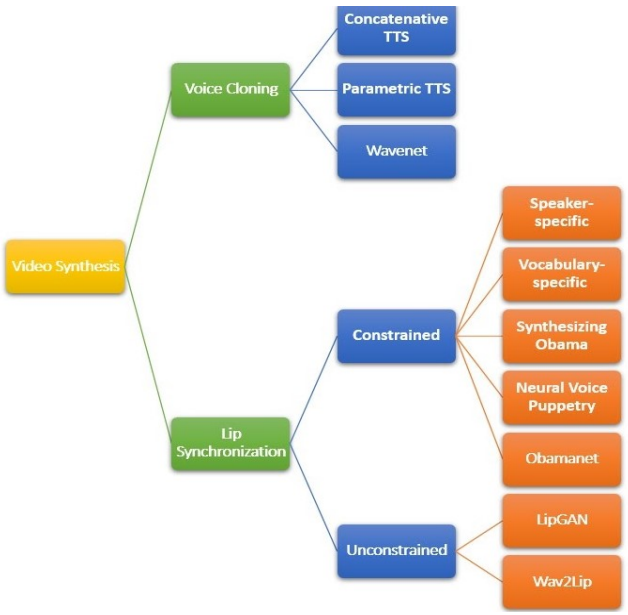

Figure 1. Different Techniques of Voice Cloning and Lip Synchronization

So, different methods and techniques are used to bring more naturalness to these audios.

There are different parameters to evaluate the quality of these TTS synthesizer systems like the preference of synthetic speech, naturalness, comprehensibility, and Intelligibility. Taking preference is the way where many test subjects are taken of all the genders and of different age groups. These subjects listen to natural and synthetic sounds and then the data of their choices are taken. The naturalness is measured by the quality of sound produced, the timing of words, the emotion in the words, and the correctness of the pronunciation. Similarly, comprehensibility is the perception of test subjects towards synthetic generated voices. Intelligibility is the way or quality of the word produced in the audio generated by the system.

3.1. Methods

With the development of science and technology, the approaches to speech synthesis are also evolving. Technology has also seen a growth in the field of data collection and data processing which has led to a boom in deep learning and artificial intelligence. The two main methods leading to the development of text-to-speech synthesis are – Concatenative TTS and Parametric TTS.

Concatenative TTS. Concatenative TTS is a technique of synthesizing sound by adding up short sounds of recorded audio clips that are combined together to form a speech. Many of speech fragments which are small in size are collected and added in the database.

These fragments are recorded from a single user. The speaker records a large number of voices from whole sentences to syllables and these speeches are further broken down and tagged and these segmented speeches are stitched together to form sentences [6]. One of the major advantages which this method/technique gives is that it preserves the speaker's voice as well as maintains high quality in terms of intelligibility. The limitation of using this technique is that the system is very slow and takes a lot of time in generating the output. This is because of the large number of databases and the hard coding involved to form these voices into sentences. Secondly, the sound is more robotic and less natural. Because of adding up different speeches to form a sentence the output is emotionless and lays stress on random parts of the string. It also failed on switching to different speakers, adding new words, and altering the emotion of speech [6].

Parametric TTS. In order to overcome the limitation of concatenative TTS, a new method is developed. It is called Parametric TTS. The idea here is to figure out the parameters behind a speech. Then we train our model based on these parameters to generate a speech. At first, the text is extracted of the linguistic features like phonemes, and Secondly, the extract vocoder features from it. Out of these parameters, we figure out the features that are important to recognize human speech and then train our model on these features and use it in audio processing. All data required for data processing is stored in the model parameters and the content and symbols can be controlled by modeling [7]. Parameters such as speech, basic frequency, sound levels, etc. vary in overtime performance. While producing a waveform, the vocoder alters features and measures speech parameters such as paragraph, speech rate, tone, and more. Using parametric TTS, we get better audio with more naturalness. Here we are able to generate audio invoices of different speakers. We have more flexibility to change the pitch of emotional change. It has less development time and can be trained on fewer data like we need around 4 hours of the sound of the speaker to train the model [7]. Parametric TTS also has limitations though it may sound more natural than concatenative, but it's more like a robot. For even simpler words, it becomes hard to design algorithms. The sound quality generated is less and contains noise and buzzy sound.

