Improved Speech Reconstruction from Silent Video

Ariel Ephrat* Tavi Halperin* Shmuel Peleg
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

Speechreading is the task of inferring phonetic information from visually observed articulatory facial movements, and is a notoriously difficult task for humans to perform. In this paper we present an end-to-end model based on a convolutional neural network (CNN) for generating an intelligible and natural-sounding acoustic speech signal from silent video frames of a speaking person. We train our model on speakers from the GRID and TCD-TIMIT datasets, and evaluate the quality and intelligibility of reconstructed speech using common objective measurements. We show that speech predictions from the proposed model attain scores which indicate significantly improved quality over existing models. In addition, we show promising results towards reconstructing speech from an unconstrained dictionary.

1. Introduction

Human speech is inherently an articulatory-to-auditory mapping in which mouth, vocal tract and facial movements produce an audible acoustic signal containing phonetic units of speech (phonemes) which together comprise words and sentences. *Speechreading* (commonly called *lipreading*) is the task of inferring phonetic information from these facial movements by visually observing them. Considering the fact that speech is the primary method of human communication, people who are deaf or have a hearing loss find that speechreading can help overcome many of the barriers when communicating with others [5]. However, since several phonemes often correspond to a single viseme (visual unit of speech), it is a notoriously difficult task for humans to perform.

We believe that machine speechreading may be best approached using the same form of articulatory-to-acoustic mapping that creates the natural audio signal, even though not all relevant information is available visually (e.g. vocal chord and most tongue movement). In addition to the perceptual sense this approach makes, modeling the task as an acoustic regression problem has many advantages over the visual-to-textual or classification modeling: (i) Acous-

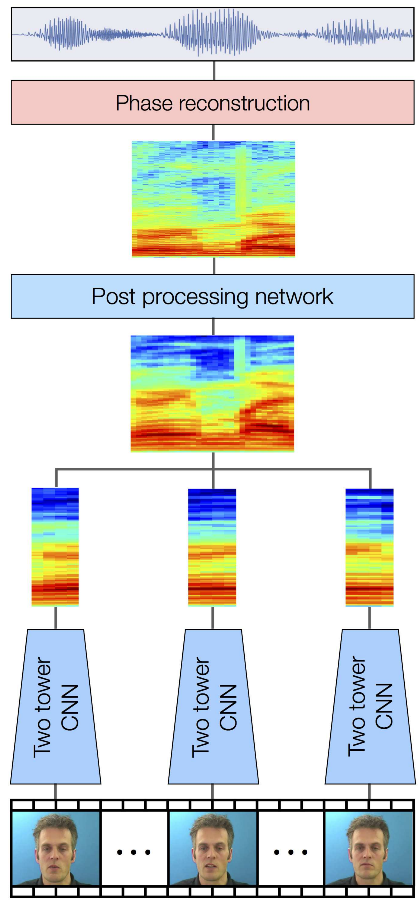


Figure 1: Proposed two-tower CNN-based model converts silent frames of a speaking person to intelligible and natural sounding speech. The CNN generates multiple mel-scale spectrograms which are converted by a post-processing network to a longer range linear-scale spectrogram. The frames-to-spectrogram model is trained end-to-end. Phase reconstruction transforms the long-range spectrogram into waveform.

*indicates equal contribution

tic speech signal contains information which is often difficult or impossible to express in text, such as emotion, prosody and word emphasis; (ii) This form of cross-modal regression, in which one modality is used to generate another modality, can be trained using "natural supervision" [30] which leverages the natural synchronization in a video of a talking person. Recorded video frames and recorded sound do not require any segmentation or labeling; (iii) By regressing very short units of speech, we can learn to reconstruct words comprised of these units which were not "seen" during training. While this is also possible by classifying short visual units into their corresponding phonemes or characters, in practice generating labeled training data for this task is difficult.

