Lipper: Synthesizing Thy Speech Using Multi-View Lipreading

Yaman Kumar
Adobe
ykumar@adobe.com

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

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

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

Yifang Yin
NUS, Singapore
yifang@comp.nus.edu.sg

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

Abstract

Lipreading has a lot of potential applications such as in the domain of surveillance and video conferencing. Despite this, most of the work in building lipreading systems has been limited to classifying silent videos into classes representing text phrases. However, there are multiple problems associated with making lipreading a text-based classification task like its dependence on a particular language and vocabulary mapping. Thus, in this paper we propose a multi-view lipreading to audio system, namely Lipper, which models it as a regression task. The model takes silent videos as input and produces speech as the output. With multi-view silent videos, we observe an improvement over single-view speech reconstruction results. We show this by presenting an exhaustive set of experiments for speaker-dependent, out-of-vocabulary and speaker-independent settings. Further, we compare the delay values of Lipper with other speechreading systems in order to show the real-time nature of audio produced. We also perform a user study for the audios produced in order to understand the level of comprehensibility of audios produced using Lipper.

Introduction

Human speech is bimodal in nature, with the two modalities coming from the auditory and optical senses. An example in this regard comes from experiments by McGurk and MacDonald in their appropriately titled paper, "Hearing lips and seeing voices" (McGurk and MacDonald 1976), where they show that subjects when shown mouth images speaking */ga/* but with the sound of */ba/*, perceived it as */da/*. However, the recent work in lipreading domain has decoupled the visual signals from the auditory ones. Instead, most of the lipreading projects treat this problem as a classification task where they consider speech videos from a restricted vocabulary of a particular language. Then, models are trained to classify those videos into a fixed number of classes made up of that limited vocabulary. However, scaling that approach to multiple languages and a complete vocabulary is a difficult task. In addition, humans do not always speak valid statements. Meaningless, gibberish or non-language and vocabulary-conformant speech (for instance, a human speaker making animal sounds) cannot be modeled using a restrictive approach like a classification model. Thus, with this in mind, we present a model, namely Lipper, which given a silent video consisting of lip-movements *reconstructs* the speech of the speaker. It does this by modeling lipreading as a regression rather than a classification task.

Distinction between Speech-reconstruction, Speech-reading, and Speech-recognition systems

Lipper is a *speech reconstruction* model. People can easily get confused between speech reconstruction, reading and recognition systems. The only commonality amongst all of these systems, pertaining to our work, is that lipreading has been shown to be effective in all the three tasks.

Taking the case of speech recognition systems, they perform the task of identifying the speaker of a speech, and therefore, are not related to this work. However, lipreading based speech-reading and reconstruction systems do share some common features. Thus, with this in mind, in explaining this work, we focus on speech-reading and speech-reconstruction systems *only*.

While on one hand, speech-reading systems involve identifying **what** a person says (using text as the identification metric), on the other hand, the objective of speech-reconstruction systems is to **generate** the *speech* of a person (using audio as the output generated). However, generation of the speech using reconstruction systems may *not* involve identifying *what* a person says. For instance, speech can be generated even for illegible tokens (such as any random permutation of characters), but due to the vocabulary dependence of speech-reading models, it becomes difficult to identify these illegible tokens. Consider another example where this might be useful, multilingual people often speak in code-switched languages (for e.g., in India, a code-switched version of *"I was in London the last month"*, can be, "*Pichle mahine*, I was in London"). For identifying that kind of speech, a speechreading system has to consider all the possible combinations of words in both the languages and then identify them. This quickly becomes un-scalable as the number of languages and the size of each vocabulary increase.

The reason for the language and vocabulary independence of speech reconstruction models like Lipper is that for the generation of a sound, lip, nose, throat and other *movements* are required, but not the vocabulary (or language) per se.

Thus, following this reasoning, one may directly translate and map lip movements to speech without referencing word or sentence mappings. This is what Lipper does. Additionally, unlike other models which have to wait for the complete sentence to get over before they can start speechreading, Lipper can begin to produce audio as soon as it detects some lip-movements, thus becoming a near real-time system. Since it is not contingent on any particular language or vocabulary mapping, thus it is a *language and vocabulary independent* model.