Wavenet. Wavenets are deep generative models. Generative models can generate new data from the data given to it while in the contrary discriminative models are one which classify objects. Generative models can also learn data like an object that looks like a "ship" will always be found near an object that looks like "water". Wavenets are auto-regressive. In this method, only the previous time-step is taken into consideration, not future time-steps. Wavenets are built on dilated CNN (Convolutional Neural Network), which is by increasing the dilation factor in CNNs(including larger previous inputs) the receptive field becomes larger [8]. By increasing the receptive fields we are able to collect information from a larger context. Voice Cloning is the new emerging technology in the current scenario. This is replacing the robotic sounds of robots with natural human voices generated from machines. With the entry of Artificial Intelligence in the field of Audio synthesis, it has achieved new heights. There is research being done to bring more naturalness to the generated audios.

3.2. Model Architecture

We are working to clone human voices in real-time with very short audio input from the user. The topic on which our research is based is Speech Vector to TTS or SV2TTS[12].

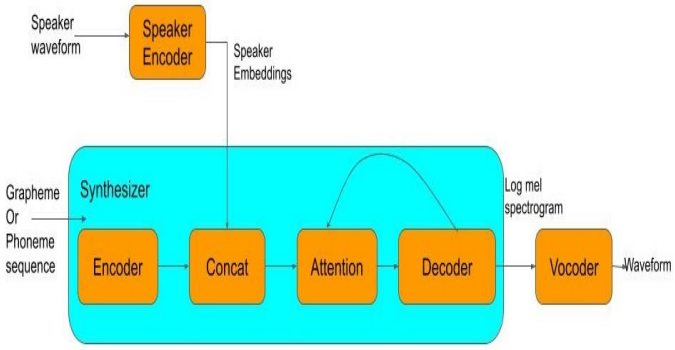

Figure 2. A General SV2TTS Architecture

The SV2TTS is mainly composed of three parts:

Speaker Encoder. This is the body text with no indent. This is the body text with no indent. This is the body text with no indent. This is the body text with no indent. This is the body text with no indent. This is the body text with no indent. Speaker Encode is the first part of the SV2TTS. It takes the audio input of the speaker, then captures various audio embeddings. The output encoder produces is focused on how the speaker sounds instead of what the speaker says. The output captures features like the voice of the speaker, its pitch, its tone, the accent, etc. While ignoring the words and the voice in the background. The encoder converts these features into low dimension vectors, also known as d-vectors. As a result of which the utterances of the speaker match only his/her voice and not of a different speaker. The embeddings are generated by breaking the

audio files in a 1.6-sound clip, which is then encoded as mel-spectrograms frames. As a result of this speaker, the encoder gets trained to recognize the audio of the speaker.

Synthesizer. The synthesizer takes user inputs and converts them into a mel-spectrogram. It takes the text input and generates a sequence . This sequence contains phonemes which are text mapped. The smallest unit of speech is called Phonemes. It separates one word from another as the word "n" in "fan" separates it from "fab", "fat", "fam".

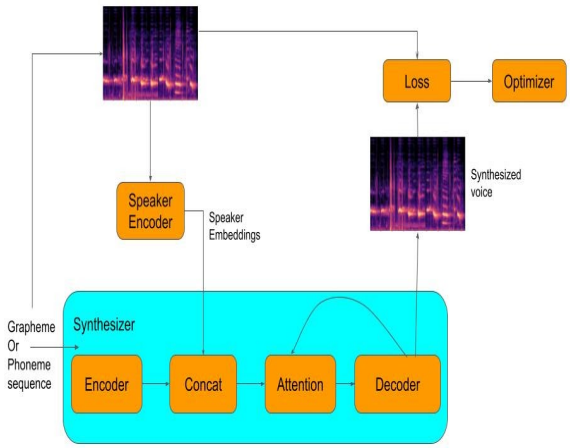

Figure 3. Synthesizer Training

The way synthesizer is trained is by providing a mel-spectrogram of speaker's audio to speaker encoder. Then speaker encodings are generated from speaker encoders. Now, the sequence of phonemes, generated from text provided to produce an audio of which in the sound of the speaker, is fed to the encoder inside the synthesizer. The encoder then concatenates the encoding of the text sequence with the speaker embedding. The decoder then generates a mel-spectrogram. The mel-spectrogram is now compared with the original spectrogram, the loss is calculated and then optimized.