Several applications come to mind for automatic video-to-speech systems: Enabling videoconferencing from within a noisy environment; facilitating conversation at a party with loud music between people having wearable cameras and earpieces; maybe even using surveillance video as a long-range listening device. In another paper we have successfully used the generated sound, together with the original noisy sound, for speech enhancement and separation [13].

Our technical approach builds on recent work in neural networks for speechreading and speech synthesis, which we extend to the problem of generating natural sounding speech from silent video frames of a speaking person. To the best of our knowledge, there has been relatively little work for reconstructing high quality speech using an unconstrained dictionary. Our work is also closely related to efforts to extract textual information from a video of a person speaking, i.e. the visual-to-textual problem, as the output of our model can potentially also be used as input to a speech-to-text model.

In this paper, we: (1) present and compare multiple CNN-based encoder-decoder models that predict the speech audio signal of a silent video of a person speaking, and significantly improve both intelligibility and quality of speech reconstructions of existing models; (2) show significant progress towards reconstructing words from an unconstrained dictionary.

2. Related work

Much work has been done in the area of automating speechreading by computers [32, 29, 44]. There are two main approaches to this task. The first, and the one most widely attempted in the past, consists of modeling speechreading as a visual-to-textual mapping. In this approach, the input video is manually segmented into short clips which contain either whole words from a predefined dictionary, or parts of words comprising phonemes, visemes [4] or characters. Then, visual features are extracted from the frames and fed to a classifier. Assael *et al*. [3], Chung *et al*. [7] and others [42, 39, 33] have all recently showed state-of-the-art word and sentence-level classification results using neural network-based models.

The second approach, and the one used in this work, is to model speechreading as a visual-to-acoustic mapping problem in which the "label" of each short video segment is a corresponding feature vector representing the audio signal. Kello and Plaut [25] and Hueber and Bailly [20] attempted this approach using various sensors to record mouth movements, whereas Cornu and Milner [9] used active appearance model (AAM) visual features as input to a recurrent neural network.

Our approach is closely related to recent speaker-dependent video-to-speech work by Ephrat and Peleg [11], in which a convolutional neural network (CNN) is trained to map raw pixels of a speaker's face directly to audio features, which are subsequently converted into intelligible waveform. The differences between our approach and the one taken by [11] can by broken down into two parts: improvement of the encoder and redesign of the decoder.

The goal of our encoder modification is to improve analysis of facial movements, and consists of a preprocessing step which registers the face to a canonical pose, the addition of an optical flow branch, and swapping the VGG-based architecture with a ResNet based one. Our decoder is designed to remedy a major flaw in [11], namely the unnatural sound of the reconstructed speech it produces. To this end, we use the sound representation and post-processing network of [43], which introduces longer-range dependency into the final speech reconstruction, resulting in smoother, higher quality speech. Section 4 expounds on the above differences, and Section 6.1 contains a comparison of the results of [11] to ours.

Our work also builds upon recent work in neural sound synthesis using predicted spectrogram magnitude, including Tacotron (Wang *et al.*) [43] for speech synthesis, and the baseline model of NSynth (Engel *et al.*) for music synthesis. While Tacotron focuses on building a single-speaker text-to-speech (TTS) system, our paper focuses on building a single-speaker video-to-speech system.

This work complements and improves upon previous efforts in a number of ways: Firstly, we explore how to better analyze the visual input, i.e. silent video frames, in order to produce an encoding which can be subsequently decoded into speech features. Secondly, while prior work has predicted only the output corresponding to a single video frame, we jointly generate audio features for a sequence of frames, as depicted in Figure 1, which improves the smoothness of the resulting audio. Thirdly, [11] focused on maximizing intelligibility at the expense of natural sounding speech on a relatively limited-vocabulary dataset. We aim to overcome the challenges of the more complex TCD-TIMIT [16] dataset, while optimizing for both intelligibility

