LipSpeaker: Helping Acquired Voice Disorders People Speak Again

Yaohao Chen[1], Junjian Zhang[1], Yizhi Zhang[2], and Yoichi Ochiai[1]

[1] Digital Nature Group, University of Tsukuba, Ibaraki, Japan
{yaohao.chen, tyookk, wizard}@digitalnature.slis.tsukuba.ac.jp
[2] Applied Analytics, Columbia University

Abstract. In this paper, we designed a system called LipSpeaker to help acquired voice disorder people to communicate in daily life. Acquired voice disorder users only need to face the camera on their smartphones, and then use their lips to imitate the pronunciation of the words. LipSpeaker can recognize the movements of the lips and convert them to texts, and then it generates audio to play.
Compared to texts, mel-spectrogram is more emotionally informative. In order to generate smoother and more emotional audio, we also use the method of predicting mel-spectrogram instead of texts through recognizing users' lip movements and expression together.

Keywords: Accessibility, Disabled people, Lipreading

1 Introduction

Currently acquired voice disorder people communicate with others mainly in three methods. The first is through sign language. The second is communicate through paper and pen. The last method is to use smartphones or computers as medium.

However, all three methods have their own flaws. The first method requires others to be proficient in sign language but few people know sign language. The second method requires literate people but not all the people are literate. Furthermore, it is inconvenient to create writing environment on the road. The third method requires the users to master the basic keyboard input which is not applicable to all the people.

In order to solve the above problems, we have designed a new interactive solution- LipSpeaker: a system that uses the movements of the user's lips to generate speech. What the user need to do is simply face the camera on his smartphone. LipSpeaker uses the facial landmark detector to capture images of the user's lips. With the time sequence frame of the lips captured as input, the deep neural network can generate the text of the user's speech. With LipSpeaker, acquired voice disorder users can communicate with other people without the need for sign language, literacy or any keyboard input.

2 Related Work

Benefited from the development of deep neural networks in recent years, the field of lipreading has also been greatly developed. Among the well-known contributions are LipNet [1] of Yannis.M.Assae et al. and Lip Reading in the Wild [2] by Joon Son Chung et al.

Word Error Rate (WER) is an important indicator in training and evaluating the accuracy of lipreading using GRID corpus [3] dateset. Compared to M.Wand's WER in Lipreading with long short term memory [4] in 2016 is only 20.4%, the WER in LipNet is 4.8%. In Lip Reading in the Wild, the WER even reached 3.0%. At the same time, Yannis.M.Assae et al. indicated that the accuracy of lipreading using deep neural network is 4.1 times higher than that of artificial lipreading. Proving that using deep neural networks to predict text through the movement of the lips is feasible.

At this stage, we also use the GRID corpus dataset to train deep neural networks. The text results are predicted by inputting the motion sequence frame of the user's lips into the lipreading deep neural network model, and the audio playback is synthesized through Text-To-Speech (TTS) system. However, there is a disadvantage in generating audio in this way: the tone of the user's speech will be filtered directly into a text while our ultimate goal is to generate emotional audio based on the user's lip movements.

Inspired by Tacotron2 [5] by Jonathan Shen et al., we are trying to predict mel-spectrogram using the lipreading deep neural network model instead of predicting texts. Together with mel-spectrogram, we can generate emotional audio with WaveNet [6].

3 IMPLEMENTATION

The implementation of LipSpeaker is divided into two major steps: training phase and evaluation phase. Training phase runs on Ubuntu. We use tensorflow as a framework for deep learning, training the lipreading deep neural network with the GRID corpus dataset, and obtaining a well-trained model after training. In the evaluation phase, we convert the well-trained model into the model of Apple's deep learning framework Core ML, and then run the model onto the phone.

Since loading and running a well-trained model on the smartphone produces a relatively large amount of computation, it will result in high performance requirements for the smartphone. In order to alleviate the burden of computation on Text-To-Speech (TTS), we use Apple's AVSpeechSynthsizer provided by AVFoundation to generate audio. Compared to the TTS system such as Tacotron2, AVSpeechSynthsizer improves the generation speed and reduces the amount of computation at the expense of audio fluency and naturalness. AVSpeechSynthsizer is sufficient at this stage to verify the validity of the system.

