Cover Sheet

Title of Proposed Project
Is it fake or not

Project Summary

Speech perception is a crucial function of the human auditory system, but speech is not only an acoustic signal-visual cues from a talker's face and articulators (lips, teeth, and tongue) carry considerable linguistic information. These cues offer substantial and important improvements to speech comprehension when the acoustic signal suffers degradations like background noise or impaired hearing. However, useful visual cues are not always available, such as when talking on the phone or listening to a podcast.. In this project, we consider a task of such: given an arbitrary audio speech and one lip image of arbitrary target identity, generate synthesized lip movements of the target identity saying the speech. To perform well in this task, it inevitably requires a model to not only consider the retention of target identity, photo-realistic of synthesized images, consistency and smoothness of lip images in a sequence, but more importantly, learn the correlations between audio speech and lip movements. Our final goal is training a model which is robust to lip shapes, view angles and different facial characteristics (e.g. beard, hair). Solving this task is crucial to many applications, e.g., enhancing speech comprehension while preserving privacy. Another application is that we can synthesize the cartoon images to make sure their facial movements are consistent with the speech. Currently in cartoon movie, the facial movement of characters is not paired with the sound at all.

Table of contents

Cover Sheet 1
Project Summary 2
Table of Contents 2
NSF Proposal 3
1. Objectives and novelties of the proposed work. 3
2. Potential benefits of the proposed work..... 3
3. Methods.....3
3.1 Image data preprocess..... 3
3.2 Audio data prepross 4
3.3 Information fusion 4
3.4 3D video generation.....4
4. Dataset.....4
References: 5

NSF ITRG Project Proposal

September 06th 2018

1. Objectives and novelties of the proposed project

There are many computer vision research [1,2,3] work on multi-modal tasks. For example, generating images from text, generating images from skeleton , object detection for autonomous driving using Lidar point cloud and camera images. However, how to get paired data is a main challenge in those multimodal research. For example, when we use lidar point cloud and rgb image to do object detection or localization, calibrating lidar [4] and camera will cost plenty of time and course lot's of problems like data un-synchronization, different receptive field for different machines. Same as skeleton/text and image, we need to annotate tons of texts or skeletons with paired image. But in the wild world, there are two modalities that always well-aligned and easy to get. Audio and visual data. For instance, when you here sound of the train whistle, you can also see that the training is coming, when it rainy, you can also hear the rhythm of the sound of raindrops. When you hear a speech from someone, you can see their mouse or facial changes along with their sound.

To perform well in this task, it inevitably requires a model to not only consider the retention of target identity, photo-realistic of synthesized images, consistency and smoothness of lip images in a sequence, but more importantly, learn the correlations between audio speech and lip movements. And our final model should be robust to lip shapes, view angles and different facial characteristics.

2. Potential benefits of the proposed work

This research project will benefits research area and industrial area. As to research society, currently there is no efficient and high quality method to fuse the information from two different modalities. If we fuse two information in low level (raw data), we need to consider different data structures and it's is really computation consuming. If we fuse the information in high level, we will sacrifice many important information. So in this project, we will discuss different fusion methods and evaluate their performance. Those fusion methods can be transferred to other multimodal tasks. As to industry society, our model is designed to run in real time so that it can be used as an assistive listening device.

3. Methods

3.1 Image data preprocess. The correlation between speech and image is on human face. So other part of the image (e.g. background, chair, hair, clothe) are redundant information and

sometimes will fuse the model. So we need to crop the face region from the original image frames and apply affine transformation to align the images.

3.2 Audio data prepross. There are many redundant information in raw audio data. And it is hard for us to extract valid information from raw audio data. The proposed solution is that we can apply Short Time Fourier Transformation or other methods to transfer raw audio data into time-frequency pattern (see Figure 1). Then we can easily extract meaningful information from time frequency pattern.