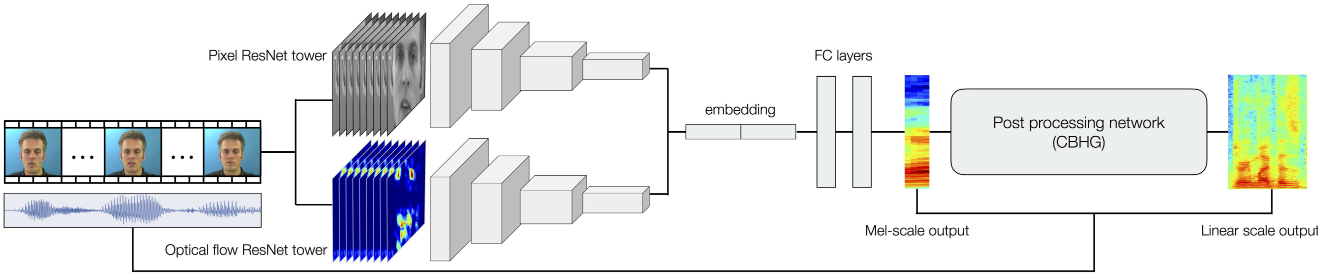


Figure 2: This figure illustrates our model's two-tower CNN-based encoder-decoder architecture. First, a face is detected in the silent input video. The face undergoes in-plane registration, cropping and resizing. One tower inputs grayscale face images, and the other inputs optical flow computed between frames. An embedding is created by concatenating the outputs of each tower. It is subsequently fed into the decoder which consists of fully connected layers which output mel-scale spectrogram, and a post-processing network which outputs linear-scale spectrogram.

and natural-sounding speech.

3. Data representation

3.1. Visual representation

Our goal is to reconstruct a single audio representation vector S_i which corresponds to the duration of a single video frame I_i. However, instantaneous lip movements such as those in isolated video frames can be significantly disambiguated by using a temporal neighborhood as context. Therefore, the encoder module of our model takes two inputs: a clip of K consecutive grayscale video frames, and a "clip" of $(K-1)$ consecutive dense optical flow fields corresponding to the motion in (u, v) directions for pixels of consecutive grayscale frames.

Each clip is registered to a canonical frame of reference. We start by detecting 5 facial points (two eyes, nose, and two tips of the mouth). We use the points on the eyes to compute a similarity transform alignment between each frame and the central frame of the clip. Following [11], we then crop the speaker's full face to a size of $H \times W$ pixels, and we use the entire face region rather than using only the region of the mouth. This results in an input volume of size $H \times W \times K$ scalars. The second input, dense optical flow, adds an additional volume of $H \times W \times (K-1) \times 2$ scalar inputs. It has been proven that optical flow can improve the performance of neural networks when combined with raw pixel values for a variety of applications [36, 12], and has even been successfully used as a stand-alone network input [34]. Optical flow is positively influential in this case as well, as we show later.

3.2. Speech representation

The challenge of finding a suitable representation for an acoustic speech signal which can be estimated by a neural network on one hand, and synthesized back into intelligible audio on the other, is not trivial. Use of raw waveform as network output was ruled out for lack of a suitable loss function with which to train the network.

Line Spectrum Pairs (LSP) [22] are a representation of Linear Predictive Coding (LPC) coefficients which are more stable and robust to quantization and small deviations. LSPs are therefore useful for speech coding and transmission over a channel, and were used by [11] as output from their video-to-speech model. However, without the original excitation, the reconstruction using unvoiced excitation (random noise) results in somewhat intelligible, albeit robotic and unnatural sounding speech.

Given the above, we sought to use a representation which retains speech information vital for an accurate reconstruction into waveform. We experiment with both spectrogram magnitude and a reduced dimensionality mel-scale spectrogram as our regression target, which can subsequently be transformed back into waveform by using a phase reconstruction algorithm such as Giffin-Lim [15].

