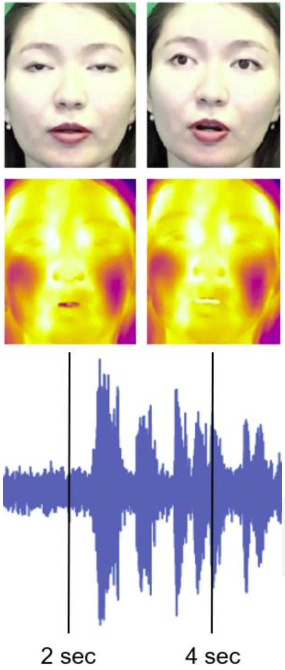
NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

Audio Visual Speech Recognition Using Visual and Thermal Images

Zhaniya Koishybayeva, Candidate for Master of Data Science Degree

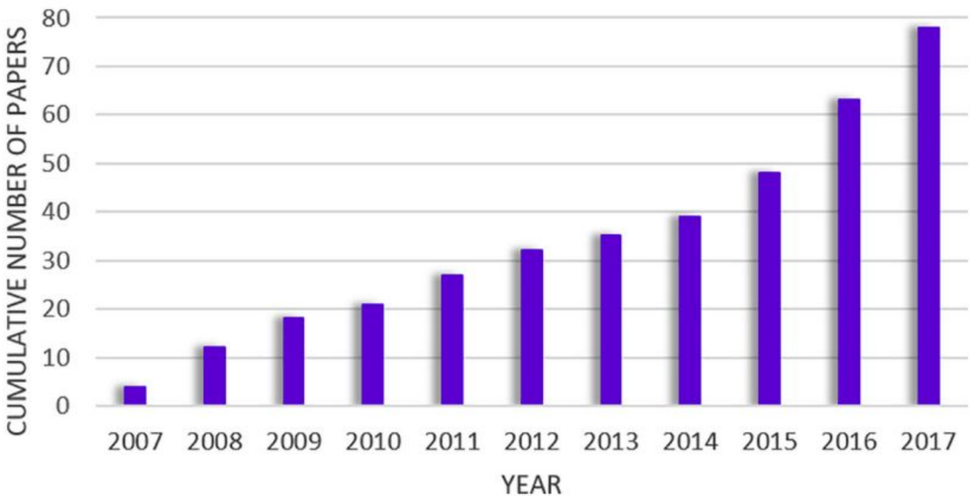
Objectives

- To examine if the performance of lipreading systems can be improved by including thermal image to the visual image stream
- Adapt and modify existing architecture
- Use different data input types: visual, thermal and combined
- Evaluate the performance with standard metrics (WRR, STOI, ESTOI, PESQ)



Related works

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES



Cumulative number of lipreading papers (2007-2017) [1]

[1] Fernandez-Lopez, A. and Sukno, F.M., 2018. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, pp.53-72.

Related Works

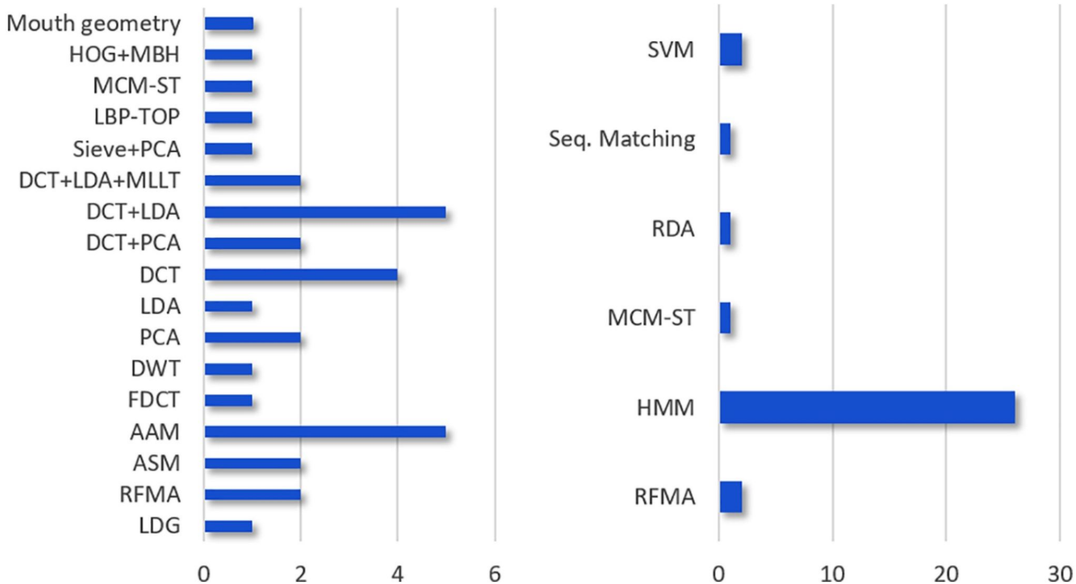
"Lip Reading Using Video and Thermal Images" (2006) [2]

