INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Hush-Hush Speak: Speech Reconstruction Using Silent Videos

Shashwat Uttam[1*], *Yaman Kumar*[2*] , *Dhruva Sahrawat*[3*], *Mansi Aggarwal*[4], *Rajiv Ratn Shah*[3], *Debanjan Mahata*[5], *Amanda Stent*[5]

[1]MIDAS@IIITD, NSUT, India
[2]Adobe, India
[3]MIDAS, IIIT-Delhi, India
[4]MIDAS@IIITD, DTU, India
[5]Bloomberg, USA

shashwatu.co@nsit.net.in, ykumar@adobe.com, dhruva15026@iiitd.ac.in, r18522mansi@dpsrkp.net, rajivratn@iiitd.ac.in, dmahata@bloomberg.net, astent@bloomberg.net

Abstract

Speech Reconstruction is the task of recreation of speech using silent videos as input. In the literature, it is also referred to as *lipreading*. In this paper, we design an encoder-decoder architecture which takes silent videos as input and outputs an audio spectrogram of the reconstructed speech. The model, despite being a speaker-independent model, achieves comparable results on speech reconstruction to the current state-of-the-art *speaker-dependent* model. We also perform user studies to infer speech intelligibility. Additionally, we test the usability of the trained model using bilingual speech.

Index Terms: speech reconstruction, human-computer interaction, speech recognition, multi-view

1. Introduction

Research on automatic speech recognition has mainly treated the task as one of *classification* - choose a word from a vocabulary that best matches the acoustic signal. However, this limits the capability of the speech recognizer to the given vocabulary. If instead, we treat the task as *reconstruction* [1], we can model a much greater variety of speech, including out of vocabulary or out of language speech. Unfortunately, most current speech reconstruction approaches are speaker-dependent [1, 2, 3, 4, 5], meaning that a separate model is needed for each speaker. Speech reconstruction approaches that are speaker-independent are pose-dependent, thus limiting their usage to a specific speaker view [6, 7].

By contrast, the solution we present in this paper is both speaker-independent and multi-view. Additionally, it does not use any manually annotated labels of any kind; instead, in our approach the system directly learns from '*natural supervision*' where the target prediction is not some human annotation but a natural signal [2, 8], thus requiring no human involvement.

The system is composed of two *main* parts: a deciding logic and multiple speech reconstruction networks. The multiple reconstruction networks differ from one another on the basis of the different views and number of videos they can cater to. The deciding logic, based on the received input videos, decides on a particular speech reconstruction network which will work best for the received videos.

The contributions of this paper are:

1. We present a view and identity invariant speech reconstruction system.
2. The system automatically chooses the best view combination and speech reconstruction model to employ based on the input videos and their corresponding views.
3. The system is independent of any language or vocabulary.
4. Despite being a speaker-independent system, it performs on par with state of the art speaker-dependent systems.

2. Experiments

2.1. Dataset

We use the Oulu VS2 database [9]. It is a multi-view audio-visual dataset consisting of videos of 53 speakers of various ethnicities. Videos in five different poses are given and the pose ranges from 0° to 90°. The talking speed and head movements have not been controlled in the dataset, thus making it close to what one would expect in a real world scenario. Oulu is composed of three sub-datasets - Oulu sentences containing TIMIT sentences [10], Oulu phrases consisting of 10 common phrases and Oulu digits containing a random sequence of digits. We use all the three datasets for training and testing.

2.2. Proposed Model

We use four components in the overall speech reconstruction system:

1. Pose Classifier - The pose classifier model uses transfer learning to classify lip-poses. This module is derived from [5]. It consists of a VGG-16 model [11] pretrained on ImageNet images [12] followed by one dense layer with 1024 units and then by one softmax layer with five units. The VGG-16 model helps in extracting visual features from the lip region images.