4. Model architecture

At a high-level, as shown in Figure 2, our model is a comprised of an encoder-decoder architecture which takes silent video frames of a speaking person as input, and outputs a sound representation corresponding to the speech spoken during the duration of the input.

It is important to note that our proposed approach is speaker-dependent, i.e. a new model needs to be trained for each new speaker. Achieving speaker-independent speech reconstruction is a non-trivial task, and is out of the scope of this work.

4.1. Encoder

The encoder module of our model consists of a dual-tower Residual neural network (ResNet[18]) which takes the aforementioned video clip and its optical flow as inputs and encodes them into a latent vector representing the visual features. Each of the inputs is processed with a column of residual block stacks. Each tower comprises ten consecutive $conv1\times1 - BatchNormalization - Relu - conv3\times3 - BatchNormalization - Relu - conv1\times1 - BatchNormalization - Add - Relu$ blocks consisting of $128-128-128-256-256-256-256-512-512-512-512$ kernels. Following the last layer, a global spatial average (5×4) is performed, after which the two towers are concatenated into one 1024-neuron layer which is essentially a latent representation of our visual features.

4.2. Decoder and post-processing

The latent vector is fed into a series of two fully connected layers with 1024 neurons each. The last layer of our CNN is of size $n \times l$, where l is the number of audio windows predicted given a single video clip, and n corresponds to the size of the sound representation vectors we wish to predict. For example, an output of 80 coefficients of Mel frequency with $hopsize = windowsize/4$ and window size of 500, results in an $80 \times 4 = 320$ dimensional output vector. The output of our CNN is fed into the post-processing network used by [43], consisting of one CBHG module, which is described as a powerful module for extracting representations from sequences. The CBHG module comprises several convolutional, Highway [38] and Bidirectional GRU [6] layers whose goal is to take several consecutive sound representations as input and output a higher temporal resolution version. The input clips are then packed in mini-batches of T consecutive samples. As in the implementation of [43], the post processing network takes these consecutive mel-scale spectrogram vectors as input, and outputs a T consecutive linear-scale spectrogram. The entire model is trained end-to-end with a two-term loss, one on the decoder output and one on the output of the post-processing network. Although the model is trained end-to-end, we keep all convolutional layers of the CNN frozen during training of the second network.

4.3. Generating a waveform

We consider several methods for generating a waveform from our model's predicted mel and linear scale sound features. The first is the spectrogram synthesis approach used by [43, 10] in which the Griffin-Lim algorithm is used to reconstruct the phase of the predicted linear-scale spectrogram. Inverse STFT is then used to convert the complex spectrogram back into waveform. We found that the result is intelligible and smooth, albeit somewhat unnatural and robotic sounding.

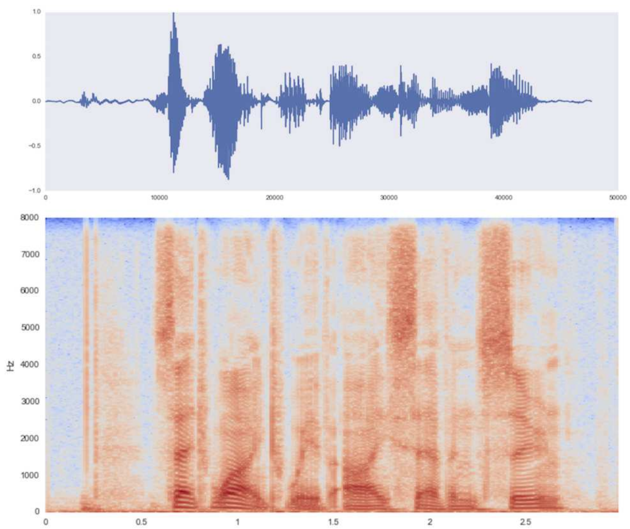
Therefore, we also consider an example-based synthesis method similar to the one used by [30], in which we replace predicted sound features with their closest exemplar in the training set. We search for the nearest neighbor to both mel-scale and linear-scale predicted features, as measured by L_2 distance, and replace it with the neighbor's corresponding linear-scale spectrogram feature. The full spectrogram is then converted into waveform using the procedure described above. We find that in most cases, mel-scale gives better results, which are more natural-sounding, but less smooth than using the predicted linear spectrogram itself.