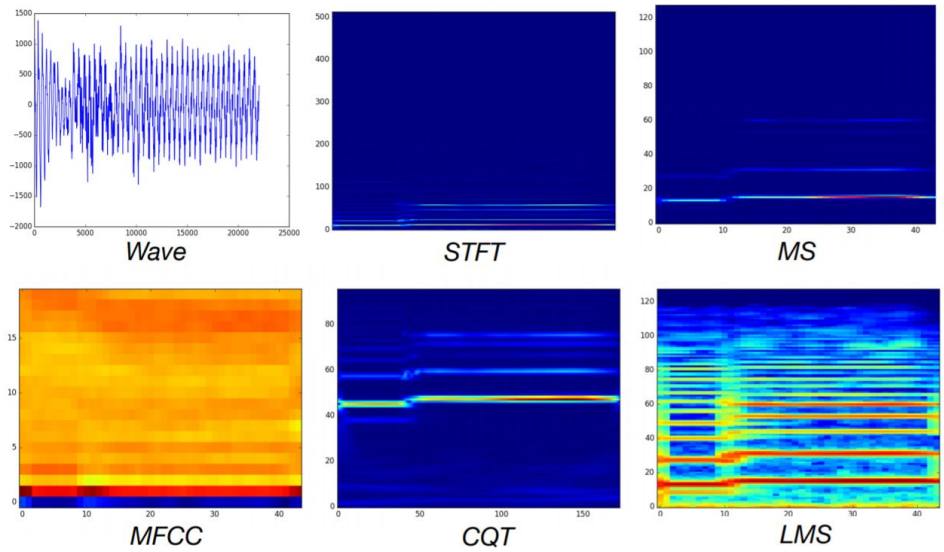

Figure 1: Outputs of different transformation on raw wave data [11]

3.3 Information fusion. The main challenge for this project is how can we fuse the information from two different modalities. There are many existing methods[1, 2, 3] work on this problem. However those methods can not be adapted in our project directly. The information space of audio signal is $\mathbb{R}^{f \times t}$. f represent frequency information and t indicate temporal information. However, the information space of one image is $\mathbb{R}^{h \times w}$. h and w are image size. We find that those two data are in different space, so we need to model the part and then we can do fusion. Currently my proposed fusion method including concatenation, addition, convolution operation. We can discuss it after some thoughtful experiments.

3.4 3D video generation. After fusion, we need to use optical flow [5] to estimate the motion from fused information. The estimated motion is the input of the generation network. In last four years, Generative adversarial network [6] shows its capability in generation tasks. We will adapt this idea and design our own 3D GAN structure with audio visual correlation loss.

4. Dataset

There are many face datasets [7,8,9], but there are not too much dataset contain paired audio and 3D meshes. So in this project, we can use some existing state-of-the-art method to generate fake data for training. For example, we can download the BBC news dataset [10].

There are paired images and audio signal. Then we can estimate the 3D meshes from RGB image (see Figure 2).

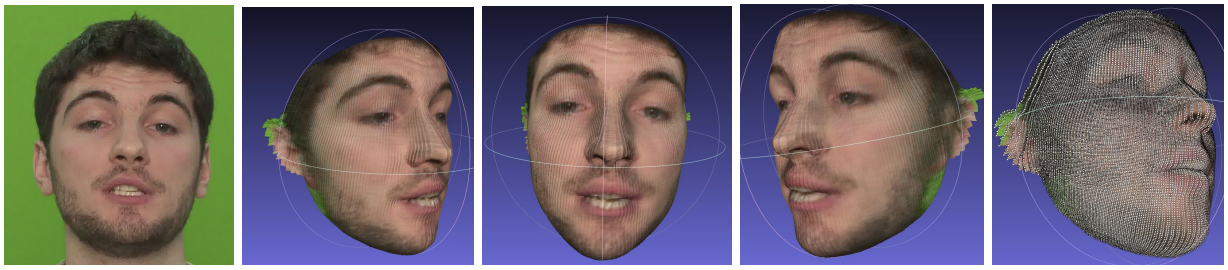

Figure 2: The output example of 3D meshes from one single RGB image

Reference

[1] D. Lahat, T. Adal, C. Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. In Proc. of the IEEE, Institute of Electrical and Electronics Engineers, 2015, Multimodal Data Fusion.
[2]] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In Proc. NIPS, 2012
[3] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl., 2000.
[4] Li, G.; Liu, Y.; Dong, L.; Cai, X.; Zhou, D. An Algorithm for Extrinsic Parameters Calibration of a Camera and a Laser Range Finder Using Line Features. In Proceedings of IEEE International Conference on Intelligent Robots and Systems, 2007
[5] B.K.P. Horn and B.G. Schunck. Determining optical flow. Artificial Intelligence, 17:185–203, 1981.
[6] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)
[7] http://www.robots.ox.ac.uk/~vgg/data/vgg_face/
[8] http://www.wilmabainbridge.com/facememorability2.html
[9] https://faces.dmi.unibas.ch/bfm/
[10] http://mlg.ucd.ie/datasets/bbc.html
[11] L. Chen, S. Srivastava, Z. Duan and C. Xu. Deep cross-modal audiovisual generation. arXiv preprint arXiv:1704.08292, 2017.