- Dataset with 3 speakers uttering 5 word
- Eigen image + DP matching

- Visual: 76.0%
  Thermal: 44.0%
  Both: 80.0%

[2] Saitoh, T. and Konishi, R., 2006, October. Lip reading using video and thermal images. In *2006 SICE-ICASE International Joint Conference* (pp. 5011-5015). IEEE.

Related works



Number of times that each feature technique has been used (left) and number of times that each classification method has been used (right) from 2007 to 2017 [1]

[1] Fernandez-Lopez, A. and Sukno, F.M., 2018. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, pp.53-72.

Related Works

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

"Lipreading with Long Short-Term Memory" (2016) [3]

Dataset	Feature Extractor	Classifier	WRR
GRID	Eigenlips	SVM	70.6%
	HOG	SVM	71.3%
	Feed-forward	**LSTM**	**79.6%**

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

Related Works

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

“Lipreading Using Temporal Convolutional Networks.” (2020) [4]

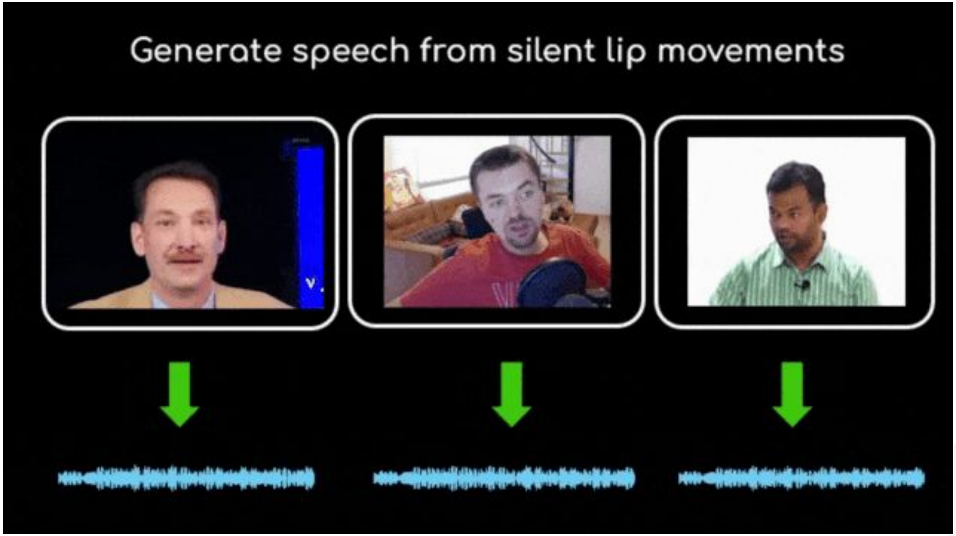
Dataset	Feature Extractor	Classifier	WRR
LRW	CNN	ResNet+TCN	Video: 85.30% Audio: 98.46% Both: 98.96%

[4] Martinez, B., Ma, P., Petridis, S. and Pantic, M., 2020, May. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6319-6323). IEEE.