5. Model details

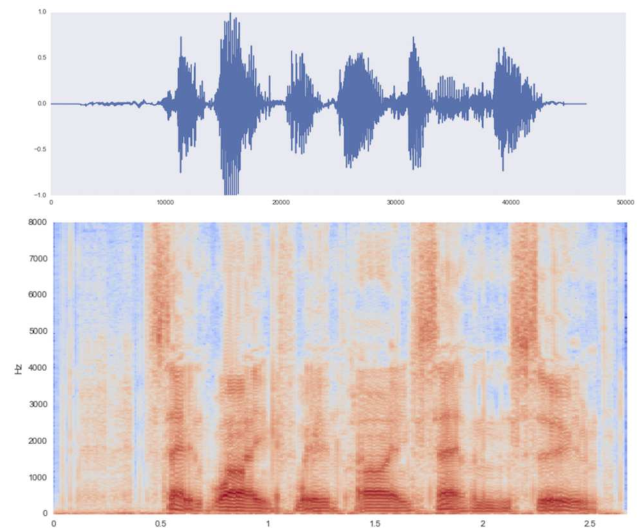
We use the method of [40], with the code provided from their website, to detect facial features. The speaker's face is cropped to 160×128 pixels. Using $K = 9$ frames as input worked best. This results in an input volume of size $160 \times 128 \times 9$ scalars, from which we subtract the mean. We use the method of [27] with python wrapper provided by [31] to compute an optical flow vector (u, v) for every image pixel. Optical flow is not normalized as its mean is approximately zero, and its std is in the range of $0.5 - 1.5$ pixels. [23, 31].

We use the code provided by [10] to compute log magnitude of both linear and mel-scale spectrograms which are peak normalized to be between 0 and 1. For videos with a frame rate of 25 FPS we downsample the original audio to 16 kHz and use 40 ms windows with 10 ms frame shift. For videos with a frame rate of 29.97 FPS we downsample the original audio to 14985 Hz and use 33.3 ms windows with 8.3 ms frame shift.

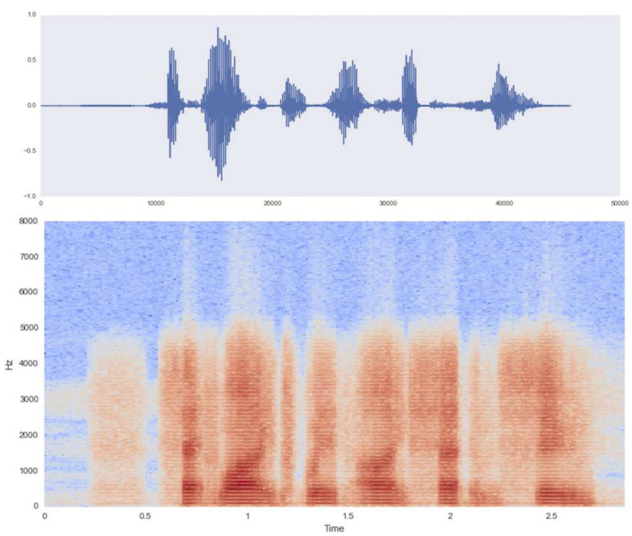
Our network implementation is based on the *Keras* library [1] built on top of *TensorFlow* [2]. Network weights are initialized using the initialization procedure suggested by He *et al.* [17]. Before each activation layer Batch Normalization [21] is performed. We use Leaky ReLU [28] as the non-linear activation function in all layers but the last two, in which we use the hyperbolic tangent (tanh) function. Adam optimizer [26] is used with an initial learning rate of 0.001, which is reduced several times during training. Dropout [37] is used to prevent overfitting, with a rate of 0.25 after convolutional layers and 0.5 after fully connected ones. We use mini-batches of 16 training samples each and stop training when the validation loss stops decreasing (around 100 epochs). The network is trained with backpropagation using mean squared error (L_2) loss for both decoder and post-processing net outputs, which have equal weights. To improve the temporal smoothness of the output, after generating spectrogram coefficients for T consecutive frames ($1...T$), we move one step forward and do the same for ($2...T+1$), which creates an overlap of $T-1$ frames, thus creating exactly T predictions for each frame. We then calculate a weighted average over the predictions for a given frame using a Gaussian.