In this paper, we present an exhaustive set of experiments on Lipper for speech-reconstruction. Lipper can take multiple views into account and then lipread them to produce sound. First, in Section Speaker-Dependent Results, we present thorough experiments for exploring the quality of speech-reconstruction on all possible combinations of all the views. Second, in Section Speaker-Independent Results, using the best-view combination, we explore the results of Lipper for speaker-independent settings. Third, in Section Speaker Dependent OOV Results, we note the results of Lipper for out-of-vocabulary (OOV) phrases. Fourth, in Section Comparison of Delays, we compare the delays for speech-reconstruction and speech-reading systems thus demonstrating the difference of time taken between them. Fifth, in Section User-Study, we do a user study for the speech-comprehensibility of the sound generated. Sixth, in Section Demonstration of Reconstructed Audios, we provide links to some of the videos which show speaker-dependent and speaker-independent results produced using Lipper. Seventh, though the audios produced by Lipper might be noisy in some cases, but they do capture and contain the content of the speech of a speaker. Thus, in order to show that, we present text-classification results in Section Text-Prediction Model on the encoded audios produced. In the end, before concluding the paper, we present the future research directions in lipreading domain.

Related Work

Despite much research in lipreading domain, it is still seen as a classification task in which, given some silent videos, a model has to classify those videos into a limited and fixed size of lexicon (Lucey and Potamianos 2006; Ngiam et al. 2011; Lee, Lee, and Kim 2016; Zimmermann et al. 2016; Assael et al. 2016; Chung et al. 2016; Petridis et al. 2017; Chung and Zisserman 2017; Shah and Zimmermann 2017). There have also been a few works on speech-reconstruction as well (Cornu and Milner 2015; Ephrat and Peleg 2017; Kumar et al. 2018a; 2018b). However, the problem of view and pose-variation has been dealt by a very few lipreading systems (Zhou et al. 2014).

Most of the authors have worked on frontal-view lipreading only. Lipreading on just frontal view is a major problem since a speaker cannot be expected to always face the camera while speaking. In the speechreading domain, there have been a few works which have worked on some views other than the frontal view (Lucey and Potamianos 2006; Kumar, Chen, and Stern 2007; Lan, Theobald, and Harvey 2012; Lee, Lee, and Kim 2016; Saitoh et al. 2016) but dealing with pose-variation is still a challenge. In addition, the problem is compounded by the absence of multi-view datasets. A very limited number of datasets exist for fostering research in multi-view lipreading. One of such datasets is Oulu-VS2 (Anina et al. 2015) which provides five different views of speakers shot concurrently. On this dataset, combination of multiple poses was tried for speechreading by (Petridis et al. 2017) and for speech-reconstruction by (Kumar et al. 2018a; 2018b). Given visual feeds from multiple cameras, the authors showed that combining multiple views would result in better accuracy in speechreading (Lucey and Potamianos 2006; Zimmermann et al. 2016; Lee, Lee, and Kim 2016; Petridis et al. 2017) and better speech quality (Kumar et al. 2018a; 2018b) for speech-reconstruction.

Another task in lipreading domain is dealing with pose-variation. Usually, different models are made for dealing with different poses (Lucey and Potamianos 2006; Kumar, Chen, and Stern 2007; Lan, Theobald, and Harvey 2012; Lee, Lee, and Kim 2016; Saitoh et al. 2016). The other approach for dealing with pose-variation is to extract pose-invariant features and then use them in speech-reading (Lucey and Potamianos 2006; Lucey, Sridharan, and Dean 2008; Lan, Theobald, and Harvey 2012; Estellers and Thiran 2011). However, the chief limitation of these systems is their low accuracies which prevents their usage. There have been very few works on speech reconstruction using single view visual feed (Ephrat and Peleg 2017; Cornu and Milner 2015). However, as noted earlier, neither were the systems tested on pose variation, nor speaker-independent settings or on multiple views. We show Lipper's performance on pose-invariant multi-view speech-reconstruction.

Lipper: Design and Development

In this section, we describe the architecture of Lipper (as shown in Figure 1). Primarily, Lipper is composed of a view classifier followed by a STCNN+BiGRU (a combination of Spatio-Temporal Convolutional Neural Network and Bidirectional Gated Recurrent Units) network (as shown in Figure 2). As shown in the diagram, the view-classifier takes input from multiple cameras, and then, maps the speaker view (can be from frontal view, i.e., 0° to profile view, i.e., 90°) to the nearest pose from the pose-set $\{0^\circ, 30^\circ, 45^\circ, 60^\circ$ and $90^\circ\}$. Once this has been mapped, based on the view mapping provided by the classifier, the decision logic decides on two issues:

1. Which view combinations to utilize. This is based on the experiments shown in Section Speaker-Dependent Results. The decision logic may or may not decide to utilize all the available data.
2. Based on the above decision, it chooses the appropriate speech-reconstruction model which takes the multi-view visual input feed and generates speech.

