Comparison of DCT and Autoencoder-based Features for DNN-HMM Multimodal Silent Speech Recognition

Licheng Liu, Yan Ji, Hongcui Wang, Bruce Denby
Tianjin Key Laboratory of Cognitive Computation and Application, Tianjin University, Tianjin, China

Motivation

- Mappings between articulatory movements and different units of speech
- DNN-HMMs on ultrasound-based SSR
- Approaches for non-acoustic feature extraction

Significance

- Applications in different areas and situations
- Deep Learning on ultrasound-based SSR as a pioneer

Method

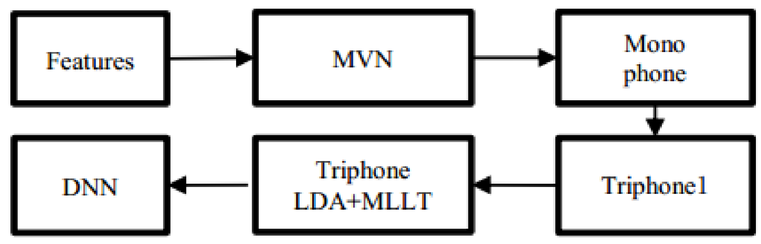
Data Acquisition

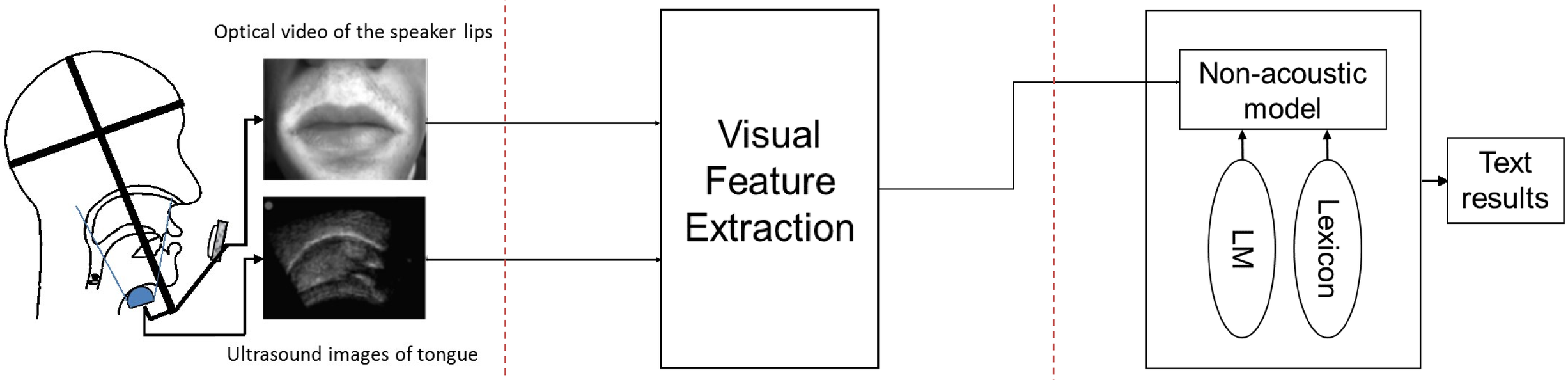
- a 128-element micro convex ultrasound probe
- a CMOS industrial camera
- Ultrasound focal depth: 7cm
- Frame of image streams: 60fps

Non-acoustic feature extraction - Autoencoder

- Pre-train a stack of Restricted Boltzmann Machines (RBMs) to obtain initial weights that are closed to global minima;
- "Unroll" the RBMs to build an encoder and a decoder;
- Perform back-propagation to fine-tune the model.

Silent Speech Recognition





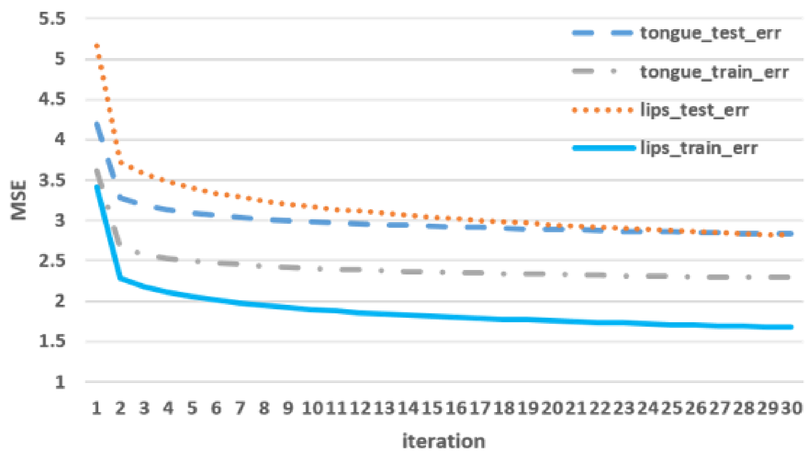
Experiment & Results

Speech corpora - TIMIT text

- 47 lists, each containing 50 sentences
- 5-8s for each utterance and contain 300-500 image frames
- 320*240 for ultrasound and 640*480 for video
- Test set: 100 sentences selected from WSJ0 5000-word