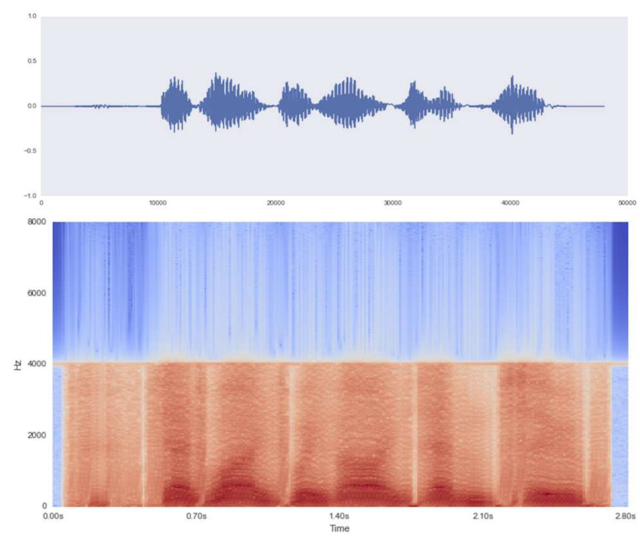
(a) Original



(b) Ours



(c) Le Cornu *et al.* [9]



(d) Vid2speech [11]

Figure 3: Visualization of original vs. reconstructed speech results on one GRID speaker $S3$ utterance. Original waveform and spectrogram are shown in (a). (b) shows the results of our reconstruction. The results of [9] and [11] are depicted in (c) and (d), respectively. Best viewed on a color display.

6. Experiments

Datasets Previous works performed experiments with the GRID audiovisual sentence corpus [8], a large dataset of audio and video (facial) recordings of 1000 sentences spoken by 34 talkers (18 male, 16 female). Each sentence consists of a six word sequence of the form shown in Table 1, e.g. "Place green at H 7 now". Although this dataset contains a fair volume of high quality speech videos, it has several limitations, the most important being its extremely small vocabulary. In order to compare our method with previous ones, we too perform experiments on this dataset.

In order to better demonstrate the capability of our model, we also perform experiments on the TCD-TIMIT dataset [16]. This dataset consists of 60 volunteer speakers with around 200 videos each, as well as three *lipspeakers*, people specially trained to speak in a way that helps the deaf understand their visual speech. The speakers are recorded saying various sentences from the TIMIT dataset [14], and are recorded using both front-facing and 30 degree cameras.

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	minus W		now
place	red	in			please
set	white	with			soon

Table 1: GRID sentence grammar.

Evaluation We evaluated the quality and intelligibility of the reconstructed speech using four well-known objective scores: STOI [41] and ESTOI [24] for estimating the intelligibility of the reconstructed speech and automatic mean opinion score (MOS) tests PESQ [35] and VisQOL [19], which indicate the quality of the speech. While all objective scores have their faults, we found that these metrics correlate relatively well with our perceived audio quality, as well as with our model's loss function. However, we strongly encourage readers to watch and listen to the supplementary video on our project web page [1] which conclusively demonstrates the superiority of our approach.