The system generates two types of outputs: audio and associated text. The audio is generated after decoding the encoded audio produced using the neural network. The text is generated by taking encoded audio as input in another neural network which performs classification of the encoded audio into predefined categories.

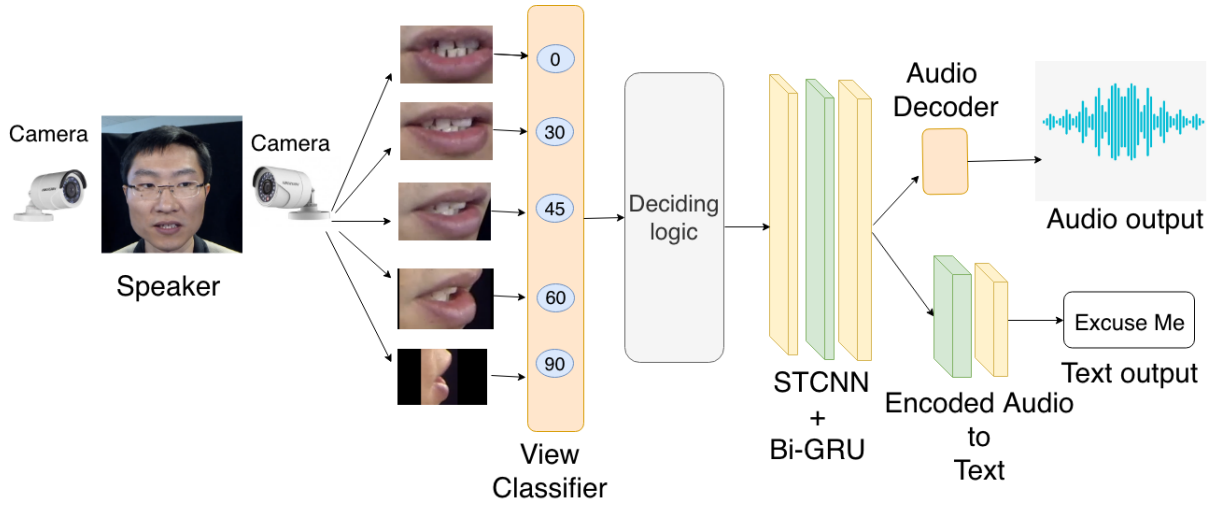

Figure 1: End-to-end diagram for Lipper

True ° / Pred. °	0°	30°	45°	60°	90°
0°	1546	13	1	0	0
30°	76	1399	78	7	0
45°	3	19	1500	38	0
60°	0	0	20	1538	2
90°	0	0	0	2	1558

Table 1: Confusion matrix for the view classifier

Classifier Model

The classifier model uses transfer learning to classify lip-poses. It consists of a VGG-16 model (Simonyan and Zisserman 2014) pretrained on ImageNet images followed by one dense layer with 1024 units and then by one softmax layer with five units. The VGG-16 model helps in extracting the visual features from the lip region images. Since this is a multi-class classification problem, we use the cross-entropy loss to train the system. A visual representation of the classification model is given in Figure 3. For training the model, we used lip-region images of size 224x224. While training, we use the batch size as 100 and then we train the system for 30 epochs with Adam optimization. The confusion matrix of the classifier is given in the Table 1. For training and testing, we used a uniform class distribution with equal number of samples from each of the classes. The overall accuracy as calculated from Table 1 is 96.7%.

Speech-Reconstruction Model

STCNN+BiGRU network helps Lipper to deal with video based data. STCNN layers extract the visual features while BiGRU layers help it to take care of temporal nature of the data. For the audio features, we use Linear Predictive Coding (LPC) (Fant 2012) for representing audio. LPC order was found out to be optimal at 24. Line Spectrum Pairs (LSPs) (Itakura 1975) can represent LPC coefficients in a quantization robust manner. The network takes input images of lip-region of size 128x128 and produces the output as LPC+LSP encoded audio. The network consists of 7 layers of STCNN (of size 32, 32, 64, 64, 128 and 128 respectively) followed by 2 layers of BiGRU (of size 64 and 32 respectively). This is finally followed by the output layer of size 50 which produces the encoded audio.

