

Review on Speech Synthesis from Lip Movements using Deep Neural Network

Ms. Shilpa Sonawane¹ and Dr. P. Malathi²

¹Rajarshi Shahu College of Engineering, Tathawade, SPPU University, Pune, India
Email: Shilpa.sonawane8@gmail.com

²D.Y. Patil College of Engineering, Akurdi, SPPU University, Pune, India
Email: malathijesudason@ymail.com

Abstract—Speech synthesis using machine learning has received much attention in the past decades. Recent development in speech synthesis applications make use of deep learning techniques as it yields far better results when compared to others in variety of applications. The perception of speech signal is dependent on both acoustic and visual information. This paper is review of the current available databases, video and audio features for speech synthesis. In this paper, we describe speech reconstruction from lip movements based on deep neural network model. This paper also highlights the performance and limitations of available methods.

Index Terms— RNN, CNN, LPC, PESQ, MOS.

I. INTRODUCTION

Speech synthesis is joint embedding of auditory and visual information. The systems which control the computers using speech command are affected by ambient noise and acoustic conditions. The recent work focuses on both audio and visual modalities to recognize the speech commands. The performance is improved by developing the system which is not sensitive to ambient noise and change in acoustic conditions, does not require the user to make a sound such system is known as Visual speech recognition (VSR). VSR utilizes the visual information of the movement of the lips, teeth and somehow tongue of the speaker [9].

Speech synthesis from visual features has many applications in the field of biometric identification, silent-movie processing, silent dictation in public areas, speech transcription for cases where audio is not available and speech recognition in noisy environment [10]. One of the major issues to reconstruct the speech form visual knowledge is audio and lip motions are not consistent. It is a challenging task for machine lip reading system as it requires information of spatiotemporal features from the video [9].

The recent advances in machine learning are helpful in speech reading systems. In the past, most of these systems output text but there are many applications such as videoconferencing in noisy environments that would benefit from the use of video to speech systems. The possible way to develop video to speech system is text to speech conversion. The text-based systems require transcribed datasets which are hard to obtain as they require manual annotation and results in unnatural speech because such systems prevent from capturing emotions and intonation of the speaker.

This paper summarizes the related work of speech synthesis using deep learning model in Section II, whereas information regarding the methodology used is presented in Section III. Section IV demonstrates the results, where Section V concludes the paper.

II. RELATED WORK

There is considerable amount of literature on speech synthesis. Various approaches have been suggested the two phases of lipreading system: learning visual features, and prediction. Much work on the speech synthesis has been carried out which generates audio from text. In classical approach, sentence-level sequence prediction is done rather than word classification. In many cases it is found that unnecessary lag exists between what is spoken and what is seen. Thus, audio generated would not be in synchronous with the video and it leads to rejection in real-time applications [25]. Experiment is conducted on states of HMM and Gaussian mixture component in which information fusion with weighting factor is used to integrate the different lip features to achieve higher performance. The study shows that integrated discriminate vector after feature fusion outperforms than geometrical features vector [4].

The method for visual speech recognition is implemented from lip movement features. The visual features of mouth region of the speaker are extracted for better recognition rate. GLCM algorithm is implemented to extract four features like entropy, energy, contrast, and correlation. Artificial neural network classifier is used for classification of visual properties. It is found that the performance of the system is more accurate than the other conventional methods. It is mentioned that it can be possible to add both audio and video input characteristics to increase the performance of the system [15].

The speech reading system is implemented on segmented input video which contain either whole words from a predefined dictionary, or part of the words comprising phonemes. Then visual features are extracted from images and fed to classifier which classifies words or sentences. The author implemented video to speech reconstruction system which make use of convolutional neural network to generate the sound features based on neighbouring frames. Audio signal is generated from learned speech features [10]. A recent review of the literature on lipreading found that LipNet model is implemented that maps variable-length sequence of video frames to text using convolution network. LipNet learns spatial-temporal visual features and it shows higher performance for longer words [17].

Lip reading models, being recognition-based models, are constraint by a fixed size vocabulary. For similar sounding words, lipreading performance deteriorates. Recurrent neural network (RNN) has emerged as a useful tool for modelling such sequential data, but training these models requires, large scale annotated datasets. The proper synchronization of video with lip movements as well as reconstruction of the audio is required for real time application [2]. visual speech recognition requires modelling of lip-motions. Hence, an accurate structural representation of lips can be vital for improving their performances. There is difficulty to obtain pitch and intonation of the speaker form lip movement. They have to be artificially generated to produce intelligible results [17].

WaveNet uses autoregressive generation technique for speech generation in which each audio sample is generated from all the previous samples in the audio. The real recordings of human speakers are used for network training. After training, the network is sampled to generate utterances. At each step probability density of sample value is computed by the network. This value is fed back to the input to predict new sample. Generating samples one at a time leads to more time complexity, but it is essential for generating realistic-sounding audio [18].

A major issue of WaveNet model is it requires features from the existing TTS systems. This problem can be resolved by tacotron model in which model is trained on audio and text pairs.

