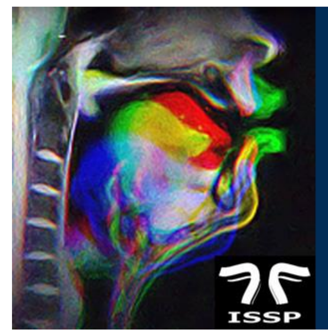
Ultrasound Tongue Gestural Sequence Classification Using Convolutional Auto-encoder and Recurrent Neural Network

Kele Xu[1,2], Tamás Gábor Csapó [3,4], Dawei Feng[1,2], Haibo Mi[1,2]

1. School of Computer, National University of Defense Technology
2. National Key Laboratory of Parallel and Distributed Processing, Changsha, China
3. Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
4. MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary



12th International Seminar on Speech Production

14 - 18 December 2020

1. Background

◆Speech is the vocalized form for the human-to-human communication, which is the most common and useful interface for human daily communication.

◆Traditional natural speech present some problems.

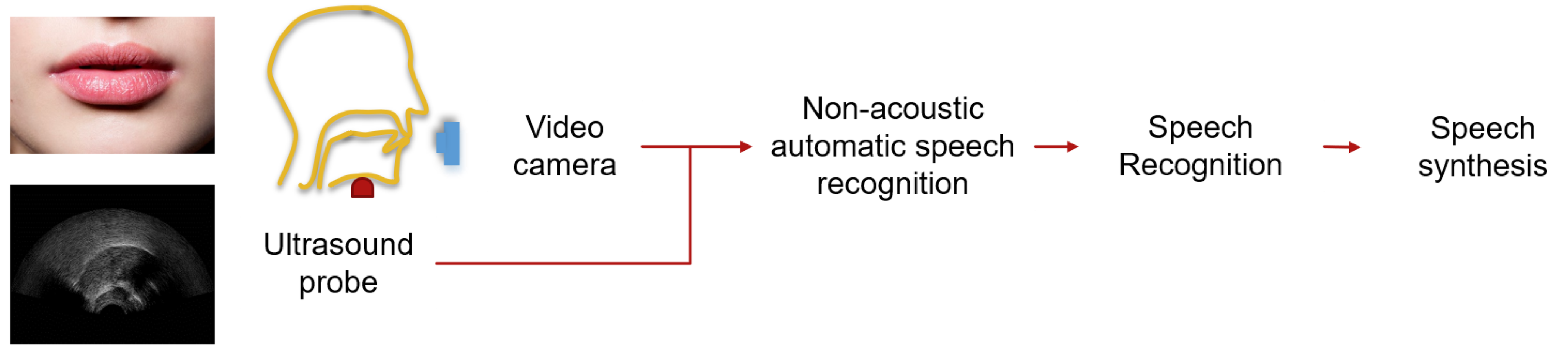
- ✓Speech is one-to-many modality, which can give rise to problems of users' interference and communication security;
- ✓If there is a high level of background noise, the quality of speech communication degrades rapidly;
- ✓The speech modality may be impossible when a speaker is incapacitated by illness or injury, either temporarily (laryngitis, flu, etc.) or permanently (cancer, laryngectomy, pulmonary insufficiency, accident, etc.);
- ✓Speech communication may be impossible when the parties involved do not share a common language.

2. Ultrasound-based Silent Speech Interface

◆"Silent Speech Interfaces" (SSI) is a system which uses the non-audible signals recorded during speech production to perform speech recognition and synthesis tasks.

◆Compared to other imaging modalities, ultrasound imaging is noninvasive, less expensive than other imaging systems, and convenient.

◆Ultrasound can track the tongue movement with relatively good spatial (e.g. 800×600 pixels) and temporal resolution (around 100 frame-per-second)