We use 60 epochs for training and 20 epochs for fine-tuning the network. We first sample audio at a sampling rate of 20,000 and then encode it using LPC order of 24. This is then fed to the network along with the image sequence in timesteps of 5 during the training time.

For different experiments Lipper's training happens in different formats:

1. For speaker dependent experiments, Lipper was trained on all possible combination of views (5 views available in total, thus $\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5}$ possible combinations) for each individual speaker such that out of three phrases for a given speaker in the database (Anina et al. 2015), it was trained on two and one was kept for testing the system. The results corresponding to that are given in Section Speaker-Dependent Results.

2. For speaker independent settings, the best view combination as obtained from the experiments above was taken. The model then was trained on all but two speakers for all their phrases. Then, it was tested on the two speakers left out for the speech generated. The results corresponding to that are given in Section Speaker-Independent Results.

3. For out-of-vocabulary settings, Lipper was trained individually for all the speakers. The training strategy followed was such that it was trained ten times and in each iteration, one of the ten phrases was left out of the training data, and was included in the test data. Thus, in each iteration, Lipper was tested on a phrase which it had never seen. The results corresponding to that are given in Section Speaker-Dependent OOV Results.

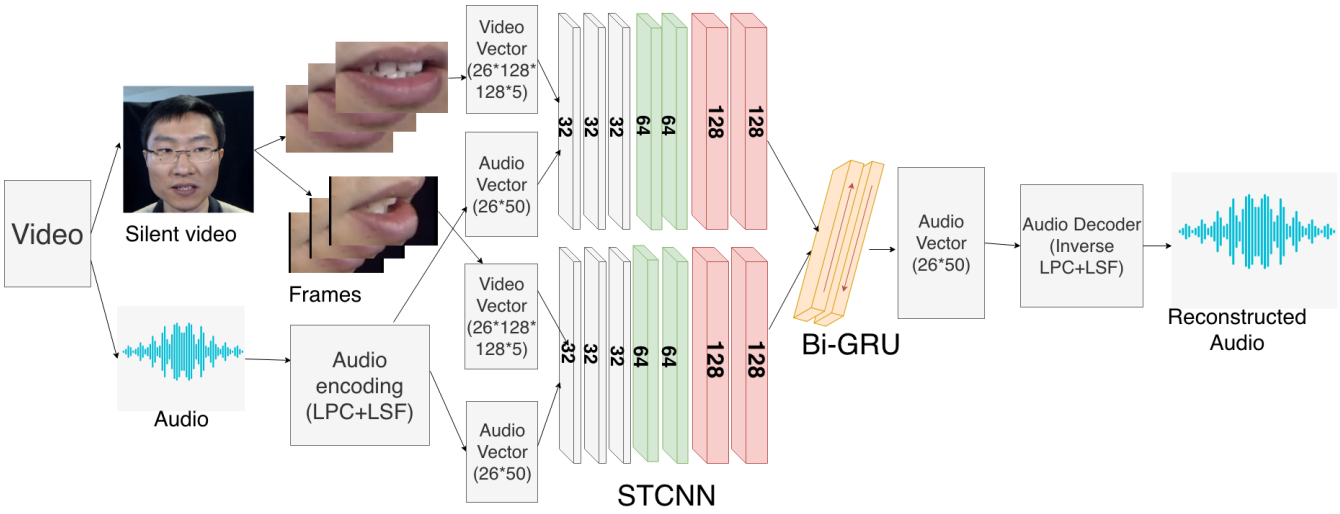

Figure 2: STCNN and BiGRU based architecture used for speech reading and reconstruction

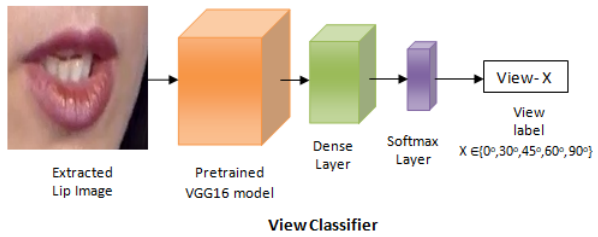

Figure 3: View-classifier model for Lipper. It classifies lip region images into five categories from the set {0°, 30°, 45°, 60° and 90°}

Text-Prediction Model

The text-prediction model takes encoded audio of the reconstructed speech as input and classifies those encoded audios into text classes. For collecting the input for this model, we used a pre-trained speech-reconstruction STCNN+Bi-GRU network and then obtained the output of *all* the silent videos present in the database.