So, it is adaptable to any dataset. The characters of the text are provided as input and forwarded them through different neural network submodules and produces spectrogram of the audio to generate the speech signal [16].

III. METHODOLOGY

The architecture of speech synthesis system using deep neural network consists of extraction visual and audio features, training of network and classification. The architecture of deep neural network is as displayed in Fig. 1. The input to the model is images of speaker speaking something and their corresponding audio. A convolution operation is used to find local visual features from input images. The size of visual feature is huge. Pooling operation is used to reduce the dimensionality of visual features which leads to less amount of time for network training. In training, the network learnt visual features for that specific subject and maps them to the corresponding audio.

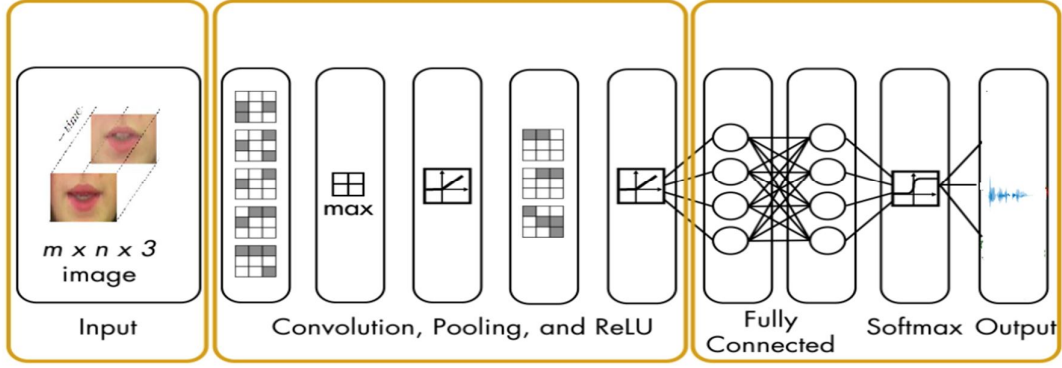


Fig. 1 Deep neural architecture for speech synthesis

A. Database

Speech synthesis researchers mostly worked on GRID corpus and OuluVS2 database. OuluVS2 is an audio-visual database of 53 speakers with multiple views. The database consists of speakers uttering ten short English phrases are listed in Table I.

TABLE I: LIST OF PHRASES UTTERED BY SPEAKERS IN OULUVS2

Sr. No.	Phrases	Sr. No.	Phrases
1	Excuse me	6	See you
2	Goodbye	7	I am sorry
3	Hello	8	Thank you
4	How are you	9	Have a good time
5	Nice to meet you	10	You are welcome

GRID is a large multi speaker audiovisual sentence database which consists of audio and video recordings of 1000 sentences spoken by each of 34 talkers with 18 males and 16 females. Video files are provided in two formats such as normal quality (360x288) and high quality (720x576). Each video has a duration of 3 sec and a frame rate of 25 frames per second. The audio and video tracks have the same duration with 50 kHz sample frequency [10]. TCD-TIMIT is rich database consists of audio and video recordings of 62 speakers uttering 6913 sentences. Video was recorded from two angles: straight on, and at 30 degrees [7].

B. Visual and Audio Features

The research in speech synthesis shows that visual feature extraction of lips, mouth etc is time consuming procedure. In deep neural network, convolution layer is responsible for extraction of appropriate visual features from the images [1]. In order to reconstruct the audio signal from image, CNN learns the essential visual features. Recent research shows that Linear Predictive Coding (LPC) is used to generate speech signal. The key parameter is order of LPC for generation of the best quality speech. Spectrogram of audio signal is useful to find parameters like frequencies and their bandwidth. These parameters are utilized in speech modelling techniques to reconstruct the speech signal [23].

IV. RESULTS

The performance of speech synthesis system is evaluated by various parameters. The quality and intelligibility of the speech can be evaluated by Perceptual Evaluation of Speech Quality (PESQ) and mean opinion score (MOS). Tacotron achieves an MOS of 3.82, which is better than the parametric system [16]. WaveNet achieved 5-scale MOS in naturalness above 4.0, which were significantly better [18]. To measure the performance of LipNet, the word error rate (WER) and the character error rate (CER), standard metrics for the performance of ASR models are calculated. It achieved an average of 47.7% WER.

V. CONCLUSION

This paper has presented a review on the speech synthesis using deep learning techniques. A popular speech synthesis database, GRID was used in many research work. But it consists of extremely small vocabulary. In

most cases, the network model is tested in both speaker dependent and speaker independent scenarios. In both cases, estimated speech quality and intelligibility is generally higher due to deep convolutional neural network. The future work is increasing intelligibility of speech reconstruction from more training data including emotions and to test a model to unknown speakers. It can also be used in speech enhancement domain.

