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

LipSound: Neural Mel-spectrogram Reconstruction for Lip Reading

Leyuan Qu, Cornelius Weber, Stefan Wermter

Knowledge Technology, Department of Informatics, University of Hamburg,
Vogt-Koelln-Str. 30, 22527, Hamburg, Germany

{qu,weber,wermter}@informatik.uni-hamburg.de

Abstract

Lip reading, also known as visual speech recognition, has recently received considerable attention. Although advanced feature engineering and powerful deep neural network architectures have been proposed for this task, the performance still cannot be competitive with speech recognition tasks using the audio modality as input. This is mainly because compared with audio, visual features carry less information relevant to word recognition. For example, the voiced sound made while the vocal cords vibrate can be represented by audio but is not reflected by mouth or lip movement. In this paper, we map the sequence of mouth movement images directly to mel-spectrogram to reconstruct the speech relevant information. Our proposed architecture consists of two components: (a) the mel-spectrogram reconstruction front-end which includes an encoder-decoder architecture with attention mechanism to predict mel-spectrogram from videos; (b) the lip reading back-end consisting of convolutional layers, bi-directional gated recurrent units, and connectionist temporal classification loss, which consumes the generated mel-spectrogram representation to predict text transcriptions. The speaker-dependent evaluation results demonstrate that our proposed model not only generates quality mel-spectrograms but also outperforms state-of-the-art models on the GRID benchmark lip reading dataset, with 0.843% character error rate and 2.525% word error rate.

Index Terms: Mel-spectrogram reconstruction, lip reading, visual speech recognition

1. Introduction

Recently, automatic speech recognition (ASR) has accomplished a quantum leap, and advanced models are proposed achieving improved performance on a variety of benchmarks [1–4], reaching human parity on some tasks [5]. However, in realistic environments, the performance of ASR systems suffers from significant degradation because of environmental noise or ambient reverberation [6–9].

Inspired by human bimodal perception [10] in which both visual and auditory information are used to improve the comprehension of speech, a lot of effort has been spent on lip reading to predict text transcriptions directly from visual cues and improve the robustness of ASR [11–16]. The visual signal is invariant to acoustic noise and complementary to auditory representation [17, 18], and the visual contribution becomes more important as the acoustic speech-to-noise ratio is decreased [19].

Approaches to lip reading generally fall into two categories: (a) handcrafted visual feature extraction, in which many methods have been proposed based on visual signal processing algorithms. For instance, Discrete Cosine Transform [20], Discrete Wavelet Transform [21], Active Appearance Models [22]; (b) automatic feature extraction using neural networks. This has become the dominant technique in this task, for example, using convolutional auto-encoder [23], spatio-temporal convolutional neural networks [14], long short-term memory [24], and residual networks [12].

Although advanced feature engineering and powerful deep neural network architectures have been proposed, lip reading still cannot be competitive with speech recognition from audio. It is mainly because the visual modality carries less relevant information for recognition than audio. Furthermore, some phonemes are visually identical but different and discriminative in audio. For example, in English the minimal pairs /b/ and /p/, where /b/ is a voiced sound and /p/ is an unvoiced sound. They are different in audio-based speech representation, while the two phonemes are modeled as the same unit in traditional lip reading systems, since /b/ and /p/ are produced with the same visually apparent lip and tongue movement.

Inspired by the success of the Tacotron2 [25] which only requires text as input to predict the mel-spectrogram for speech synthesis, in this paper we propose to map the image sequences of the mouth region directly to mel-spectrogram to reconstruct the relevant acoustic information.

Our proposed model architecture consists of two main components, as shown in Figure 1: (i) a recurrent encoder and decoder with an attention mechanism front-end that generates mel-scale spectrograms from image sequences of video. Unlike end-to-end lip reading models, this component can be trained with large amounts of non-annotated video data. (ii) a lip reading back-end that maps the generated mel-spectrogram to text directly. We conduct the evaluation of the overall model on the lip reading benchmark GRID dataset.

