VID2SPEECH: SPEECH RECONSTRUCTION FROM SILENT VIDEO

Ariel Ephrat   Shmuel Peleg
The Hebrew University of Jerusalem

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

IEEE 2017 International Conference on Acoustics, Speech and Signal Processing

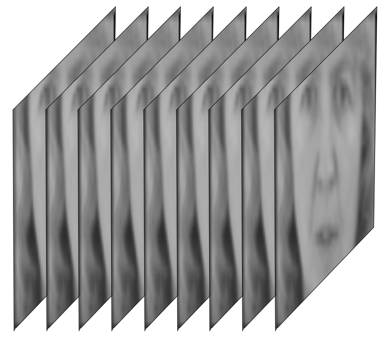
Speechreading Is Difficult

speech-reading ***n.*** : use of non-auditory clues as to what is being said, acquired by observing the speaker's facial expressions, lip and jaw movements. Formerly called **lip reading***

- Two approaches for automating speechreading:
  1. Classification: Output is word from predefined vocabulary, phoneme
  2. **Regression: Output is audio signal**
- Regression advantages:
  - No input pre-segmentation
  - Vocabulary-agnostic
  - Learn using "natural supervision"
  - Ability to output emotion, prosody
- Possible uses:
  - Videoconference from noisy environment
  - Surveillance video as "listening" device

* Adapted from Medical Dictionary for the Health Professions and Nursing and Mosby's Medical Dictionary

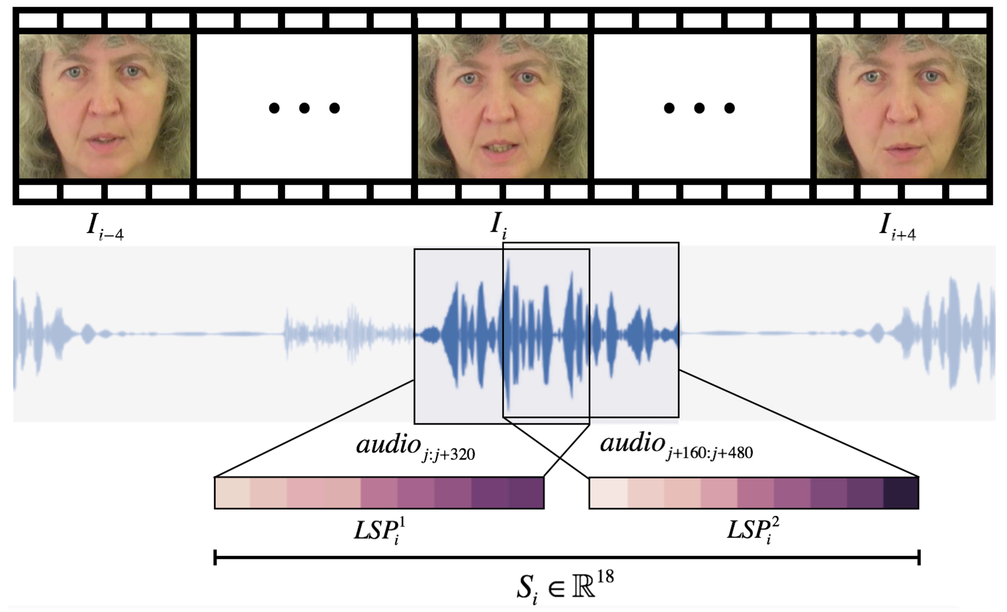
Visual Representation (Input)



- Speaker's face cropped and rescaled to 128 x 128 pixels
- K consecutive grayscale frames (K=9 worked best)

⟹ CNN input volume of 128 x 128 x 9 numbers

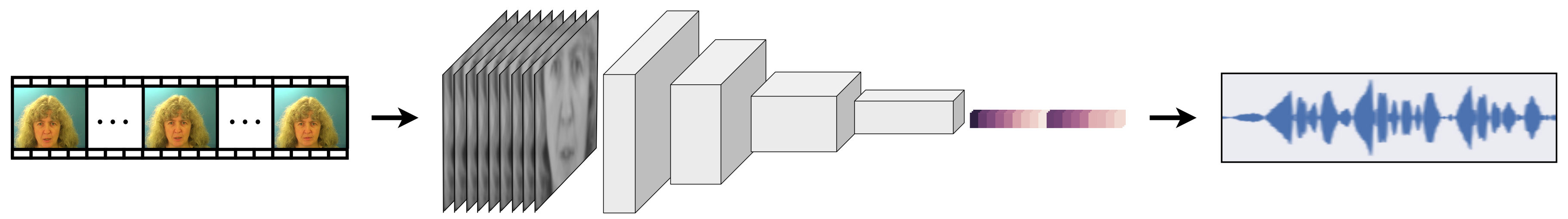
Speech Representation (Output)



