An attempt on Multimodal ASR

Using video images to improve accuracy

Joan Puigcerver i Pérez
joapuipe@upv.es

Motivation

- Most ASR systems use two sources of information to build their models.
  - Acoustic information
  - Prior knowledge about the spoken language

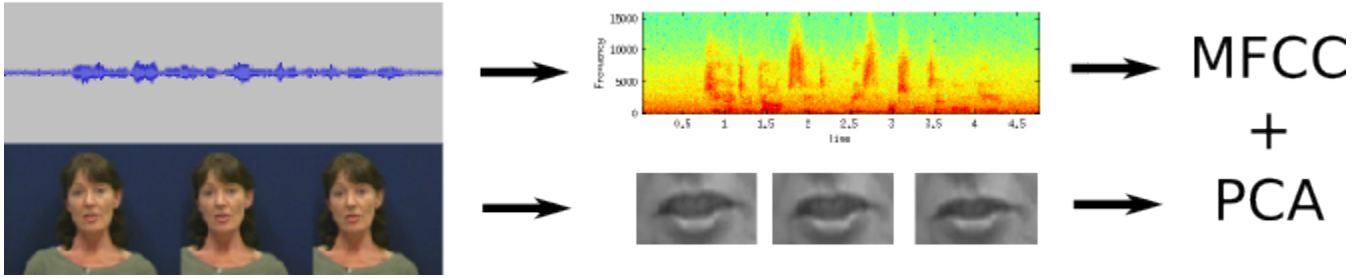
- However, there exist scenarios where we can use an additional source:
  - Visual information from the mouth, eyes, gestures, etc.

- Interesting in the transcription of:
  - TV news
  - Movies
  - Video lectures
  - ...

Motivation

- Deaf people use lipreading to help their communications.
- Some phonemes may be misunderstood by its acoustic representation, but are easily distinguishable using the lips:
  - [n], [m]
- However, the pronunciation of a phoneme depends on other mouth parts besides the lips:
  - tongue
  - glottis
  - air pressure
- Only 30-40% of English phonemes can be distinguished using only the lips. But, can the visual information help the recognition or it will just add noise?

Multimodal ASR system



- Audio features: 12 MFCC + Energy + First derivatives + Second derivatives
  - The regular 39-dimensional audio frames
- Video features: K-Principal Components of the detected lips
  - K was determined experimentally using the validation set
- The audio sampling frequency is much higher than video:
  - 32KHz vs. 25Hz
  - The final audio and video frames are linearly aligned

Dataset

- Hard to find a public and free dataset composed by both Audio + Video.

- The vidTIMIT corpus:
  - 43 different speakers (19 females, 24 males)
  - 10 sentences per speaker from the TIMIT corpus
  - Audio recorded at 16 bits and 32KHz
  - Video recorded at 512 x 384 pixels and 25fps
  - Video segmented in frames and stored as JPEG (90% qual.)
  - Only 27 minutes of audio!
  - Not all native-English speakers

- It does not include the transcription of the sentences, but you can get them from the TIMIT corpus.

Experiments

- Translectures-UPV Toolkit (a.k.a. AKToolkit)
- HMMs trained using monophonemes
  - 3 states per phoneme
- Gaussian mixtures as emission probability distribution
  - 8 components. Chosen by minimizing the validation error
- In addition to the 39 audio features, 4 PCA components used
  - Number of PCA components tuned using the validation set
- Bigram LM with Kneser-Ney smoothing using the training sentences

Results

	WER (%)	OOV (%)
Only Audio	45.14	10.51
Audio + Video	44.68	10.51

WER on validation set. Training: 370 sentences, Validation: 30 sentences.

	WER (%)	OOV (%)
Only Audio	56.50	12.60
Audio + Video	54.07	12.60

WER on test set. Training: 370+30 sentences, Test: 30 sentences.

Discussion

- Using visual information helped ASR in this experiment.
  - 2.43% WER reduction on Test
- Very few Principal Components used to represent the mouth.
  - Very few training data available, with higher dimensionality GMMs weren't estimated well enough.
- It's expected that with more data and more components, the improvement will be better.
  - Specially for noisy environments.
- Train/Valid/Test partition is not the standard one for this dataset, but I needed more training data!

Questions

Thanks!