During training, instead of consuming the predicted output from previous time steps, we use teacher-forcing training strategy [26] to utilize the ground truth speech spectrogram as input, which is different from the traditional lip reading architectures that map the sequence of images directly to text transcriptions. Besides, the temporal dependencies between consecutive acoustic frames enable the front-end model not only to reconstruct the segmental features, for example formants, but also the supra-segmental information, for instance speaking styles. The reconstructed speech-relevant information significantly improves the performance of the lip-reading back-end.

2. Dataset and Preprocessing

GRID [27] is a current benchmark and biggest open source lip reading dataset, which consists of in total 34000 videos from 34 speakers (16 female and 18 male) on sentence-level. The sentences have a fixed 6-word structure and are generated by a restricted grammar: $command^{(4)}$ + $color^{(4)}$ + $preposition^{(4)}$ + $letter^{(25)}$ + $digit^{(10)}$ + $adverb^{(4)}$, where the superscript means the number of candidate words, for example "Set red by Z two

http://dx.doi.org/10.21437/Interspeech.2019-1393

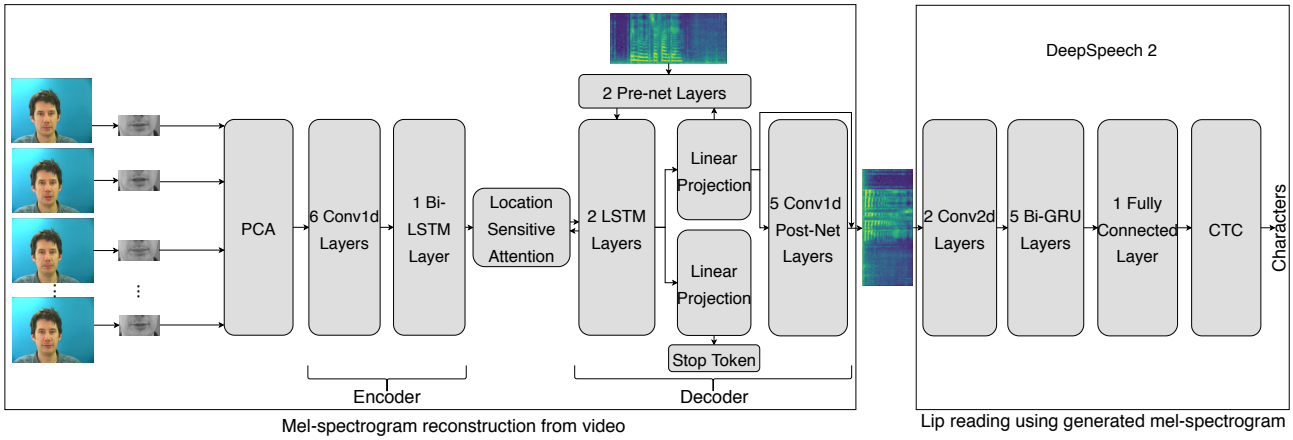

Figure 1: *LipSound model architecture. The front-end (left) is used for mel-spectrogram reconstruction and the back-end (right) is used for character recognition. Together, they perform lip reading.*

now". The category details are listed in Table 1.

Table 1: *GRID dataset word categories*

Categories	**Candidate Words**
Command	bin, lay, place, set
Color	blue, green, red, white
Preposition	at, by, in, with
Letter	A, . . . , Z (W excluded)
Digit	zero, . . . , nine
Adverb	again, now, please, soon

GRID consists of 34 speakers, but there are only 33 speakers' videos available with a total of 32669 sentences. We randomly select 255 sentences from each speaker used for evaluation. The remaining sentences are divided into training and development sets with a ratio of 9:1. Both training and test sets contain samples of all speakers, leading to speaker-dependent results.