- Audio downsampled to 8 KHz
- 8th order LPC and LSP on half-overlapping 40ms waveform segments
- Concatenate every 2 successive LSP vectors

⟹ Network output of 18 numbers

Speech Reconstruction Model



- VGG-like convolutional neural network:
  - 5 *conv3-conv3-maxpool* blocks followed by 2 *fc* layers
- Trained with MSE loss
- *Unvoiced* excitation to synthesize speech from network output

GRID Corpus [24]

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

- Audio-visual recordings of 34 speakers
- Each has 1000 3-second videos @ 25 FPS containing 6-word sequence of form shown above

[24] Martin Cooke et al., "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, 2006

Evaluation

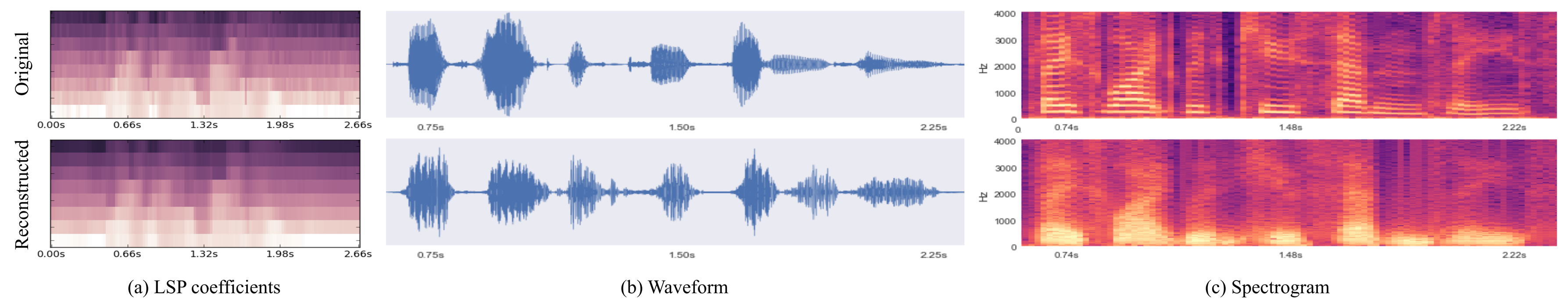
- Used Amazon MTurk for human intelligibility testing
- Followed protocol used by [10]
- Workers were given GRID vocabulary
- Each job was transcription of either:
  1. **Audio only** - reconstructed speech with no video
  2. **Audio-visual** - reconstructed speech with original video frames
  3. **Out-of-vocabulary Audio-Visual**
- Over 400 videos (38 distinct) transcribed by 23 people

amazon mechanical turk beta

[10] Thomas Le Cornu and Ben Milner, "Reconstructing intelligible audio speech from visual speech features," in *Conference of the International Speech Communication Association (Interspeech)*, 2015

Results

Visualization of Original vs. Reconstructed Speech



- Vertical columns of (a) are actual output of CNN
- In (c), unvoiced excitation causes lack of formants (horizontal lines)

1. Reconstruction from full dataset

Dataset:
- Train on 800 videos from one GRID speaker (60K frames)
- Test on remaining 200 videos

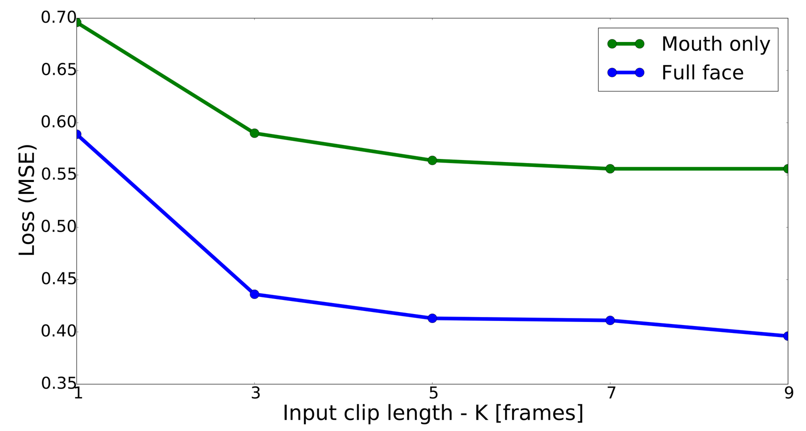
	[10] S4	Ours S4	Ours S2
Audio-only	40.0%	**82.6%**	-
Audio-visual	51.9%	**79.9%**	79%

2. Reconstruction of out-of-vocabulary (OOV) words

Dataset:
- Train on (S4) videos containing only 8 spoken digits
- Test on videos containing 2 OOV digits
- Results averaged across 5 splits

	Digits 0-9 OOV	Digits 0-9 None out	Chance
Audio-visual	51.6%	93.4%	10.0%

3. Learning from mouth only vs. full face



- Face region error is 40% lower than mouth only
- Disambiguation effect of using temporal context is clear