The network has four fully connected layers (with sizes as 1000, 500, 100 and 10) with dropout (0.5) after the layers with sizes 500 and 100. The loss function used was cross-entropy loss and the optimizer as Adam. The network was trained with batch size as ten and number of epochs as twenty.

For making a text predicting model, we tried two different train-test data configurations:

1. In the first configuration, we randomly divided all the encoded audios of all the speakers into train, test and validation data with the ratio as (70, 10 and 20) respectively.
2. In the second configuration, we took *all* the encoded audios of 70% of the speakers as the training data, and divided the rest 30% speakers' audios into 10% and 20% to be used as validation data and test data respectively.

	Config-1	Config-2	(Petridis et al. 2017)
Accuracy	97.0	78.5	95.6

Table 2: Accuracy of the two configurations of the text-prediction model trained compared with the best results as reported in (Petridis et al. 2017)

The results for both these configuration compared with those of (Petridis et al. 2017) (which is the state-of-the-art speachreading model on OuluVS2) are presented in Table 2. We present the best accuracy reported in the paper by (Petridis et al. 2017) for comparison. As shown in the table, data configuration-1 performs much better than even the best-performing model of (Petridis et al. 2017). We believe this is due to the train-test configuration of Lipper itself. Randomly classifying the video into one of the ten classes present in the database would have led to 10% accuracy. Thus, even considering the second configuration, the accuracy of the text-prediction model is non-trivial. This implies that the audios produced by Lipper, although might be noisy in some cases, capture the content of the speech of a speaker.

Evaluation

Database

For training and testing Lipper, we use *all the speakers* of OuluVS2 database (Anina et al. 2015) for speech-reconstruction purposes. OuluVS2 is a multi-view audio-visual dataset with 53 speakers of various ethnicities like European, Indian, Chinese and American. These 53 speakers speak 10 phrases with five cameras recording them simultaneously from five different angles. The angles considered are {0°, 30°, 45°, 60° and 90°}. The speakers speak at different pace and also stop in between while speaking. A listing of all the phrases is given in Table 3. Thus, this dataset serves well for the task at hand. This database has been used

in other similar studies as well (Ong and Bowden 2011; Zhou, Zhao, and Pietikäinen 2011; Pei, Kim, and Zha 2013; Rekik, Ben-Hamadou, and Mahdi 2014; 2015; Petridis et al. 2017; Kumar et al. 2018a; 2018b).

Table 3: List of phrases uttered by speakers in OuluVS2

S.No.	Phrases	S.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

Evaluation Metric

We choose Perceptual Evaluation of Speech Quality (PESQ) (Rix et al. 2001) as the evaluation metric to judge the quality of the sound. This metric has been used by other speech-reconstruction works as well (Ephrat and Peleg 2017; Ephrat, Halperin, and Peleg 2017; Kumar et al. 2018a; 2018b). PESQ is a ITU-T recommended standard for evaluating speech quality of 3.2 kHz codecs (Recommendation 2001). For comparing two audios, PESQ first level aligns them, then after passing them through filters time aligns them. It then passes the audios through auditory transform and finally extracts two distortion parameters which denote the difference between the transform of the two signals. Finally, these signals are aggregated in frequency and time, and are mapped to a MOS (mean opinion score). The range of PESQ varies from -0.5 to 4.5, where speech quality increases with increasing score.

Results

This section presents the results obtained for speaker-dependent, speaker-independent, out-of-vocabulary settings for Lipper. Additionally, it also reports the results for delay measurements on Lipper.

Speaker-Dependent Results

Speaker Dependent Single-View Results The results for Lipper when trained and tested on single-view visual feeds are shown in Table 4. We have compared our results with (Ephrat and Peleg 2017) which train a similar speech-reconstruction network and tested it on single-view visual feed. As can be seen from the table, our results are better for all the views than (Ephrat and Peleg 2017) but the general trend of PESQ scores is similar in both the systems. In both the models, frontal view outperforms all other views and obtains a PESQ score of 2.002 and 1.72 respectively.

Speaker Dependent Multi-View Results Table 5 presents the PESQ scores for all the possible two-view combinations. As can be seen, the combination of 0° and 45° outperforms all the other possible combinations. Closely following it is the combination of 0° and 30°. It should also be noted that the PESQ scores, in general, for all the possible views have been benefited after a combination with some other view. For instance, 30° in combination with 0°, experiences a gain in PESQ by over 6%. Thus with respect to placement of two cameras, in regards to obtaining best quality of audios, one should place the cameras at an angle of 45° between them. This, as shown by the table, would produce an audio which would carry the maximum quality.