All videos are 3 seconds long and with a frame rate of 25fps (total 75 frames). All frames are converted to gray images from original RGB color. To collect training data for the mel-spectrogram generator, we use FFmpeg to extract audio and downsample to 16kHz. We use Dlibs Python bindings [28] to detect 68 facial landmarks which are used to crop an area of 100x50 pixels around the mouth from each frame. Then we regularize the pixel value to [-1, 1] and normalize the mean value and standard deviation of all images to 0 and 1 respectively. The input shape of the mel-spectrogram generator is 75x512, where 75 (3x25) is the number of frames extracted from videos and 512 is the dimension of each mouth area compressed by PCA.

3. Model Architecture

3.1. Front-end: mel-spectrogram generator

The mel-spectrogram generator is inspired by the Tacotron2 which only uses character embedding and corresponding speech waveform as input for training, without requiring any linguistic, handcrafted features or domain expertise knowledge. When combining it with the Wavenet vocoder [29], the Tacotron2 achieves high-quality sounds which can be comparable to human natural speech. In the mel-spectrogram generator, we replace the character sequences with video representation as the model's inputs and remove the word embedding layer since it is unnecessary in our case.

As shown in Figure 1 left rectangle, our mel-spectrogram generator consists of an encoder, a decoder and an attention mechanism, which is similar to the Tacotron2 system. The encoder compresses the image sequences from the mouth area into latent vectors and the attention mechanism learns to align the encoder and decoder time steps and focuses on the most relevant information for the current step. Finally, the decoder consumes the attention context vector and all the previous information to predict the mel-spectrogram step by step.

The images of size 100x50 are compressed to 512 dimensions by PCA and then directly fed into 6 layers of a 1D convolutional neural network with 512 filters and a kernel size of 5. These CNN layers have a similar function as the N-grams used in natural language processing to capture the temporal information in multiple adjacent frames. Each CNN layer is followed by batch normalization and rectified linear unit (ReLU) activation functions. The outputs from CNN layers are consumed by one bi-directional LSTM layer.

Attention mechanisms have become standard in encoder and decoder architectures since they reduce computational complexity and let the model focus on the most relevant information. We use location-sensitive attention [30] to direct the information flow from the encoder to the decoder, which focuses both content and location information to predict the next decoding time step and yields smoother alignments.

The decoder consists of two LSTM layers, 2 fully connected layers (pre-net) and 1 linear projection layer, which evaluates an output mel-spectrogram one frame at a time. Only during training the ground truth mel-spectrogram extracted from the corresponding speech waveforms is fed into the pre-net layers. We use the predicted mel-spectrogram from previous time steps when inferring. Then the output from pre-net is concatenated with attention context vectors as the inputs of the following 2 LSTM layers. The output is used to generate mel-spectrogram by one sigmoid projection layer at this time step. In the meantime, the output from the LSTM layers is also consumed by another sigmoid linear layer to predict the stop token. This is useful for the inference phase since all sentences in the training set are zero-padded to have the same dimensionality. The stop token predicts when to terminate decoding and avoids

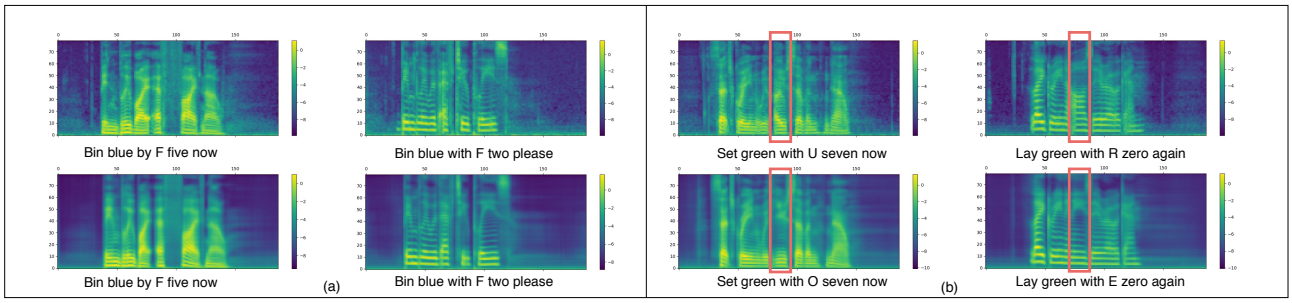

Figure 2: *The comparison of real (top row) and generated (bottom row) mel-spectrogram. (a) correct generation. (b) generated mel-spectrogram with word substitution (as marked with red rectangles).*

