Lipper: Speaker Independent Speech Synthesis Using Multi-View Lipreading

Khwaja Mohd. Salik
MIDAS Lab, NSUT-Delhi
khwajam.co@nsit.net.in

Yaman Kumar
MIDAS Lab, IIIT-Delhi
yaman@nsitonline.in

Rohit Jain
MIDAS Lab, NSUT-Delhi
rohitj.co@nsit.net.in

Swati Aggarwal
NSUT-Delhi
swati@nsit.ac.in

Rajiv Ratn Shah
MIDAS Lab, IIIT-Delhi
rajivratn@iiitd.ac.in

Roger Zimmermann
NUS, Singapore
rogerz@comp.nus.edu.sg

Abstract

Lipreading is the process of understanding and interpreting speech by observing a speaker's lip movements. In the past, most of the work in lipreading has been limited to *classifying* silent videos to a *fixed* number of text classes. However, this limits the applications of the lipreading since human language cannot be bound to a fixed set of words or languages. The aim of this work is to reconstruct intelligible acoustic speech signals from silent videos from various poses of a person which Lipper has never seen before. Lipper, therefore is a **vocabulary and language agnostic, speaker independent and a near real-time model that deals with a variety of poses of a speaker**. The model leverages silent video feeds from multiple cameras recording a subject to generate intelligent speech of a speaker. It uses a deep learning based STCNN+BiGRU architecture to achieve this goal. We evaluate speech reconstruction for speaker independent scenarios and demonstrate the speech output by overlaying the audios reconstructed by Lipper on the corresponding videos.

Introduction

Lipreading is the process of understanding and interpreting speech by observing a speaker's lip movements. Lipper is a *multi-view speech reconstruction* model unlike most of the work in this field in the past which deal with classifying speech videos into restricted text classes. It is easy to get confused between speech reconstruction, recognition and reading systems. Speech recognition systems help in identification of the speaker of a speech. Speech-reading systems involve identifying **what** a person says. The 'what' part in these classification models is given by text-class of the video. Speech-reconstruction systems **generate** the *speech* of a person. They deal with the problem by considering it as a regression and not a classification problem. Speech-reconstruction normally does not identify speech but just reconstructs it. The reason for this is that speech can be produced even for those sounds for which one may not have any vocabulary in a particular language (e.g, a Television plays the sounds in a show without knowing the language or vocabulary). Due to these reasons, speechreading models suffer from the following limitations: non-real time output and language and vocabulary dependency. Speech-reconstruction systems are not marred by these since they map lip movements directly to sound.

Lipper, to the best of our knowledge, is the **first system** to attempt **speaker independent multi-view speech reconstruction**. Both *speech reconstruction using lipreading* and *multi-view lipreading* have not seen much research. There have been only three deep learning based previous works which have leveraged multiple views to build lipreading based speechreading (and **not speech reconstruction**) models (Petridis et al. 2017; Zimmermann et al. 2016; Lee, Lee, and Kim 2016). These systems being speechreading systems, suffer from all the constraints of a speechreading system as pointed out above. Moreover, few authors have worked on speech reconstruction systems (Ephrat and Peleg 2017; Cornu and Milner 2015; Kumar et al. 2018a; 2018b) and the major limitations with these models were, neither did they work for speaker-independent settings nor did they utilize multiple views. Due to these shortcomings, they effectively ignored the pose problem with real-world visual feeds. We address that problem and also produce the results on all speakers without re-training the model from scratch for every new speaker we have to use it for. For the complete results and explanation of the model, readers are encouraged to refer to (Kumar et al. 2019).

Lipper: Design and Development

Lipper uses deep learning based neural network model for speech reconstruction. We build STCNN+BiGRU (Spatio-Temporal CNN in conjugation with Bidirectional Gated Recurrent Units) architecture (as shown in Figure 1) to utilize multiple visual feeds of different views to finally reconstruct the speech of a speaker. While STCNN layers help the system to extract visual features, BiGRU layers help it to take care of the time dependencies in the speech videos. The architecture is composed of seven layers of STCNN followed by two layers of Bi-GRU layers whose output is subsequently fed to a dense layer that produces the final output.

Audio Features: Raw audio wave cannot be used for training of the model due to a lack of suitable loss function. Lipper uses Linear Predictive Coding (LPC) for representing the audio speech (Fant 2012). LPC is a technique to represent the compressed spectral envelope of speech using a linear predictive model. It produces high-quality speech using a low bit-rate. The order P of LPC was varied and chosen so as to get the best quality speech. Through experiments, the LPC order was found to be giving optimal results at the value 24.