2. Decision Network - The decision network is responsible for choosing the right encoder for the input video sequence. The design of this network is inspired by [5]. It chooses based on three factors: the number of videos capturing this speaker, the views of the speaker captured in the different videos and an internal logic (given in Table 2) which maps views to the best performing speech reconstruction networks. For instance, let's assume we have 3 videos of a speaker possibly captured in 3 different views. First, using the pose classifier, the deciding logic maps each videos to its closest views. Let's assume the pose

*Equal Contribution

http://dx.doi.org/10.21437/Interspeech.2019-3269

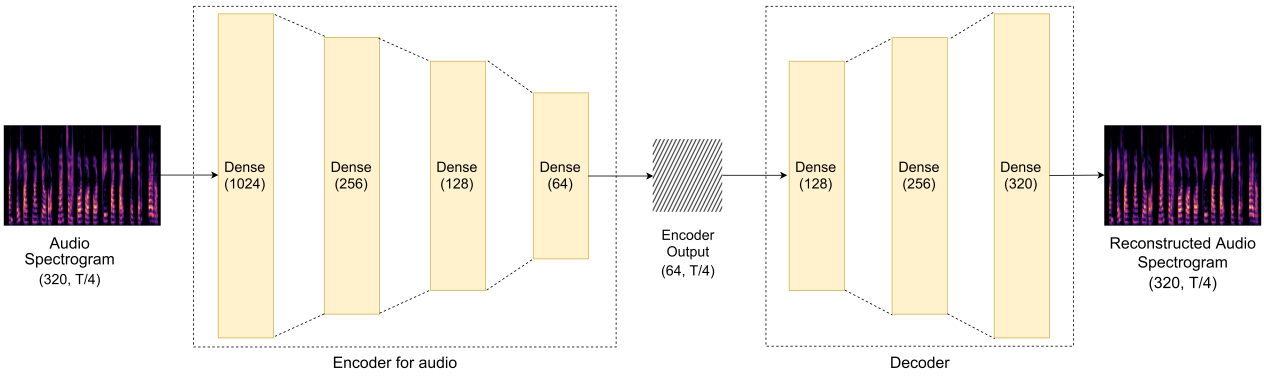

Figure 1: *Audio auto-encoder system showing speech reconstruction. T represents the temporal dimension of the audio spectrogram.*

Table 1: *Results for the audio-autoencoder model*

Bottleneck Size	PESQ	Corr2D
32	2.52	0.95
64	2.76	0.97
128	2.90	0.98

classification network maps them to 0°, 30° and 60°. Based on this information, the decision network makes all the possible 1-view, 2-view and 3-view combinations:- {{0°},{30°},{60°}, {0°, 30°},{60°, 30°} , {0°, 30°, 60°}}. Finally, the decision network has to choose a particular combination to give to the speech reconstruction network. It does this based on precomputed results, presented in Table 2. Thus, in this specific example, the best performing combination among all the possible combinations would be {0°, 30°, 60°}. This would have an expected PESQ[1] score of 1.926 and corr2D[2] of 0.821. This combination of videos is passed on to the third component.

3. Audio Autoencoder - We represent audio using mel-frequency spectrograms. We choose this particular representation since it can represent speech without depending too much on speaker dependent characteristics. Given an input spectrogram, the autoencoder is responsible for learning how to reconstruct that spectrogram. We set 80 frequency bins per mel spectrogram. In the encoder portion of the autoencoder, this gets converted to a bottleneck representation which is then fed to a decoder in order to get the spectrogram representation back. To create a same-sized temporal dimension vector as the video sequence, we stack four audio data points together to create a $(320, T/4)$ shaped vector, where T represents the number of audio data points sampled in the spectrogram. This model is presented in Figure 1. As shown in the figure, the spectrogram representation of size 320 is iteratively downsampled to an encoding of size 64. This is then fed to a decoder module which subsequently up-samples it back to size 320. The network consists of a input layer of size 320 followed by a dense layer containing 1024 neurons. The numbers of neurons in the subsequent layers decrease in simple autoencoder fashion to a bottleneck of size 64. The decoder portion of the network mirrors these layers to get back the output of size 320.

4. Video Encoder - This encoder's input sequence is a pre-processed video sequence consisting of multiple views stacked along depth as input. The videos are divided into slices, consisting of 6 frames each. The encoder network has seven 3D convolution layers (32, 32, 32, 64, 64, 128 and 128) with their sizes in increasing fashion. The last one is followed by an LSTM layer of size 512. Each 3D Convolution layer is followed by a 3D max pooling operation. The output of the LSTM layer is passed through a dense layer and finally to the output layer. The purpose of this network is to capture spatiotemporal features from the input and generate an encoded representation of the audio bottleneck features corresponding to the input video.