Feature extraction

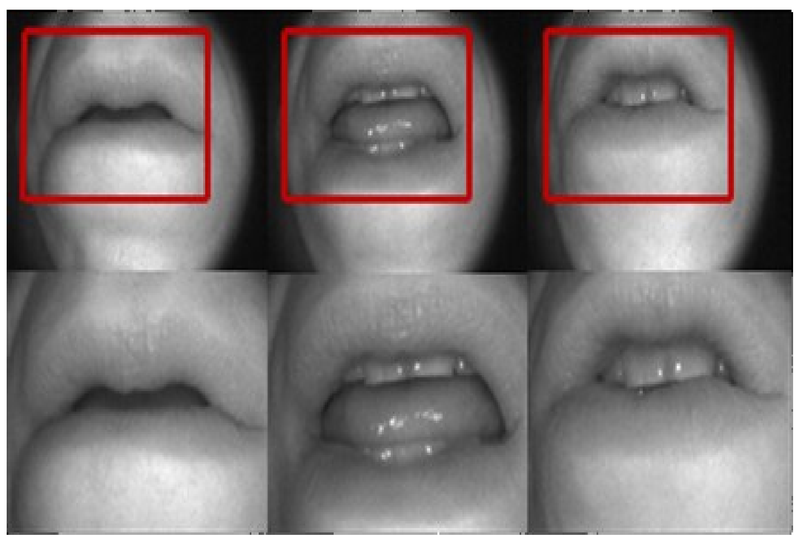
- Region of Interest (ROI) selection
- Bi-cubic interpolation image resizing
- Symmetric 3500(3000)-1000-500-250-30 model training
- Code layer combination of tongue and lip



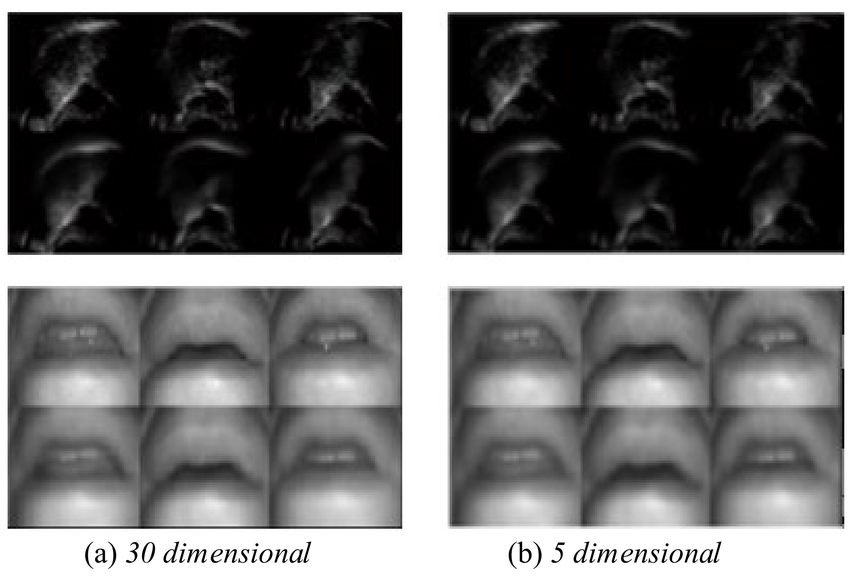
Mean Square Error (MSE) after each iteration of back-propagation.

WER of	DCT (%)			
Dimension	30	20	10	5
monophone	58.94	59.43	60.80	98.14
triphone1	59.24	55.33	55.62	98.44
triphone2b	43.50	44.87	46.73	100
DNN	36.75	37.15	39.49	99.90
WER of	Autoencoder (%)			
Dimension	30	20	10	5
monophone	69.79	65.10	63.44	65.20
triphone1	68.72	59.63	51.81	50.44
triphone2b	46.73	46.33	45.94	48.09
DNN	36.56	37.54	37.54	38.71

Results



ROI selection for raw images.



(a) *30 dimensional* (b) *5 dimensional*

Reconstructed images of tongue and lip

Conclusions

- Two types of features achieve similar WER performances
- Use of DNN-HMM is beneficial for video-based silent speech recognition for the first time
- DAE is able to create compact features that appear to retain saliency even at surprisingly low dimensionality

Acknowledgements

This research is supported by the National Nature Science Foundation of China (No. 61303109) and 985 Foundation from China's Ministry of Education (No. 060-0903071001).