Table 6 presents the mean PESQ scores for all possible three view combinations. The views combination 0°, 45° and 60° outperforms all the combinations and presents the best results obtained till now. This is a stupendous increase of over 32% for both 45° and 60° when considering their single view PESQ scores only. Moreover, even while considering the combination of 45° and 60°, their association with 0° leads to a non-trivial increase of more than 18%. Although, it can be noted that not all possible 3 view combinations experience a gain over their 2 view counterparts. This might be because of less training data available due to which a larger network could not be trained appropriately.

Results for all possible 4-view combinations are shown in Table 7. In most cases, there is not a major increase in PESQ scores from the three-view combinations or in some cases, even a decrease in the scores is observed. As has been stated above, the reason for this decline in performance can be due to non-availability of adequate data for training the larger network. However, some view combinations have better scores than their individual view counterparts.

Table 8 shows the result obtained on all-view combinations. As can be seen, although the network at this stage outperforms all of its single view counterparts but does not perform as good as the best possible three view combination of 0°, 45° and 60°. We have not compared 2-views, 3-views, 4-views and 5-views combinations with models by other authors since as mentioned in the Section Related Work, there were no models previously who have worked on combinations of multiple views.

Speaker Dependent OOV Results

One of the major strong points for any speech reconstruction system is their ability to reconstruct speech on phrases which were not present in the training set. For a system which treats lipreading as a classification task, this is not possible since essentially, these systems have to mark any video with the limited classes that they consider. However, human language (or, in general, sounds made by humans) presents a very wide vocabulary and cannot be modeled easily with speechreading models.

In the Table 9, we present the PESQ scores using the overall best model (0°, 45° and 60° combination), for each of the phrases from the Table 3 considered as out-of-vocabulary in different iterations. The OOV PESQ scores though lesser than their speaker-dependent counterparts are not inconsequential.

Speaker Independent Results

This section presents the speaker independent results obtained using Lipper. 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

Table 4: Mean readings for **single-view** PESQ scores for Lipper and (Ephrat and Peleg 2017)

Views	0 degree	30 degree	45 degree	60 degree	90 degree
Lipper	**2.002**	1.750	1.642	1.744	1.804
(Ephrat and Peleg 2017)	**1.72**	1.57	1.48	1.46	1.52

Table 5: Mean readings for **double-view** PESQ scores

View Union	**PESQ**	**View Union**	**PESQ**
0°+30°	2.125	30°+60°	1.842
0°+45°	**2.130**	30°+90°	2.021
0°+60°	1.952	45°+60°	1.960
0°+90°	1.982	45°+90°	1.930
30°+45°	1.991	60°+90°	1.920

Table 6: Mean readings for **triple-view** PESQ scores

View Union	**PESQ**	**View Union**	**PESQ**
0°+30°+45°	1.975	0°+60°+90°	1.987
0°+30°+60°	2.112	30°+45°+60°	1.931
0°+30°+90°	2.005	30°+45°+90°	1.903
0°+45°+60°	**2.315**	30°+60°+90°	1.965
0°+45°+90°	1.814	45°+60°+90°	1.838

Table 7: Mean readings for **quadruple-view** PESQ scores

View Union	**PESQ**
0°+30°+45°+60°	2.11
0°+30°+45°+90°	1.916
0°+45°+60°+90°	**2.147**
0°+30°+60°+90°	2.071
30°+45°+60°+90°	1.948

Table 8: Mean reading for **all-view** PESQ scores

Combination of views	**Mean PESQ scores**
0°+30°+45°+60°+90°	**2.086**

Table 9: Mean readings for single view PESQ on the OuluVS2 database for out of vocabulary (OOV) phrases

Phrases	**Mean PESQ scores**
Excuse Me	1.79
Goodbye	1.66
Hello	1.82
How are you	1.84
Nice to meet you	1.57
See you	1.64
I am sorry	1.68
Thank you	1.55
Have a good time	1.46
You are welcome	1.60

Table 10: Readings on the best-view combination PESQ scores for speaker-independent settings