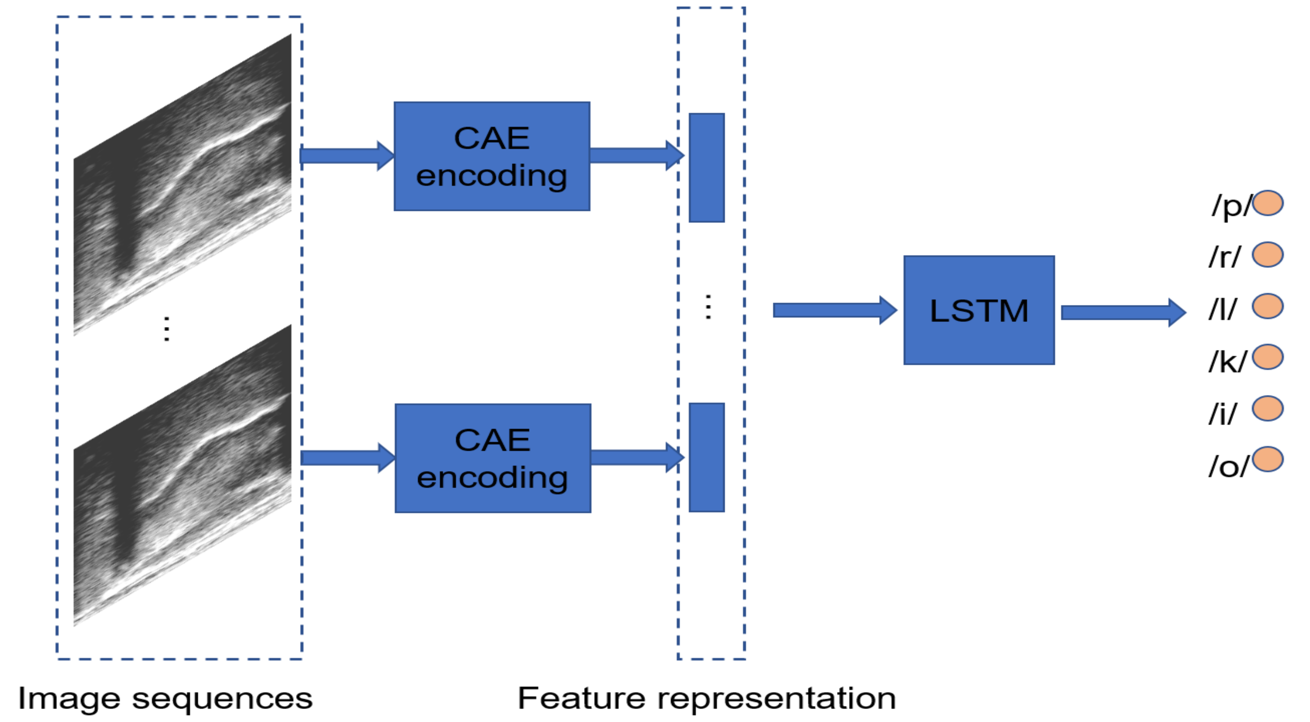


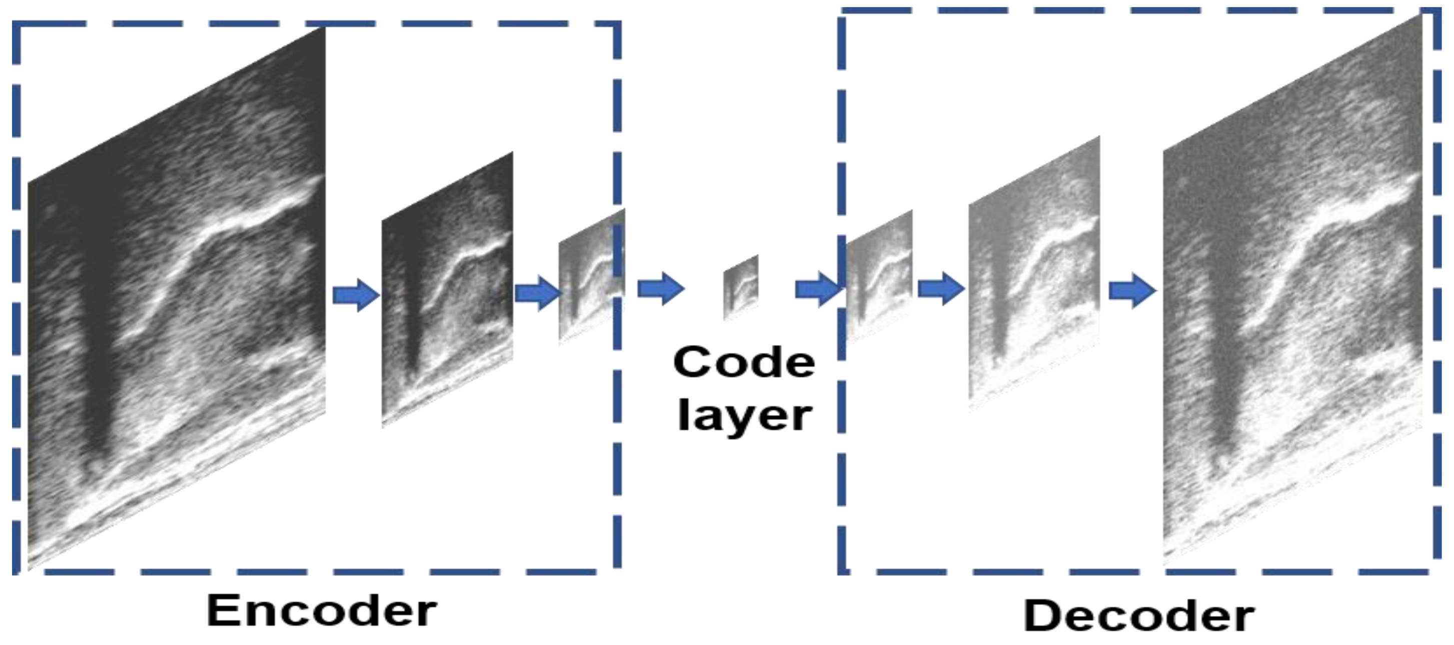
3. Methodology

◆Most of previous attempts aimed to distinguish between different ultrasound frames, this paper explores the sequence classification issue, whose length is variational.

◆In this paper, we present the approaches used for the sequence's classification task, which can be divided into two main parts:

1. Unsupervised feature extraction (or dimension reduction) using the convolutional autoencoder;
2. Supervised sequence classification using the recurrent neural network





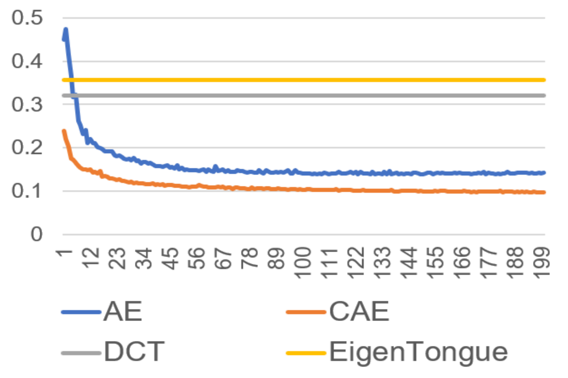
4. Experimental Results

➢We argue that the Convolutional Neural Network (CNN) may be more suitable to extract the visual information from ultrasound images

➢We employ all the single images in the training dataset to train the CAE. The employed CAE adopts the conventional architecture, in which the encoder consists of 3 convolutional layers and 3 max pooling layers.