Table 2: *Results for speech reconstruction on all view combinations*

View Union	PESQ	Corr2D
0°	1.920	0.813
30°	1.915	0.804
45°	1.801	0.812
60°	1.787	0.810
90°	1.826	0.800
0°+30°	1.941	0.816
0°+45°	1.947	0.815
0°+60°	1.847	0.820
0°+90°	1.751	0.810
0°+30°+45°	1.883	0.818
0°+30°+60°	1.926	0.821
0°+30°+90°	1.931	0.814
0°+45°+60°	1.964	0.816
0°+45°+90°	1.939	0.815
0°+30°+45°+60°	1.931	0.819
0°+45°+60°+90°	1.904	0.812
0°+30°+45°+60°+90°	1.844	0.813

The complete picture of the trained model is shown in the Figure 2. One or more video sequences is taken as input and is given to the deciding logic. The deciding logic then filters the videos based on the conditions presented above and also selects a particular video encoder module best suited for the task. The selected video encoder module encodes the video sequence into a bottleneck representation. This encoding is then fed to the decoder module of audio autoencoder which then outputs a spectrogram representation of the speech present in the video.

[1]Perceptual Evaluation of Speech Quality
[2]2D Correlation Analysis

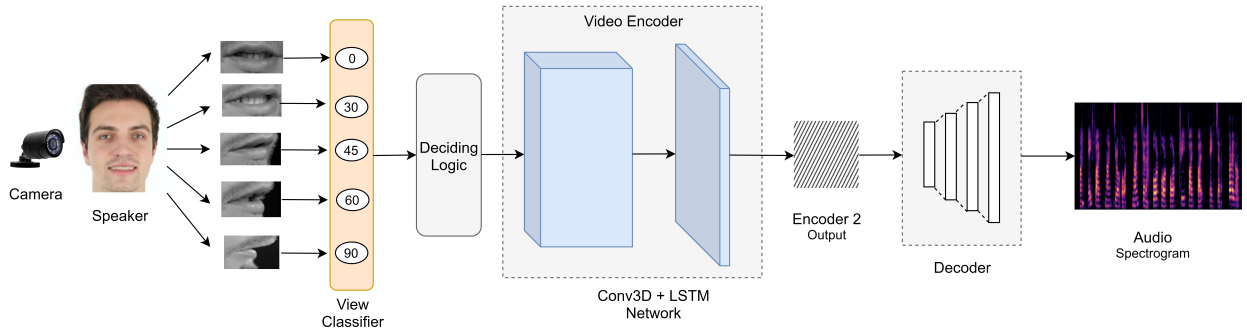

Figure 2: *End-to-end diagram for the proposed model*

2.3. Training of the network

Data Preprocessing: The videos are first converted to grayscale and are then normalized. As mentioned earlier, all videos are divided into slices, consisting of 6 frames each. The frames of the videos are of the shape $(128, 128)$ and contain the lip region of the face. The audio files are re-sampled from 48000 Hz to 22050 Hz, converted to mel spectrograms consisting of 80 bins, and split into slices. The length of the audio slices is selected as 24 such that both video and audio slices are 0.2 seconds long and are perfectly in sync.

The entire network is implemented in Keras [13]. We train the audio autoencoder and the video encoder separately. The audio autoencoder is trained with a batch size of 256. The bottleneck sized feature vectors for the audio samples of size 64 are extracted to be used as the target features for the video sequence encoder. The video sequence encoder is trained with a batch size of 64 for 120 epochs using the Adam optimizer [14] with learning rate initialized to 0.0001 and decreased subsequently as required. We use the difference of correlation and mean square error (MSE) as the loss function. The value of correlation for a perfect reconstruction is 1 and that of MSE is 0. For both models, we use the combination of digits and phrases datasets in Oulu VS2 split into 85:10:5 for training, validation and testing respectively.