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

speaker dependent settings. The results for the male and female speakers (Speakers 38 and 39, respectively), are presented in the Table 10. It can be observed that the results for speaker dependent models are significantly better than the speaker independent one. We believe this is so since Lipper does not only depend on lip movements of individual speakers but also their voices. The lip-movements of the speakers although are different, but carry the commonality of movement while speaking the same words. However, since Lipper depends not only on lip-movements but also voices of the speakers and since the voice of each speaker is different, thus the model does not perform well on the PESQ score evaluation for unknown speakers. Thus, in speaker independent settings, Lipper is not able to learn the person-specific voice features which are crucial for getting high scores using PESQ as the evaluation metric. This can explain the noticeable difference between the speaker dependent and independent results obtained using Lipper. Additionally, 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 thus forming a bias for Lipper in favour of male voice generation.

Comparison of Delays

One of the major advantages of Lipper is it being a near real-time system. In this section, we compare the end-to-end delay in getting speech from Lipper and a similar work of that of speechreading by (Petridis et al. 2017) thus confirming the validity of Lipper being a small delay speech-reconstruction system. The comparison of delay values is reported in Table 11. It is to be noted that the delay values for speechreading work of (Petridis et al. 2017) depend on the phrase spoken, longer the phrase higher is the delay.

User-Study

Although PESQ is a standard numeric measure which can give an idea of the virtue of a speech and can give a reference metric for comparison but from our experiments, it was observed that the measurements produced by PESQ were not always perfect. Even for some noisy audios, the PESQ scores were high and the vice-versa was also found out to

Table 11: Mean readings for Delay values (in secs) on Lipper and (Petridis et al. 2017)

Video	Lipper(s)	Petridis et al.(s)
Excuse Me	0.169	1.26
Goodbye	0.169	0.94
Hello	0.169	0.98
How are you	0.169	1.24
Nice to meet you	0.169	1.4
See you	0.169	1.34
I am sorry	0.169	1.44
Thank you	0.169	1.09
Have a good time	0.169	1.61
You are welcome	0.169	1.95

Table 12: User studies for the reconstructed audios

Study	Accuracy	Variance
Audio-only	80.25	2.72
Audio-visual	81.25	1.97

be true. Due to these observations, a user study was done to understand the intelligibilty of the audios. Two types of user studies were done for doing this analysis:

1. In the first study, we gave reconstructed speech of 25 speakers to 10 different annotators who were given 4 options consisting of 1 true and 3 random options (amongst the 10 classes). Each annotator was asked to listen to the audios produced by the model as many times as he likes and then choose the best option amongst the 4 classes given.
2. In the second study, in accordance with a real video-conference environment, we showed the speaker's videos to the annotators with the reconstructed audio playing along with the video. Then, the annotators were asked to choose among 10 phrase classes for the audio-video sequence that they just listened to.

The results for both the user studies are given in Table 12. In the audio-only study, a dice throw could get 25% of the annotations right but the annotation accuracy turned out to be 80.25 with annotator accuracy variance value as 2.72%. In the audio-visual study, although a dice-throw would have led 10% annotations correct, Lipper achieved an annotation accuracy of 81.25 with inter-annotator accuracy variance as 1.97%.

Demonstration of Reconstructed Audios

Just numeric results cannot do justice to reconstructed *speech* output. Thus, we have made a video listing consisting of all the reconstructed speech outputs as part of a *Youtube* channel. The readers are encouraged to view the video playlist at https://www.youtube.com/playlist?list=PL9rvax0EIUA4LNaXSeVX5Kt6gu2IBBnsg. This playlist contains videos consisting of speech reconstructed using speaker dependent model, speaker independent model, OOV phrases and videos of some non-dataset speakers who speak multiple languages (Hindi, English and Chinese). The non-dataset speakers and the languages they speak show the language and vocabulary independence of the model. For them, we train speaker-dependent models, and first carefully get their lip-region videos and then reconstruct the speech using the model generated. 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 complete *sync* with the video. In addition, the speaker's voice is comprehensible and can be understood.

Conclusion and Future Work

Future Research Directions

As explained in this paper, not much research has happened in speech reconstruction domain. Thus, there are a lot of areas where speech-reconstruction system can be improved.

As shown in the Section Demonstration of Reconstructed Audios, the audios are robotic in nature. One of the main reasons for this is that voice is generated not just using mouth, but also using nose, throat and tongue. Since Lipper takes only lip-region into account, thus the voice generated cannot have emotion, prosody or modulation. Therefore, speech reconstruction systems have to work to make the audios more real-life.