Related Works

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

“Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis” (2020) [5]



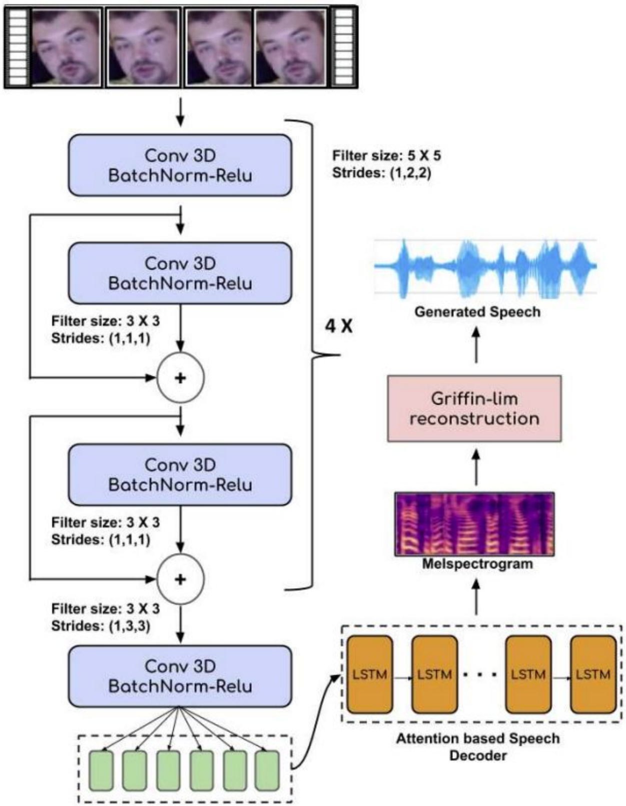
[5] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V., 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13796-13805).

Lip2Wav Advantages

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

- State-of-the-art lip2speech model
- Works for both shot in fixed environment and speaking in the wild datasets
- ROI - the entire face
- Speaker specific, only one subject necessary
- Up-to-date deep learning methods

Lip2Wav Architecture



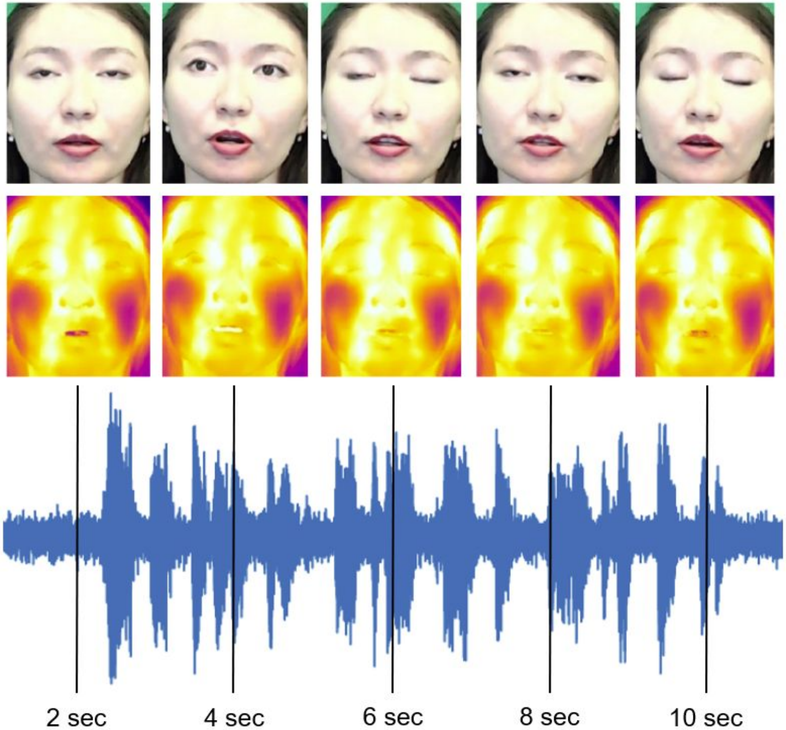
- Preprocessing:
  - Video to frames
  - Crop the ROI
- Encoder:
  - 3D-CNN
  - Skip connections
  - Batch normalizations
- Decoder:
  - Tacatron 2
- Ground truth - melspectogram