always generating the same duration and silence padding for sentences of short duration. Finally, to further improve the mel-spectrogram quality, the predicted mel-spectrogram is fed into 5 convolutional layers with residual connections, named post-net. Both the mel-spectrogram output from the linear layer and post-net are used for lip reading back-end model training.

3.2. Back-end: lip reading system

We use DeepSpeech 2 [1] ASR system as back-end module to transcribe spectrogram into text, as shown in Figure 1 (right rectangle), which begins with two layers of 2D convolutions, followed by five layers of gated recurrent units (GRU) [31] and a fully connected output layer. Finally, we use the connectionist temporal classification (CTC) loss [32] to calculate the difference between the predicted transcriptions and the ground truth.

4. Experiments

In this section, we introduce two audio based speech recognition systems as gold standard and conduct experiments to reconstruct mel-spectrogram from videos with the GRID dataset. The predicted spectrograms are evaluated on lip reading tasks.

4.1. Setups for mel-spectrogram generator and lip reading

The feature prediction experiments are conducted on a single NVIDIA 1080Ti GPU card with a fixed mini-batch size of 30. We used the Adam optimizer [33] with an initial learning rate of 0.001 and anneal the learning rate with a value of 1.1 after every 50000 iterations.

The input features for our lip reading systems are mel-spectrograms. The back-end neural networks are trained with the CTC loss function, using the stochastic gradient descent optimization strategy along with a mini-batch of 30 utterances per batch. We use 40 epochs and pick the model that performs best on the development set to quantify on the test set. Learning rates are chosen from [1e-4, 6e-4] and a learning rate annealing algorithm is used with the value of 1.1 after each epoch. The momentum is 0.9. Batch normalization is used to optimize models and accelerate training on hidden layers. All architectures described in this paper do not use any language models.

4.2. Audio gold standard models for lip reading

We use Word Error Rate (WER) and Character Error Rate (CER) as the evaluation metric. There are two strong audio gold standards in this work. Audio gold standard 1 is trained from scratch using only the mel-spectrogram features extracted directly from the original training set. Audio gold standard 1 achieves better performance, with 0.8% CER and 2.1% WER, than all lip reading systems [14–16] that only use the visual modality as input. This result also verifies that the speech modality contains more useful information for recognition than the visual modality.

To further improve the performance, we use the 960 hours LibriSpeech [34] training set to pretrain the audio gold standard model. LibriSpeech is a large open source speech corpus and a widely-used speech recognition benchmark. The pretrained acoustic model achieves 11.43% WER on the LibriSpeech clean evaluation set after 13 epochs. After fine-tuning 18 epochs on the pretrained LibriSpeech acoustic model, the audio gold standard 2 gets significant improvement with 0.2% CER and 0.6% WER on the GRID test set.

5. Results and Discussion

5.1. Results of mel-spectrogram reconstruction

As shown in Figure 2, we visualize the original mel-spectrogram (top row) extracted from the corresponding audio as references. We also show the mel-spectrogram samples (bottom row) generated from our feature prediction front-end.

Figure 2 (a) shows two correctly generated samples. As shown in the figure, the mel-spectrogram predicts highly similar details to the original one, with similar starting and ending time, low frequency and formants. We tend to attribute this performance to the good alignment learned by the attention mechanism between the encoder and decoder time steps. But the high frequencies seem fuzzy and are not as clear as in the original mel.