Database

We use the speakers present in the OuluVS2 database for the current analysis (Anina et al. 2015) as was used in (Kumar et al. 2018a). Cameras recorded the subjects present in the dataset from five different angles: 0°, 30°, 45°, 60°, and, 90°. The speakers have different ethnicities thus allowing Lipper to be trained for various accents, tones, *etc.*

Training-Testing Configuration

For the speaker-**independent** experiments, we trained the model on all the videos for the fifty speakers and tested the system on all the videos for the rest two.

Single Stream Training Each visual input stream is first trained independently. Each network corresponding to the five views is trained for 80 epochs with a batch-size of 10. With experimentation, the timesteps parameter of BiGRU was set to 5. Once the output is obtained, it is then decoded and compared with original audio for its quality.

Multi-Stream Training Once the single streams have been trained, then the obtained individual networks are stacked together with the outputs of BiGRUs concatenated to perform multi-view training. Finally, the entire network thus obtained was trained jointly on the multi-view visual input feeds. The output thus obtained from the last layer was decoded and compared with the original audio.

Speaker Independent Results

Due to the paucity of space, we do not evaluate the speaker independent results on every combination of multiple views. We choose those combinations which prove to be the best in the comprehensive experiments conducted. The results for the male and female speakers (Speakers 38 and 39, respectively), are presented in the Table 1. It can also be noted that results for male speaker are better than the female one, we believe this is so since the number of male speakers in the dataset are in a majority.

Table 1: Readings for **best-view combinations** PESQ scores for Speaker Independent models

View Union	Male	Female
0°	1.90	1.76
0°+45°	2.03	1.85
0°+45° + 60°	1.94	1.86
0°+45° + 60° + 90°	1.91	1.82
0°+30°+ 45° + 60° + 90°	1.91	1.83

Demonstration of Reconstructed Audios

Just numeric results cannot do justice to reconstructed *speech* output. Thus the readers are encouraged to view the video playlist at http://bit.ly/2qNqFns showcasing all the reconstructed audios obtained. Please use headphones to be able to listen to the reconstructed speech better. It is worth noting that in the demonstration[1], the audio is in *sync* with the video in addition to the speaker's human voice being intelligible.

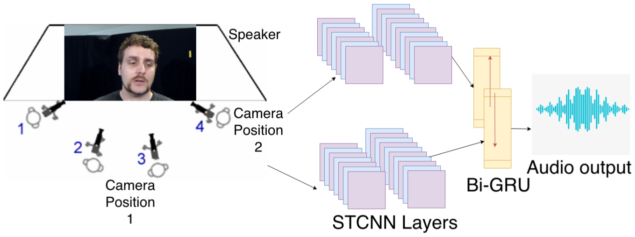

Figure 1: STCNN and BiGRU based architecture used for speech reading and reconstruction

References

Anina, I.; Zhou, Z.; Zhao, G.; and Pietikäinen, M. 2015. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015,*, volume 1, 1–5. IEEE.

Cornu, T. L., and Milner, B. 2015. Reconstructing intelligible audio speech from visual speech features. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Ephrat, A., and Peleg, S. 2017. Vid2speech: speech reconstruction from silent video. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017*, 5095–5099. IEEE.

Fant, G. 2012. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter.

Kumar, Y.; Aggarwal, M.; Nawal, P.; Satoh, S.; Shah, R. R.; and Zimmermann, R. 2018a. Harnessing ai for speech reconstruction using multi-view silent video feed. In *2018 ACM Multimedia Conference on Multimedia Conference*, 1976–1983. ACM.

Kumar, Y.; Jain, R.; Salik, M.; ratn Shah, R.; Zimmermann, R.; and Yin, Y. 2018b. Mylipper: A personalized system for speech reconstruction using multi-view visual feeds. In *2018 IEEE International Symposium on Multimedia (ISM)*, 159–166. IEEE.

Kumar, Y.; Jain, R.; Salik, K. M.; Shah, R. R.; Yin, Y.; and Zimmermann, R. 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In *AAAI.*

Lee, D.; Lee, J.; and Kim, K.-E. 2016. Multi-view automatic lip-reading using neural network. In *Asian Conference on Computer Vision*, 290–302. Springer.

Petridis, S.; Wang, Y.; Li, Z.; and Pantic, M. 2017. End-to-end multi-view lipreading. *arXiv preprint arXiv:1709.00443.*

Zimmermann, M.; Ghazi, M. M.; Ekenel, H. K.; and Thiran, J.-P. 2016. Visual speech recognition using pca networks and lstms in a tandem gmm-hmm system. In *Asian Conference on Computer Vision*, 264–276. Springer.

[1] We play the reconstructed videos with the best 3-view (0°, 45° and 60° combination) three times so that readers can easily understand the audio.