➢Different length of the code layer are tested. The Mean Square Error (MSE) is used as a metric to assess reconstruction errors

.



Frame-wise MSE

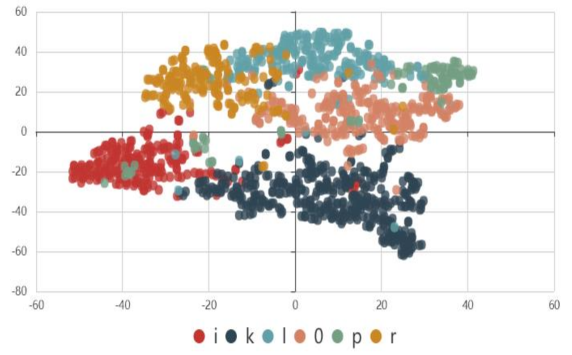

t-SNE visualization of the bottleneck features of CAE

Table 1: The accuracy of speaker-dependent sequence classification using B-mode ultrasound tongue imaging

Method	Accuracy for Female 1 (%)	Accuracy for Female 2(%)	Accuracy for Male 1(%)	(Mean + Standard variance)
EigenTongue+RNN	62.5	61.4	60.7	61.5±0.74
DCT+RNN	63.9	64.2	61.3	63.1±1.30
DAE+RNN	76.8	78.2	75.9	77.0±0.95
CAE+RNN	82.9	83.1	81.3	82.4±0.81

Table 2: The accuracy of speaker-independent sequence classification using B-mode ultrasound tongue imaging.

Method	Accuracy (%)
EigenTongue+RNN	37.3
DCT+RNN	45.8
DAE+RNN	75.5
CAE+RNN	78.2