We transform the generated mel-spectrogram back to waveform using Griffin Lim algorithm [35] with 50 iterations. The Griffin Lim algorithm is widely used to restore phase information for waveform reconstruction. We can verify that the mel-spectrogram generator can learn different speakers' voices. But by directly hearing the generated sounds, we find that some of the isolated letters have been substituted by another letter, as shown in Figure 2 (b). For example, the letter 'O' in the sentence 'Set green with O seven now' is replaced by 'U' and the letter 'R' in 'Lay green with R zero again' is replaced by 'E'. However, the other words are inferred correctly. Unlike the words with multiple letters, the isolated letters are independent of context. For the same left and right context, for example 'with+*+seven' where * means A-Z (excluding W), there are 25 possibilities. It is difficult for the decoder to infer a correct letter using the same context information. Besides, after checking the original sounds, we found that speakers tend to have a short pause before producing isolated letters. The short si-

lence affects attention alignment radically. Figure 3 shows the influence of silence on alignment. The beginning and end of the sentence is silence, causing a divergent alignment instead of an intensive yellow line. More audio samples are available on https://soundcloud.com/user-612210805/sets/video-to-mel.

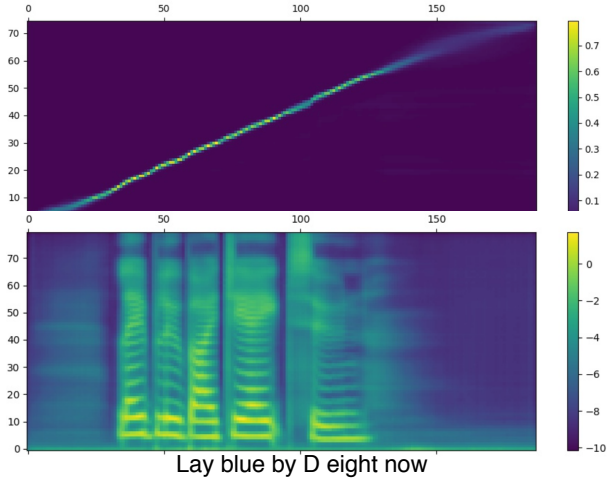

Figure 3: *Alignment between encoder and decoder time steps. Top: the attention mechanism alignment curve (yellow diagonal line). Bottom: mel-spectrogram generated from post-net.*

5.2. Results of lip reading

As expected, the gold standard models trained on audio outperform the models [14–16] trained only on visual information. This shows again that speech carries more useful information for recognition.

Table 2: *CER and WER comparison on the GRID lip reading dataset. All cited works use visual information as model inputs. Audio gold standard 1 is trained on the GRID audio dataset. Audio gold standard 2 is pre-trained on the LibriSpeech acoustic model. NoLM: no language models are used.*

Model	CER (%)	WER (%)
Audio:		
Gold standard 1-NoLM	0.811	2.053
Gold standard 2-NoLM	0.180	0.564
Visual:		
LipNet-NoLM [14]	2.0	5.6
LipNet [14]	1.9	4.8
WAS [15]	-	3.3
LCANet [16]	1.3	2.9
LipSound-NoLM	1.532	4.215
LipSound with pretrain-NoLM	**0.843**	**2.525**

Table 2 shows the comparison between our proposed LipSound and previous works. All cited works [14–16] predicted text transcriptions from videos directly. LipNet [14] is trained in an end-to-end fashion on a sentence-level which makes use of spatio-temporal convolutions and CTC and achieves 1.9% CER and 4.8 WER%. The WAS [15] network utilizes an encoder-decoder with an attention architecture and pretrains on the Lip Reading Sentences dataset which is a large-scale dataset for audio-visual speech recognition. The WAS model yields 3.3% WER on the GRID evaluation set. The LCANet networks introduced a cascaded attention-CTC decoder to further improve the performance and achieved 1.3% CER and 2.9% WER.