Vocoder. Now, when mel-spectrogram is generated by the synthesizer, the vocoder then converts this spectrogram to audio format. Deepmind's Wavenet model is the vocoder used in this method, it outputs a large number of samples from raw audio extracted from the text, and is state-of-the-art for TTS systems.

3.3. Validation Method

In this section we have present validation methods.

Mean Opinion Score. Mean Opinion Score or MOS is a method where human subjects are exposed to multiple voices and then they rate these voices based upon their naturalness and their perception. MOS is a method of checking validity of generated audio by a blind test on humans. The following diagram shows the results of MOS and compares two different languages - US English and Mandarin Chinese. Several audios were generated from the Concatenative method, Parametric method, wavenet and human speech. The scale of rating naturalness of voice is from 0 to 5. The human speech value should be 5 but the difference in the number shows the error rate of this method. Accordingly, we can see the results of other methods in the following diagram. As a conclusion, wavenet has given much better performance than other methods.

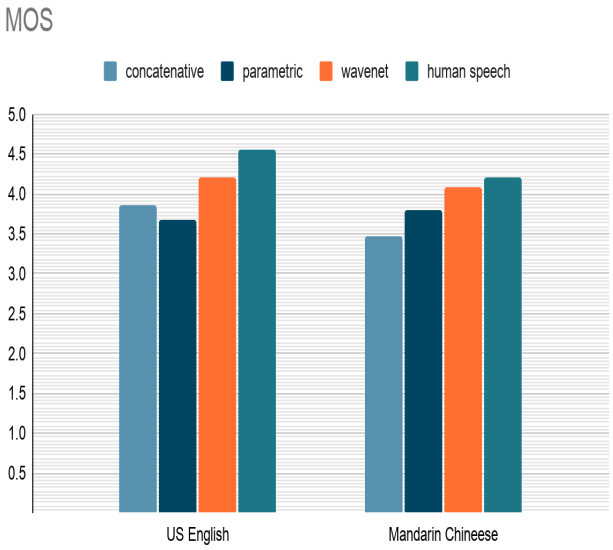

Figure 4. MOS of WaveNet's performance compared with a parametric and concatenative approach, as well as natural speech

3.4. Dataset Analysis

We are using the LibriSpeech dataset. The size of the dataset is around 305GB Approx. This is developed by OpenSLR (Open Speech and Language Resources). The Librispeech dataset is SLR12 which is the audio recording of English speech. The file format of data is in the form of FLAC(Free Lossless Audio Codec) without any loss in quality or loss of any original audio data [18]. The dataset contains more than 1000 hours of English speech at a 16kHz rate.

4. Lip Syncing

Lip syncing is matching the lips is a video to actual human lips movement for target audio. Recent years

have seen an increase in research in this domain due to the need for rapid content generation and translation techniques as the world shifts more towards digital norms. There has been research that works for specific targets [21, 22] as well as techniques working independently of speakers [23, 24]. Also, some work involves synthesizing lip-synced video from a single static image while others lip-sync video or dynamic content. Techniques working in the wild can work on speech generated by a text to speech systems [18, 24] as well. Also, lip-sync requires higher accuracy as humans can detect a video segment which is not in-sync even if unsynced video segment length is of 0.050.1 seconds [9] in duration easily.

4.1. Methods

Constrained Talking Face Generation From Speech. These are trained for a specific target person. The technique here maps the lip landmarks and waveforms of input audio. Also, these methods require large training data of specific people [18, 27], like Barack Obama. These cannot work for random audio and different people than trained data.We want to have a method working for any person and any audio data so this method will not work well for our use case. Many remarkable works have already been performed in this field. Following are a few of those existing methods.

1. Speaker-specific Talking Face Generation

   Generating talking face approach was first introduced in [4]. Long data was provided to the model of the specific person for whom the model should be trained. This approach was personalized for the Obama presidential address. So, the model trained with Obama for long hours of video and audio. And in the end, it generated the video with the word that has been provided by the text.

2. Vocabulary-specific Talking Face Generation

   This method uses a diversified person input so that the model can get the different audio feed but the problem is it has only a limited 50-100 words per speaker with a limited vocabulary. Using GRID [17] and TIMIT [19] dataset., this model was trained and evaluated. This approach was highly language-dependent.

3. Synthesizing Obama learning lip sync from audio

   For this method [16] they provided long audio and video clips of the ex-President Obama of the United States. The model created a high-quality auto-generated video with words that Obama has never spoken. The major constraint for this method was that it only works for Obama. But the performance and accuracy were high.