In order to improve the accuracy of lipreading. In the pre-processing, we did mouth detection on the input image and cropped the user's mouth area as input.

Since the training device is PC and the actual running device is smartphone. In order to ensure the consistency of mouth detection results between training and running, we use dlib [7] to crop the position of the lips.

In terms of deep neural network models, we have adopted a network structure similar to LipNet since it has three advantages over the network structure of Lip Reading in the wild. First, the overall result it gets is better. Second, its network structure is simpler and the amount of computation is much less, and third, the network structure is End-To-End, which is more suitable for running on smartphones.

See the fig.1 for the specific Network Architecture.We adopt Connectionist Temporal Classification [8] (CTC) loss as our loss function. For optimizer we use AdamOpimizer.

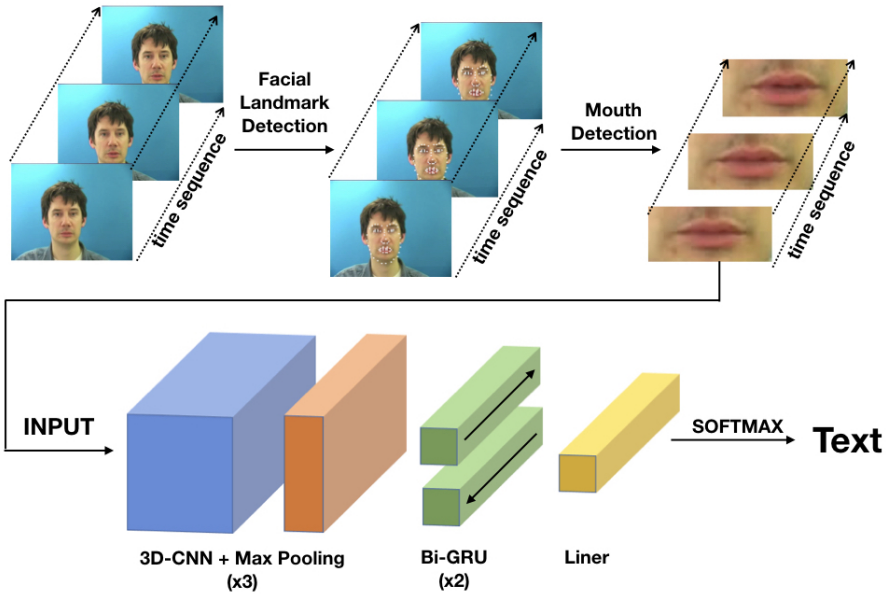

Fig. 1. Lip Reading Deep Neural Network architecture.

The accuracy of the model after training is similar to that in the LipNet. The WER of the prediction results for the overlapped speaker is about 7%, and the number for the Unseen Speaker is about 14%.

4 Future Works

We will use the same evaluation method as TTS to verify the validation of LipSpeaker using the Mean Opinion Score (MOS). In this experiment, each group

will consist of an acquired voice disorder participant and a non-disabled participant. The two users will communicate in three ways: pen and paper, keyboard input and LipSpeaker. Both users in each group score the three methods respectively with the score ranges from 1 to 5, and each group generating six numbers. Through multiple sets of experiments, the MOS of each communication method is calculated and compared to verify the effect of LipSpeaker on the acquired voice disorder participants and non-disabled participants.

As mentioned in section 2: predicting text with lip motion sequence frames can result in loss of features of the user's emotions. Therefore, we have improved the network structure inspired by Tacotron2. Tacotron2 is a system for TTS that consists of two major parts. The first part uses deep neural network to predict mel-spectrograms through text sequences. The second part uses the obtained mel-spectrogram to generate audio through another deep neural network WaveNet.

Since the mel-spectrogram is richer in the amount of features carried by texts, the generated audio from mel-spectrogram can better restore the user's emotions. Therefore, we are trying to predict the mel-spectrogram through deep neural network using the lip motion sequence frame as input. Lastly we use the trained WaveNet to generate audio with intonation. See the fig.2 for the specific Network Architecture.