REFERENCES

- [1] Konstantinos Vougioukas, Pingchuan, Stavros Petridis, and Maja Pantic, "Video-Driven Speech Reconstruction using Generative Adversarial Networks", Conference of International Speech Communication Association (ISCA), 2019.
- [2] Amir Jamaludin, Joon Son Chung, Andrew Zisserman, "You Said That?: Synthesizing Talking Faces from Audio", Springer International Journal of Computer Vision, 127:1767–1779, 2019.
- [3] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2516-2520.
- [4] Y. Kumar, R. Jain, M. Salik, R. r. Shah, R. Zim-mermann, and Y. Yin, "MyLipper: A personalized system for speech reconstruction using multi-view visual feeds", IEEE International Symposium on Multimedia (ISM), Dec 2018, pp. 159-166.
- [5] Khwaja Mohd. Salik, Yaman Kumar, Rohit Jain, Swati Aggarwal, Rajiv Ratn Shah, Roger Zimmermann, "Lipper: Speaker Independent Speech Synthesis Using Multi-View Lipreading", Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI-19)
- [6] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 6548-6552.
- [7] Yaman Kumar, Mayank Aggarwal, Pratham Nawal, "Harnessing AI for Speech Reconstruction using Multi-view Silent Video Feed", ACM Multimedia Conference on Multimedia Conference, 2018, pp 1976-1983.
- [8] Adriana Fernandez-Lopez, Federico Sukno, "Survey on Automatic Lip-Reading in the Era of Deep Learning", Imavis (2018), doi:10.1016/j.imavis.2018.07.002
- [9] Joon Son Chung, Andrew Zisserman, "Learning to Lip Read Words by Watching Videos", Elsevier Computer Vision and Image Understanding journal, 1077-3142, Volume 173, August 2018, Pages 76-85.
- [10] A Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 5095-5099.
- [11] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," ICCV 2017 Workshop on Computer Vision for Audio-Visual Media, 2017.
- [12] T. Le Cornu and B. Milner, "Generating intelligible audio speech from visual speech," IEEE/ACM Transactions on Audio, Speech, and Language Pro-cessing, vol. 25, no. 9, pp. 1751-1761, Sep. 2017.
- [13] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in British Machine Vision Conference, London, September 2017.
- [14] Supasorn Suwajanakorn, Steven M. Seitz, And Ira Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync from Audio", ACM Transactions on Graphics, Vol. 36, No. 4, Article 95, July 2017.
- [15] Amaresh P Kandagal ,V. Udayashankara , "Visual Speech Recognition system is implemented based on lip movement for Indian Languages", International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 2029-2041.
- [16] Y Wang, RJ Skerry-Ryan, D Stanton, Y Wu, RJ Weiss, N Jaitly, Z Yang, "Tacotron: Towards end-to-end speech synthesis", arXiv preprint arXiv:1703.10135, 2017.
- [17] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading", arXiv preprint arXiv:1611.01599, 2016.
- [18] Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio", A Van Den Oord, S Dieleman - arXiv preprint arXiv:1609.03499, 2016.
- [19] T. L. Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.
- [20] Itimad Raheem Ali, Ghazali Sulong & Hoshang Kolivand, "Realistic Lip Syncing for Virtual Character Using Common Viseme Set", Computer and Information Science; Vol. 8, sNo. 3; 2015 ISSN 1913-8989, E-ISSN 1913-8997.
- [21] Wesley Mattheyses, Werner Verhelst, "Audio-Visual speech synthesis: An overview of State-of-the-art", Speech communication 66,182-217, 2015.
- [22] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, Tetsuya Ogata, "Audio-visual speech recognition using deep learning", Springer Science Business Media New York 2014, DOI 10.1007/s10489-014-0629-7.
- [23] Mayank Garg, Rahul Singhal, "Speech Re-synthesis from Spectrogram Image through Sinusoidal Modelling", 2014 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), 978- 1-4799-3080-71 14
- [24] Wang Mengjun, "Geometrical and Pixel Based Lip Feature Fusion in Speech Synthesis System Driven by Visual-speech", Second International Conference on Computational Intelligence and Natural Computing, 2010.
- [25] James D. Edge, Adrian Hilton, and Philip Jackson, "Model-Based Synthesis of Visual Speech Movements from 3D Video", Hindawi Publishing Corporation EURASIP Journal on Audio, Speech, and Music Processing Volume 2009, Article ID 597267, 12 pages doi:10.1155/2009/597267.

- [26] Amar A. El-Sallam and Ajmal S. Mian, "Speech-Video Synchronization Using Lips Movements and Speech Envelope Correlation, Springer ICIAR 2009, LNCS 5627, pp. 397–407, 2009.
- [27] Sascha Fagel, G'érardBailly and Barry-JohnTheobald, "Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation", EURASIP Journal on Audio, Speech, and Music Processing Volume 2009, Article ID 826091, doi:10.1155/2009/826091.
- [28] G. Bailly, M. B'Erar, F. Elisei and M. Odisio, "Audio visual Speech Synthesis", International Journal of Speech Technology 6, 331–346, 2003.