Lip2Wav Original Results

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

Dataset	Hours/ Speaker	STOI	ESTOI	PESQ	WRR
GRID	0.8	0.731	0.535	1.772	85.92%
TCD-TIMIT	0.5	0.558	0.365	1.350	68.74%
LRW	0.03-0.08	0.543	0.344	1.197	65.80%
Lip2Wav	20	0.416	0.284	1.300	-

Dataset Collection



- SpeakingFaces LipReading
- One speaker
- 1298 phrases (2h)
- Visual and thermal image streams

Dataset Preparation

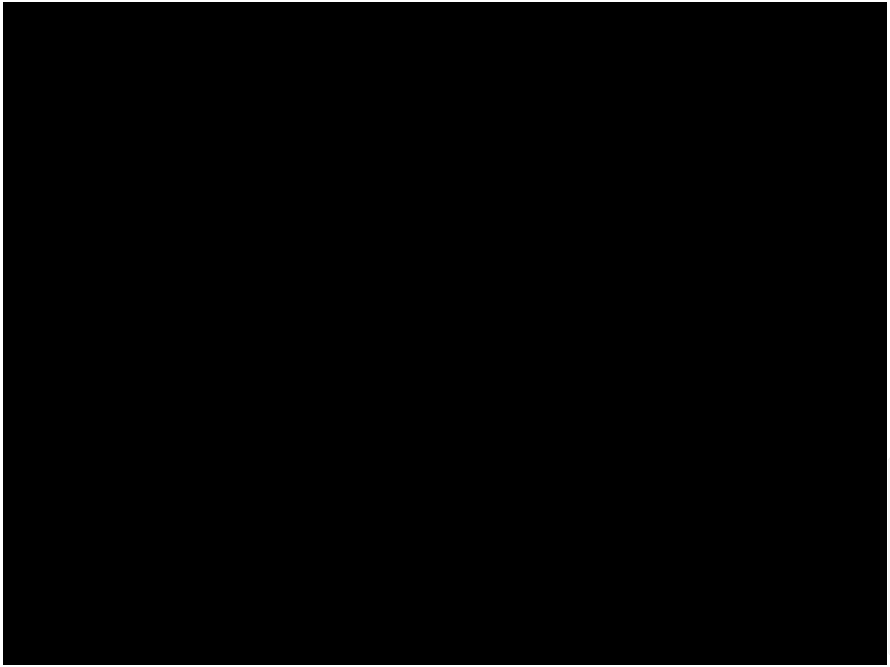
- Image aligning
  - Detecting lip landmarks on a visual frame
  - Matching the lips on the corresponding thermal image
  - Crop the same coordinates
- Audio normalization

Results and Discussion

Channels	STOI	ESTOI	PESQ	WRR
Visual	0.134	0.041	1.395	14.2%
Thermal	0.045	0.002	1.141	0.0%
Both	0.125	0.031	1.372	14.3%

Results and Discussion

"Music channels in YouTube"



Conclusion

- Lipreading performance increased in recent years
- WRR rarely exceed 85%

- Used Lip2Wav model to test the hypothesis
- Used new dataset for the model
- Conducted experiments on three different data streams
- Obtained and discussed the results

Improvements

- Improvements of the dataset:
  - Collection of additional data
  - Refinements in the dataset collection
  - Include different views of the ROI
  - Separate the utterances
- Further fine-tuning of the Lip2Wav
- Test on different lipreading systems (not necessarily lip2speech)

NAZARBAYEV UNIVERSITY
SCHOOL OF ENGINEERING AND DIGITAL SCIENCES

Thank you for your attention!