4. Neural Voice

   An audio driven video synthesis approach [30]. Providing the audio of the real person or any system generated, it results in photo-realistic video. This approach uses the 3D face model through a deep neural network. But this method works majorly for the specific person, not the generic audience

5. Puppetry Obamanet: Photo-realistic lip-syncing from text

   This method [29] works accurately with static images or video available during the training phase only. But this gets failed to provide accurate morphing of the lip movements for the random identities in the not static(dynamic), and not constrained videos for the talking face, which outputs the out lip synced video along with audio provided.

4.2. Unconstrained Talking Face Generation From Speech

These are independent of training data and can work for generic videos and audio[17,18]. This type of method is best for our use case as any user input can be transformed to desired high-quality videos. The model uses a GAN (Generative Adversarial Network) with a generator and discriminator. The model works quite well for not that dynamic videos. Upon investigating [5] the reason was inadequate discriminator loss function. The discriminator was only 56% accurate on the LRS2 dataset. Also, the discriminator was looking only at a single frame to detect out-of-sync video. This works quite well but due to furthermore research we intend to use more accurate architecture.

1. LipGAN

   Wave2Lip comes from the same researchers of LipGan and is improvised by making discriminators more accurate. This uses a 91% accurate discriminator on the same dataset making it the best approach for our use case. Also, it corrects lip-sync errors by using a modified version of SyncNet [28].

2. Wav2Lip

   Wave2Lip comes from the same researchers of LipGan and is improvised by making discriminators more accurate. This uses a

91% accurate discriminator on the same dataset making it the best approach for our use case. Also, it corrects lip-sync errors by using a modified version of SyncNet [28].

4.3. Comparison of methods

Comparison of all methods makes wave2lip the best approach for lip sync and best for our use case.

Paper	Speaker Independent	Language Independent	Smooth blending into video	Accurate for in-the-wild videos
Synthesize Obama	No	No	Yes	Yes
Obamanet	No	No	Yes	No
Puppetry	No	Yes	Yes	No
Talking Face	Yes	Yes	No	No
LipGAN	Yes	Yes	Yes	No
Wav2Lip	Yes	Yes	Yes	Yes

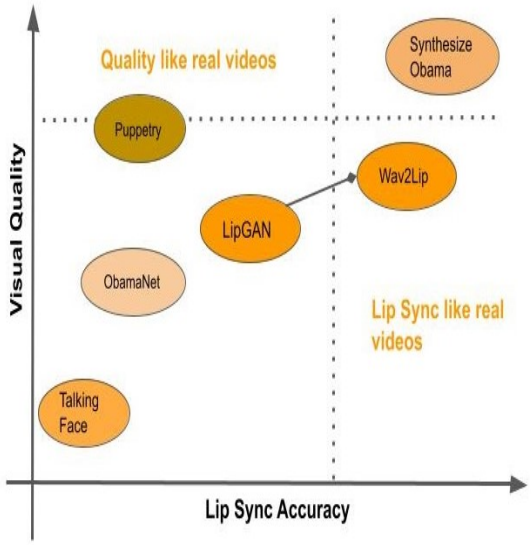

Figure 5. It shows comparative information between various existing methods. On the parameters of visual quality and lip-sync accuracy.)

5. Proposed Applications

Upon selecting the best methods of one-shot voice cloning and synchronizing the movement of lips, we will be able to synthesize the inputted videos according to the data accepted by the application. Using this video synthesizing technique certain applications can be made. Some of them are:

5.1. Customized Message Generation

Official Usage. During various corporate events and festivals, the leaders of the companies spend time wishing their employees. Due to shortage of time, every employee can't be wished by the company leader. Now, using the Custom video synthesizer which can be made using Voice Cloning and Lip-syncing, multiple videos can be generated by using the same video as input. The difference in every video will be of the name of employees. These names are taken as input from the user (here, Company leaders) as text input. These text inputs can be in the form of an excel sheet. Now multiple final videos would be generated which would be containing the voice of the user speaking the same words which were inputted. These videos will help users create multiple personalized videos very easily.