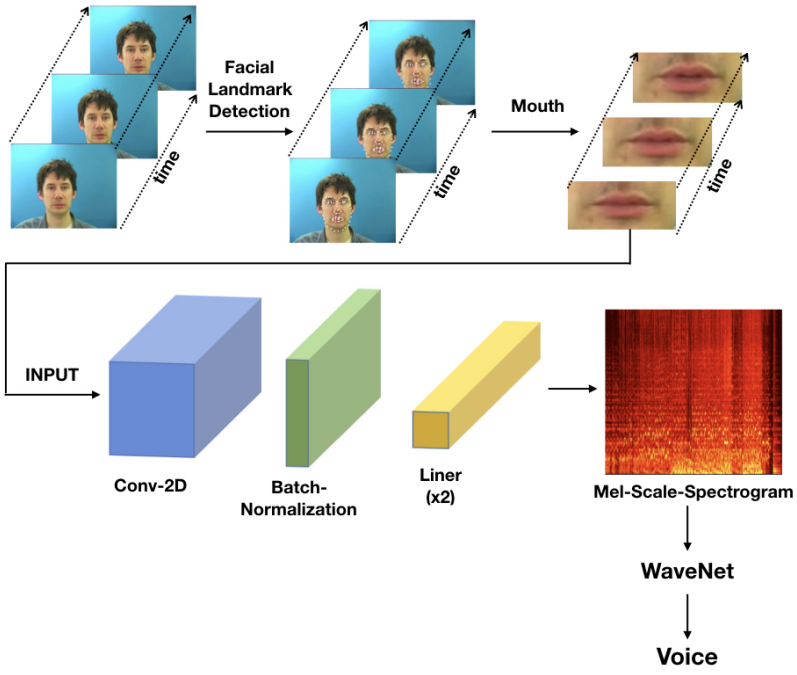

Fig. 2. The approach of predict mel-spectrogram by deep learning network architecture.

Natoki Kimura et al.'s SottoVoce [9] successfully predicted the Mel-scale spectrum using the ultrasound picture sequence frames of the tongue. Since the lip motion sequence frame has more features than ultrasound picture sequence frames of the tongue, we highly believe it is feasible for lip motion sequence frame. Up until now, we have tried to generate mel-spectrogram using 3D Convolutional Neural Networks [10] (3D-CNN). In order to achieve the expected accuracy, further experiments are still needed.

5 Conclusion

The system LipSpeaker we designed shows a new way of human-computer interaction. LipSpeaker can use the deep neural network to predict text by identifying the lip motions of the acquired voice disorder people and use it to generate speech in conjunction with TTS. LipSpeaker can help acquired voice disorders people to communicate more easily with others in their daily life. The WER of the model reached 7% in the laboratory environment, demonstrating the effectiveness of the method.

However, at the same time, the model is greatly affected by the environment. In the case of poor lighting conditions or the user's lip pictures are not clear, the accuracy will drop dramatically. The reason may be that the trained GRID corpus datasets data is obtained in an environment where the light is always sufficient and the participant is always facing the camera at the front face. In future work, we will try to add more training data to improve this situation.

Since the mel-spectrogram is richer in the amount of features carried by the text, the generated audio by mel-spectrogram can better restore the user's feelings. Inspired by the network structure of Tacotron2, we proposed to predict the mel-spectrogram through the lip motion sequence frame and use WaveNet to generate smoother audio with more intonation. In order to achieve the expected accuracy, we will conduct further experiments based on 3D-CNN.

References

1. Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. LipNet: End-to-End Sentence-level Lipreading. *arXiv e-prints*, page arXiv:1611.01599, Nov 2016.
2. Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip Reading Sentences in the Wild. *arXiv e-prints*, page arXiv:1611.05358, Nov 2016.
3. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Acoustical Society of America Journal*, 120:2421, 2006.
4. Michael Wand, Jan Koutník, and Jürgen Schmidhuber. Lipreading with Long Short-Term Memory. *arXiv e-prints*, page arXiv:1601.08188, Jan 2016.
5. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv e-prints*, page arXiv:1712.05884, Dec 2017.

6. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv e-prints*, page arXiv:1609.03499, Sep 2016.
7. Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, December 2009.
8. Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA, 2006. ACM.
9. Jun Rekimoto Naoki Kimura, Michinari Kono. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. *ACM CHI*, 2019.
10. S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan 2013.