3. Results

We evaluate both the audio autoencoder and the combination of video encoder and audio decoder. For evaluating the accuracy of the audio autoencoder in the frequency domain, we measure 2D correlation between reconstructed audio and the input audio. We also use PESQ [15] scores for measuring the quality of the reconstructed speech. The results of the experiments are shown in the Table 1. In the table, we show performance numbers for the audio autoencoder for different bottleneck sizes; we chose size 64 as the bottleneck size because it does not affect the training speed while also performs adequately well.

For the combination of video encoder and audio decoder, we use three evaluation metrics - Corr2D, PESQ and a human evaluation. The reason behind selecting Corr2D and PESQ remains the same as for the audio autoencoder. For human evaluation, we asked five judges (not authors of this paper) to assess the intelligibility of the speech generated. They were asked to assess two things: the words being spoken and the sex of the speaker. This was done in two settings: one where we gave them four options for the word sequence to choose from, and the second, where they had to transcribe themselves. Thus, for each text sequence, we took the audio of four speakers and asked all the annotators to identify the words being spoken and the sex of each speaker.

We perform three kinds of evaluation in order to better understand the model:

Table 3: *Human evaluation results for speech reconstruction on Oulu-phrases and Oulu-digits when either zero or multiple (4) view options were provided, the number of options being indicated within the brackets.*

Phrase	Sex	Speech (4)	Speech (0)
Digit Seq. - 4 0 2 9 1 8 5 9 0 4	1.00	0.90	0.85
Digit Seq. - 2 3 9 0 0 1 6 7 6 4	1.00	1.00	0.95
How are you	0.90	0.75	0.70
Thank you	0.80	0.90	0.65
Excuse Me	1.00	0.95	0.70
Avg over all samples	0.92	0.88	0.69

Table 4: *Human evaluation results for speech reconstruction on two Oulu-sentences when either zero or multiple (4) options were provided, the number of options being indicated within the brackets.*

Phrase	Sex	Speech (4)	Speech (0)
Sentence 1	0.6	0.75	0.65
Sentence 2	0.8	0.7	0.60
Avg over all samples	0.7	0.70	0.65

Table 5: *Results for speech reconstruction for Hindi audios*

Phrase	Lip Movement Conformance	Correct Pronunciation
Dhanyawaad	100%	100%
Haanji	100%	100%
Jungalee	100%	0%
Maaf Kijiye	80%	80%
Swagat	100%	0%

1. Base evaluation of reconstructed speech for data with a similar counterpart available in the training set - For this evaluation, we took some of the data present in the original dataset (but not shown to the model during its training), *i.e.*, a subset of Oulu-phrases and Oulu-digits. Since similar data was already present during training, this data should be the easiest for the model to reconstruct. We used this data for two purposes: one, to derive decision logic for the decision network and two,

Table 6: *Comparison between Lipper [4] and our proposed model using PESQ scores on the view combinations and the settings mentioned in the Lipper paper.*

Model	Lipper[4]		Our model	
View Union	**Male**	**Female**	**Male**	**Female**
0°	1.90	1.76	1.87	1.31
0°+45°	2.03	1.85	1.80	1.33
0°+45° + 60°	1.94	1.86	1.78	1.36

to get a general idea about the accuracy of the network. These results are presented in the Table 2. Subjective evaluation was done by the human evaluators and the corresponding results for some sample text sequences are given in Table 3. As shown in the table, for the common phrase "*How are you*", when given four options, the annotators were able to correctly recognize the phrase 75% of the time and the sex 90% of the time.

2. Evaluation of reconstructed speech for English data not present in the training set - A key benefits of speech reconstruction systems is that they are supposed to be vocabulary independent. With this in mind, we tested our model with the data Oulu-sentences. Unlike the data present in the training set, these text sequences are longer in nature and are derived from TIMIT sentences [10]. The vocabulary size of these sentences is 1073 words whereas the total number of words occurring in the training set is just 30 words. The results for this experiment are given in Table 4. In addition to qualitative assessment, we also performed a quantitative evaluation and got a PESQ score of 1.623 and Corr2d score of 0.816 on this data. Both the qualitative and quantitative assessment point out to the fact of this system being vocabulary independent.