Our model trained on the visually generated mel-spectrogram achieves 1.532% CER and 4.215% WER. To further improve the accuracy, we fine-tune the lip reading model on the pretrained LibriSpeech model with updating all parameters. After 15 epochs, we get better than state-of-the-art performance with 0.843% CER and 2.525% WER.

Table 3 lists the comparison between the ground truth and the predicted text transcriptions. As reported [14], the frequently confused phoneme pairs are (d, t) and (b, p), while in our results, the most frequent errors are letter substitutions, such as (A, H) where 'A' is substituted by 'H'. This indicates that our mel-spectrogram front-end has reconstructed the lost information in the visual representation, while it needs to make guesses for phonemes that are easily confused in visually.

Table 3: *Comparison between ground truth and predicted sentence by our lip reading system. Mistaken words are underlined.*

Ground truth	Predicted sentencs
Lay blue in A seven please	Lay blue in H seven please
Place red in N zero soon	Place red in A zero soon
Lay red in O seven please	Lay red in I seven please
Bin green by L seven now	Bin green by S seven now
Lay green at X six soon	Lay green at X six sooen
Set blue in R three please	Set blue in R three pleae
Lay white at C seven now	Lay white it C seven now

6. Conclusions

We proposed a novel architecture, LipSound, for lip reading in which an encoder-decoder architecture with attention mechanism is used to reconstruct mel-spectrogram from the image sequences of videos directly. The encoder encodes source image sequences into a context vector, and the decoder decodes the context vector to predict a target mel-spectrogram. The attention mechanism learns to align the encoder and decoder time steps and to concentrate on the most relevant information. The lip reading back-end consumes the generated mel-spectrogram representation to predict text transcriptions. The speaker-dependent evaluation results on the GRID benchmark dataset demonstrate that our system outperforms state-of-the-art performance.

Since the GRID lip reading corpus is designed by a restricted grammar instead of spontaneous sentences, future work will focus on spontaneous speech and speaker-independent tasks. Furthermore, we are interested to combine both visual and audio modalities to improve the robustness of speech recognition system. Future work will apply our system in real human-machine interaction scenarios.

7. Acknowledgements

The authors gratefully acknowledge Dr. Manfred Eppe, Johannes Twiefel and Di Fu for their constructive feedback and discussions, and partial support from the China Scholarship Council (CSC), the German Research Foundation DFG under project CML (TRR 169), and the European Union under project SECURE (No 642667).

8. References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[6] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1315–1329, 2016.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[8] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 547–554.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTMs recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[10] J. Besle, A. Fort, C. Delpuech, and M.-H. Giard, "Bimodal speech: early suppressive visual effects in human auditory cortex," *European Journal of Neuroscience*, vol. 20, no. 8, pp. 2225–2234, 2004.

[11] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2592–2596.

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

[13] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," *arXiv preprint arXiv:1806.06053*, 2018.

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

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[16] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 548–555.

[17] J. MacDonald and H. McGurk, "Visual influences on speech perception processes," *Perception & Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978.

[18] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[19] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[20] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Seventh International Conference on Spoken Language Processing*, 2002.

[21] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM-based automatic lipreading," in *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*. IEEE, 1998, pp. 173–177.

[22] G. Sterpu and N. Harte, "Towards lipreading sentences with active appearance models," *arXiv preprint arXiv:1805.11688*, 2018.

[23] D. Parekh, A. Gupta, S. Chhatpar, A. Y. Kumar, and M. Kulkarni, "Lip reading using convolutional auto encoders as feature extractor," *arXiv preprint arXiv:1805.12371*, 2018.

[24] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6115–6119.

[25] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[26] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[27] 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*, vol. 120, no. 5, pp. 2421–2424, 2006.

[28] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[29] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *INTERSPEECH*, 2017, pp. 1118–1122.

[30] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[31] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[32] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

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

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[35] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.