Personal Usage. Families and friends gather at festivals and various family functions. But generally, families live far away from each other due to various reasons. Now, what happens is people want to wish their loved ones to show that they care about each other. Due to busy schedules, people usually write a generic message in their family chat group. Wouldn't it be very impactful if every friend and family member receives a video wish message which is directly addressed to them. Hereby using the customized video synthesis tool this can be easily done just by feeding a generic video and list of names.

5.2. Personalized Advertisements

The advertisement industry has seen many revolutions and is still working over how to create more impact over users. Currently, the video ads which are made and published are very generic. This does not allow viewers to get connected to the product. Now, take a scenario where you are scrolling your social media feeds and the advertisement which comes contains your name every time there is a person addressed. That would be more impactful and the viewer would be more willing to watch that ad as this is directly talking to the user. With the custom video synthesis tool which uses voice cloning and lip-syncing, this can easily be done. The advertising company has to just submit the generic video. The names would be fetched from the social media username of the user.

5.3. Video Correction by video content creators

Recently we have seen a lot of video content made and uploaded across various streaming platforms. One of them is YouTube. Now, a lot of emerging YouTubers are coming as video accessibility of people has increased

due to increased ability to access good speed internet. When YouTubers are recording the videos there may be certain mistakes at some point. This will make Content Creators record the video again from the beginning. Another instance may be where a tech video is made and there is a certain update and YouTuber wants to update only that part of the video. So, whenever there is a need of correction in already recorded video, the customized video synthesis tool can be used.

5.4. Others

Apart from the usage mentioned above, some other usage can be like generating missing parts of the video call due to network connectivity. If a user is on a video call and suddenly due to poor internet connection, their video is not streaming properly but the audio of the user is captured. Then the missed video call part would be automatically created with the help of video synthesis.

6. Issues

The video synthesis tool clones the voice of the user. Also, to perfect the video synthesis, the lips are also synchronized. This imitation of personal features can raise certain social issues. Some of them are:

6.1. Legal Issues

Identity theft is a crime where one person impersonates others. Here similar voices of clients can be used where voice based recognition are there. Due to the high quality of video generation and the generated video being very close to the user's actual recorded video the tool can be misused by someone who wants to impersonate the original user.

6.2. Ethical Issues

The tool can be used to fool people. Sometimes people may get misconceptions about who actually spoke those words. Deceiving someone is plainly wrong. Defrauding people and creating fake news is ethically wrong. Impersonating someone can be a serious ethical issue in certain societies.

6.3. Moral Issues

While the tool is very productive for the people who want to reach people using very customised video messages without wasting much of their time. However, in certain relations, stratas of society using such tools can be seen as immoral. The emotions when talking to different people are different. These emotions can't be generated by our tool. Hence, having the same kind of content for various people can be seen as lack of morality in the user of the tool.

6.4. Trust Deficit

If a person's videos are easily generated. Then there would be a trust deficit in the society. People would not be able to believe in the accuracy of video.

7. Discussion

For voice cloning we have taken 3 methods. These methods are Concatenative TTS, Parametric TTS and waveNet. Concatenative TTS required large amounts of data and generated robot-like voices. To overcome limitations of Concatenative TTS [6], Parametric TTS [8] is used. Here parameters behind voice were identified. It generated more natural voice as compared to Concatinative but the generated voice contained some noise. At last waveNet [7] used an entirely different approach of deep-learning. Here, the voice generated is closer to the actual voice and isn't robot-like at all.

Furthermore, the research is going on to increase the accuracy. For Lip Synchronisation, the techniques are basically divided into two categories. First one is Constrained Talking face. The constraints here are speaker specificity and vocal specificity. Methods under this category are Synthesize Obama[16], Obamanet [29], Puppetry [30], Talking Face [4]. Another category of Lip Synchronisation is non-constraint methods. We have studied here two non-constraint methods which are LipGAN [24] and wav2lip [5]. Wav2lip is speaker independent. It also doesn't depend on languages spoken by the user. Best part about wav2lip is that it blends into videos smoothly and also works accurately for in-the-wild video.

8. Conclusion

Out of various methods to clone voice [3, 7, 8, 12, 18], the method which gives the most favorable outcome is "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (SV2TTS)". It overcomes certain drawbacks which were in previous methods for voice cloning. One of the major drawbacks is that the voice generated previously was robotic and monotonous. It lacks naturalness and does not take care of loudness and pitch of voice. Also, the previous methods required large amounts of data i.e. a large number of recorded voice of users is required. For making the video look more natural the method used for Lip Synchronising [4, 18, 24, 26, 29, 30] is the Wav2lip model. The advantages this model offers over the previous are that it is speaker independent. Wav2Lip works in the wild i.e for any user and any audio file. Apart from being the best methods till now, there are certain limitations. It works well for the English language but doesn't work well for other languages because of unavailability of data sets.