Currently, the system only works in controlled environment where speakers are not moving much and are looking into the camera at a stable angle. However, in the real world scenario, this cannot be the case. The speakers will turn and twist and their poses would vary dramatically, thus, going forward speech-reconstruction has to take that into account.

In this paper, although speaker-independent settings were explored, but as was seen, the system does not work very well on them. This is a major problem for the deployment of speech-reconstruction systems in their use-case scenarios.

Conclusion

In this paper, the authors proposed a real-time, language and vocabulary independent, multi-view accounting and speaker-independent speech reconstruction system, namely Lipper, which utilizes multi-view visual feeds to generate the speech of a speaker. Lipper extracts features directly from the pixels of the multi-view videos. It then learns those spatial features jointly along with temporal features to finally reconstruct speech of a user. The proposed system showed significant intelligibility for the audios constructed. The best combination of views was found to be the combination of 0°, 45° and 60°. This combination produced a significant gain over other possible views and their combinations. We also showed the experiments of out-of-vocabulary phrases for speech reconstruction and the delay between getting speech for speachreading and speech-reconstruction systems.

[1]We obtain the reconstructed videos with the best 3-view (0°, 45° and 60° combination). The audios are played three times so that readers can easily understand them.

Acknowledgement

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266 and by the National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, People's Republic of China, 215123.

MIDAS lab gratefully acknowledges the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

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.

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and de Freitas, N. 2016. Lipnet: end-to-end sentence-level lipreading.

Chung, J. S., and Zisserman, A. 2017. Lip reading in profile. BMVC.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2016. Lip reading sentences in the wild. *arXiv preprint arXiv:1611.05358* 2.

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.

Ephrat, A.; Halperin, T.; and Peleg, S. 2017. Improved speech reconstruction from silent video. In *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media*.

Estellers, V., and Thiran, J.-P. 2011. Multipose audio-visual speech recognition. In *Signal Processing Conference, 2011 19th European*, 1065–1069. IEEE.

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

Itakura, F. 1975. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America* 57(S1):S35–S35.

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, K.; Chen, T.; and Stern, R. M. 2007. Profile view lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, volume 4, IV–429. IEEE.

Lan, Y.; Theobald, B.-J.; and Harvey, R. 2012. View independent computer lip-reading. In *IEEE International Conference on Multimedia and Expo (ICME), 2012*, 432–437. IEEE.

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

Lucey, P., and Potamianos, G. 2006. Lipreading using profile versus frontal views. In *IEEE Workshop on Multimedia Signal Processing*, 24–28. IEEE.

Lucey, P. J.; Sridharan, S.; and Dean, D. B. 2008. Continuous pose-invariant lipreading.

McGurk, H., and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(5588):746.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Ong, E.-J., and Bowden, R. 2011. Learning sequential patterns for lipreading. In *BMVC 2011-Proceedings of the British Machine Vision Conference 2011*. The British Machine Vision Association and Society for Pattern Recognition.

Pei, Y.; Kim, T.-K.; and Zha, H. 2013. Unsupervised random forest manifold alignment for lipreading. In *IEEE International Conference on Computer Vision*, 129–136. IEEE.

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

Recommendation, I.-T. 2001. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Rekik, A.; Ben-Hamadou, A.; and Mahdi, W. 2014. A new visual speech recognition approach for rgb-d cameras. In *International Conference Image Analysis and Recognition*, 21–28. Springer.

Rekik, A.; Ben-Hamadou, A.; and Mahdi, W. 2015. Unified system for visual speech recognition and speaker identification. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, 381–390. Springer.

Rix, A. W.; Beerends, J. G.; Hollier, M. P.; and Hekstra, A. P. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 2, 749–752. IEEE.

Saitoh, T.; Zhou, Z.; Zhao, G.; and Pietikäinen, M. 2016. Concatenated frame image based cnn for visual speech recognition. In *Asian Conference on Computer Vision*, 277–289. Springer.

Shah, R., and Zimmermann, R. 2017. *Multimodal analysis of user-generated multimedia content*. Springer.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhou, Z.; Zhao, G.; Hong, X.; and Pietikäinen, M. 2014. A review of recent advances in visual speech decoding. *Image and vision computing* 32(9):590–605.

Zhou, Z.; Zhao, G.; and Pietikäinen, M. 2011. Towards a practical lipreading system. In *IEEE International Conference on Computer Vision*, 137–144. IEEE.

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.