6.1. Sound prediction tasks

GRID For this task we trained our model on a random $80/20$ train/test split of the 1000 videos of speakers $S1-3$ (male) and $S4$ (female), and made sure that all 51 GRID words were represented in each set. The resulting representation vectors were converted back into waveform using the aforementioned mel-scale spectrogram example-based synthesis (*Mel-synth*) and predicted linear spectrogram synthesis (*Lin-synth*).

Table 2 shows the results of reconstructing the speech of $S1-4$ and Table 3 shows a comparison of our best results to the best results of [9]. Figure 3 shows a visualization of our results on one $S3$ video, and compares it to the original, and the results of [9, 11].

TCD-TIMIT For this task we trained our model on a random $90/10$ train/test split of the 254 videos of Lipspeakers $1-3$ (female). This results in less than 25 minutes of video used for training, which only around 60% of the amount of data used in the previous task. In this task, many of the words in the test set do not appear in the training set. We would like for our model to learn to reconstruct these words based on the recognition of the combinations of short visual segments which comprise the words. Here too, the resulting audio vectors were converted back into waveform using mel-scale spectrogram example-based synthesis (*Mel-synth*) and predicted linear spectrogram synthesis

[1] Examples of reconstructed speech can be found at http://www.vision.huji.ac.il/vid2speech

	STOI	ESTOI	PESQ	ViSQOL
GRID S1				
Mel-synth	0.442	0.26	1.782	3.108
Lin-synth	**0.475**	**0.263**	**1.952**	**3.324**
GRID S2				
Mel-synth	0.631	0.434	2.107	3.07
Lin-synth	**0.667**	**0.462**	**2.136**	**3.316**
GRID S3				
Mel-synth	0.666	**0.398**	**1.974**	3.164
Lin-synth	**0.68**	0.354	1.904	**3.349**
GRID S4				
Mel-synth	0.644	0.429	1.809	3.092
Lin-synth	**0.7**	**0.462**	**1.922**	**3.3**

Table 2: Objective measurements of reconstructed speech for model trained on GRID speakers $S1-4$. Mel-scale spectrogram example-based synthesis (*Mel-synth*) and predicted linear spectrogram synthesis (*Lin-synth*) were used to generate waveform from the predicted audio representations.

.

	STOI	ESTOI	PESQ	ViSQOL
GRID S3				
Vid2speech [11]	0.638	–	1.875	3.13
Cornu *et al*. [9]	–	**0.437**	**2.055**	–
Ours	**0.68**	0.398	1.974	**3.349**
GRID S4				
Vid2speech [11]	0.584	–	1.19	2.691
Cornu *et al*. [9]	–	0.434	1.686	–
Ours	**0.7**	**0.462**	**1.922**	**3.3**

Table 3: Comparison to [11] and [9] using objective measurements.

(*Lin-synth*).

Table 4 holds the results of this difficult task. The reconstructed speech is natural-sounding, albeit not entirely intelligible. Given the limited amount of training data, we believe our results are promising enough to indicate that fully intelligible reconstructed speech from unconstrained dictionaries is a feasible task.

	STOI	ESTOI	PESQ	ViSQOL
Lipspeaker 1				
Mel-synth	0.274	0.116	1.309	2.533
Lin-synth	**0.318**	**0.171**	**1.384**	**2.943**
Lipspeaker 2				
Mel-synth	0.307	0.123	1.08	2.431
Lin-synth	**0.357**	**0.157**	**1.158**	**2.931**
Lipspeaker 3				
Mel-synth	0.565	0.373	1.484	2.834
Lin-synth	**0.63**	**0.447**	**1.612**	**3.201**

Table 4: Objective measurements of reconstructed speech for model trained on TCD-TIMIT Lipspeakers 1-3. Many words in the testing set were not present in the training set, which results in significantly worse results than those on GRID dataset. The Lipspeaker 3 model performs best, by a large margin.

.