Also, there are some problems in reproducing the exact accent. One of the major features which can be introduced is emotional state cloning. The same statement can be said by a user feeling different emotions. The lip synchronizing makes the lip-synced according to the vocabulary, but it doesn't take care of the emotion of the user while speaking. These limitations could be worked upon to make a state of art video synthesis tool.

References

[1] Kapur, Radhika. (2018). Role of Media in the Development of Education.

[2] Stevens C. et al., "Online experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference," Computer Speech and Language, vol. 19

[3] Oloko-oba, Mustapha T.S, Ibiyemi Samuel, Osagie. (2016). Text-to-Speech Synthesis Using Concatenative Approach. International Journal of Trend in Research and Development. 3.559–462.

[4] C. Bregler, M. Covell, and M. Slaney. (1997) Video rewrite: Driving visual speech with audio. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pages 353–360.

[5] K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar, (2020), A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild.

[6] Oloko-oba, Mustapha T.S, Ibiyemi Samuel, Osagie. (2016). Text-to-Speech Synthesis Using Concatenative Approach. International Journal of Trend in Research and Development. 3.559–462.

[7] King, Simon.(2010) "A beginners ' guide to statistical parametric speech synthesis." .

[8] A. Oord, S. Dieleman, H. Zen, K. Simonyan,O. Vinyals, A. Graves,N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. (2016) Wavenet: A generative model for raw audio. arXiv:1609.03499

[9] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. (2018) Neural voice cloning with a few samples.

[10] Ehab A. Al Badawy Siwei Lyu.(2020) Voice Conversion Using Speech-to-Speech Neural-Style Transfer INTERSPEECH.

[11] Pisoni, D. B. et al., (1985) "Perception of synthetic speech generated by rule," in Proceedings of the IEEE, pp. 1665–1676.

[12] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu (2018), Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.

[13] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. (2020). Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Comput. Surv. 53, 3, Article 63, 34 pages. DOI:https://doi.org/10.1145/3386252

[14] Deep Voice 3: 2000-Speaker Neural Text-to-SpeechICLR 20182017 W. Ping Kainan PengAndrew GibianskySercan Ö. ArikA. KannanSharan NarangJonathan RaimanJ. Miller

[15] Diamos, G., Sengupta, S., Catanzaro, B., Chrzanowski, M., Coates, A., Elsen, E., Engel, J., Hannun, A., and Satheesh, S. Persistent (2018) RNNs: Stashing recurrent weights on-chip. In ICML, pp. 2024–2033.

[16] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. (2017) Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG), 36(4):1–13.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao. (2006) An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424.

[18] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech

[19] N. Harte and E. Gillen.(2015) Tcd-timit: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia, 17(5):603–615.

[20] Chung, J., Gulcehre, C, Cho, K., and Bengio, Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555.

[21] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. (2019). Text-based editing of talking head video.

[22] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. (2019). Neural Voice Puppetry: Audio-driven Facial Reenactment. arXiv preprint arXiv:1912.05566.

[23] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. (2019). You said that?: Synthesising talking faces from audio. International Journal of Computer Vision 127, 11-12, 1767–1779.

[24] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. (2019). Towards Automatic Face-to-Face Translation. In Proceedings of the 27th ACM International Conference on Multimedia. ACM, 1428–1436.

[25] J. S. Chung and A. Zisserman. (2016) Out of time: automated lip sync in the wild. In Asian conference on computer vision, pages 251–263. Springer.

[26] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. (2017). Obamanet: Photo-realistic lip-sync from text. arXiv preprint arXiv:1801.01442.

[27] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. (2017). Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) 36, 4

[28] Joon Son Chung and Andrew Zisserman. (2016). Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV.

[29] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio. (2017) Obamanet : Photo-realistic lip-sync from ´ text. arXiv preprint arXiv:1801.01442.

[30] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. (2019) Neural voice puppetry: Audio-driven facial reenactment. arXiv preprint arXiv:1912.05566.

[31] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. (2017) Deep voice 2: Multi-speaker neural text to-speech.