3. Evaluation of reconstructed speech for Hindi data not present in the training set - Another key advantage of speech reconstruction systems is that they are language independent. This is based on the assumption that visemes are cross-lingual and map to similar phonemes [16]. Thus, a model trained on English should be able to translate its intelligence to other languages. We recorded some speakers speaking the Hindi language and gave their silent videos to the speech reconstruction system. We asked the human evaluators to assess two things: one, whether the sound reconstructed is in line with what is expected after seeing the video and two, given the Hindi phrase, does the reconstructed audio correspond to the phrase. The results pertaining to these experiments are in Table 5. We notice that even when these videos are not given to the system at training time and the ones that are given do not contain any language other than English, the system is able to associate correct phonemes with most of the visemes. As marked by the annotators (column 2), the sounds are in line with the video for *all except one* text sequences. Also (column 3), the system is able to correctly map all the Hindi-based phonemes with the corresponding visemes for three out of five text sequences. Thus, for example, according to all the annotators, the phrases "*Dhanyawaad*" and "*Haanji*" are decipherable and correctly mapped to how the phrases are pronounced by a Hindi speaker.

It is worth noting that our system performs competitively compared to the *speaker-dependent* model presented in [1]. For the view combination 30° our system's PESQ score is better than the MyLipper [1] model. In all other cases, our system does not lose more than 15% in PESQ scores, even though our system has the advantage of being generic in nature and not depending on a particular speaker.

We also compared our model against the Lipper [4] model in the speaker independent setting mentioned in their paper. The results are given in Table 6. As shown in the table, Lipper works much better than our system. We believe this is due to two reasons: one, training-testing configuration of Lipper is very different from our paper and two, Lipper is a speaker-dependent system. Lipper's training data includes 2 videos of all the speakers. Thus, during the training time itself, it has seen all the speakers. While our system sees no test-speaker at the training time.

3.1. Demonstration of Reconstructed Speech

We include several examples of reconstructed speech along with what was originally spoken in the video in the auxiliary file[3]. We have also included reconstructed Hindi speech data. In addition, the auxiliary file also contains the judgments done by the human evaluators for the various data points. The reader can evaluate our system by listening to the quality of the reconstructed audio files.

4. Conclusions

In this paper, we presented a speaker and pose independent speech reconstruction model which is also independent of any particular language or vocabulary mapping. We present speech quality, correlation and intelligibility scores on the speech reconstructed using the model. We also perform human evaluation of the speech in order to understand how humans assess the speech reconstructed using the model. We do this for the videos present in the dataset and also for some videos of speakers speaking the Hindi language.

5. Acknowledgement

This research was supported in part by SERB, Department of Science and Technology, Government of India under grant number SRP 124.

6. References

[1] Y. Kumar, R. Jain, M. Salik, R. ratn Shah, R. Zimmermann, and Y. Yin, "Mylipper: A personalized system for speech reconstruction using multi-view visual feeds," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 2018.

[2] Y. Kumar, M. Aggarwal, P. Nawal, S. Satoh, R. R. Shah, and R. Zimmermann, "Harnessing AI for speech reconstruction using multi-view silent video feed," in *Proceedings of the ACM International Conference on Multimedia*, 2018.

[3] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[4] K. M. Salik, Y. Kumar, R. Jain, S. Aggarwal, R. R. Shah, and R. Zimmermann, "Lipper: Speaker independent speech synthesis using multi-view lipreading," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[5] Y. Kumar, K. M. Salik, R. Jain, R. R. Shah, R. Zimmermann, and Y. Yin, "Lipper: Synthesizing thy speech using multi-view lipreading," in *Thirty-third AAAI conference on artificial intelligence*, 2019.

[6] A. Ephrat and S. Peleg, "Vid2Speech: Speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[3]The auxillary file along with the code is available at `https://github.com/midas-research/hush-hush-speak`

[7] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2AudSpec: Speech reconstruction from silent lip movements video," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[8] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology, Tech. Rep., 1993, NTIS Order No PB91-505065.

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

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[13] F. Chollet *et al.*, "Keras," 2015.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "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*, vol. 2. IEEE, 2001, pp. 749–752.

[16] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam, and A. Verma, "Translingual visual speech synthesis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.