	STOI	PESQ	ViSQOL
Optical flow only	0.381	1.478	2.786
Pixels only	0.67	1.949	3.179
Pixels + optical flow	0.665	1.921	3.173
Pixels + OF + postnet	**0.68**	**1.974**	**3.349**

Table 5: Results of ablation analysis on GRID speaker $S3$. Pixels provide most of the information needed, adding optical flow and post-processing network gives slightly better results

.

6.2. Ablation analysis

We conducted a few ablation studies on GRID speaker $S3$ to understand the key components in our model. We compare our full model with (i) a model using only an optical flow stream; (ii) a model using only a pixel stream; (iii) a model with no post-processing network which outputs mel-scale spectrogram. Table 5 shows the results of this analysis. Our analysis shows that pixel intensities provide most of the information needed for reconstructing speech, while adding optical flow and a post-processing network give slightly better results.

7. Concluding remarks

A two-tower CNN-based model is proposed for reconstructing intelligible and natural-sounding speech from silent video frames of a speaking person. The model is trained end-to-end with a post-processing network which fuses together multiple CNN outputs in order to obtain a longer range speech representation. We have shown that the proposed model obtains significantly higher quality reconstructions than previous works, and even shows promise towards reconstructing speech from an unconstrained dictionary.

The work described in this paper can be improved upon by increasing intelligibility of speech reconstruction from an unconstrained dictionary, and extending an existing model to unknown speakers. It can also be used as a basis for various speech related tasks such as speaker separation and enhancement.

We plan to use the visually reconstructed speech in order to enhance speech recorded in a noisy environment, and to separate mixed speech in scenarios like the "cocktail party" where the face of the speaking person is visible [13].

Acknowledgment. This research was supported by Israel Science Foundation, by DFG, and by Intel ICRI-CI.

References

[1] Keras. Software available from `https://github.com/fchollet/keras`.

[2] Tensorflow. Software available from `http://tensorflow.org/`.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[4] H. L. Bear and R. Harvey. Decoding visemes: Improving machine lip-reading. In *ICASSP'16*, pages 2009–2013, 2016.

[5] D. Burnham, R. Campbell, G. Away, and B. Dodd. *Hearing Eye II: The Psychology Of Speechreading And Auditory-Visual Speech*. Psychology Press, 2013.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

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

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[9] T. L. Cornu and B. Milner. Generating intelligible audio speech from visual speech. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[10] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

[11] A. Ephrat and S. Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[13] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg. Seeing through noise: Speaker separation and enhancement using visually-derived speech. *arXiv:1708.06767*, 2017.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[15] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[16] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Trans. Multimedia*, 17:603–615, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV'15*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[19] A. Hines, J. Skoglund, A. Kokaram, and N. Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015 (13):1–18, 2015.

[20] T. Hueber and G. Bailly. Statistical conversion of silent articulation into audible speech using full-covariance hmm. *Comput. Speech Lang.*, 36(C):274–293, Mar. 2016.

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

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

[23] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. *arXiv preprint arXiv:1612.05478*, 2016.

[24] J. Jensen and C. H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[25] C. T. Kello and D. C. Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, 116(4):2354–2364, 2004.

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

[27] C. Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML'13*, 2013.

[29] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

[30] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *CVPR'16*, 2016.

[31] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. *arXiv preprint arXiv:1612.06370*, 2016.

[32] E. D. Petajan. *Automatic lipreading to enhance speech recognition (speech reading)*. PhD thesis, University of Illinois at Urbana-Champaign, 1984.

[33] S. Petridis, Z. Li, and M. Pantic. End-to-end visual speech recognition with lstms. 2017.

[34] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[35] 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 *Proceedings of the Acoustics, Speech, and Signal Processing, 200. On IEEE International Conference - Volume 02*, ICASSP '01, pages 749–752, Washington, DC, USA, 2001. IEEE Computer Society.

[36] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[37] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[38] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[39] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[40] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[41] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[42] M. Wand, J. Koutn, et al. Lipreading with long short-term memory. In *ICASSP'16*, pages 6115–6119, 2016.

[43] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

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