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

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

[1]School of Computer, National University of Defense Technology, China
[2]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[3]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary

kelele.xu@gmail.com

Abstract

The interpretation of B-mode ultrasound tongue images is an important topic for speech production research. Unlike the previous attempts which aimed to distinguish between different ultrasound frames, this paper explores the sequence classification issue, whose length is variational. A convolutional auto-encoder is used as the unsupervised feature extraction method for the ultrasound tongue images, while a recurrent neural network is employed for the sequence classification. Both single-speaker and multi-speaker sequence classification experiments are conducted, and the results are reported in this paper.

1. Introduction

In the last decade, there has been an increasing interest in the analysis, recognition and articulatory movements. Systems that can perform the automatic articulatory-to-acoustic mapping are often referred to as "Silent Speech Interfaces" (SSI) [1]. Such an SSI system can be applied to facilitate the communication of the speech impaired (e.g. patients after laryngectomy), and in situations where the speech signal itself cannot be recorded (e.g. extremely noisy environments). Ultrasound tongue imaging (UTI) can capture tongue movements non-invasively during natural speech production, making it a potential choice for SSI. 2D ultrasound has been employed in phonetic research for more than 30 years. Typically, during UTI experiments, the ultrasound probe held under the chin will record a mid-sagittal view of the tongue, and those recordings are a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air. Compared to other articulatory capture techniques, UTI can track the tongue movement with relatively good spatial, (e.g. 800×600 pixels) and temporal resolution (around 100 frame-per-second). However, its main drawback is that: for phonetic research, the tongue contour has to be extracted first, for which still there is a lack of accurate automatic methods [2].

In our previous work, we were testing the automatic classification of midsagittal ultrasound tongue gestures using a convolutional neural network (CNN) [3] in an end-to-end manner. In the current paper, we extend our previous results through a more challenging tongue classification experiment. Unlike the previous attempt to classify single frames, the current study aims to classify the ultrasound tongue image sequences, whose length can be variational [4]. The paper is organized as follows: Section 2 describes the data acquisition, whereas Section 3 presents the methods used for the classification task, which can be divided into two parts: unsupervised learning using convolutional auto-encoder, and sequence classification using recurrent neural network. Section 4 presents the results and the conclusion is drawn in Section 5.

2. Data acquisition and pre-processing

The dataset used in this paper is similar to the data used in [3], but the preprocessing method is different as our goal is to classify image sequences of varying lengths. Same as in our previous study and as in, we focused on the image sequences classification for six phones (/p/, /r/, /l/, /k/, /i/, /o/). In synchrony with the speech signal, the tongue movements were recorded in midsagittal orientation using a "Micro" ultrasound system (Articulate Instruments Ltd.) with a 2-4 MHz / 64 element 20 mm radius convex ultrasound transducer at 82 frames per second. The datasets of labeled ultrasound tongue images sequences were relatively small, with only 320 labeled sequences for each subject (three subjects are used, one male subject and two female subjects), and a total of 960 sequences are employed for the experimental studies. The length of the sequence is variational, and the average length is 13 for all sequences. 50% of total sequences (480 sequences, 160 per subject) are randomly selected as the training data, while the remaining data (480 sequences) are used for evaluation.

3. Methodology

In this section, we present the approaches used for the sequences 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.

3.1. Convolutional autoencoder

In this paper, we explore the feasibility of using convolutional auto-encoder (CAE) [5] to extract the information in an unsupervised manner. In more detail, an CAE consists of two components: an encoder and a decoder (as shown in Figure 1). The encoder takes an image as an input and, by applying one or more parametrized nonlinear transformations, converts it into a new compact representation (code layer). The decoder takes this representation and learns to reconstruct the original frame as close as possible. The code layer also serves as a compressed representation of an ultrasound tongue image.

3.2. Recurrent neural network-based classification

Recurrent neural networks (RNNs) [6] are a type of artificial neural network designed to recognize patterns in sequences of

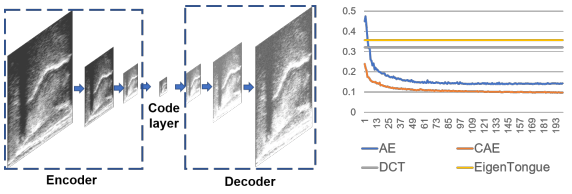

Figure 1: *Architecture of CAE and the reconstruction mean-square error (MSE) using different kinds of methods.*

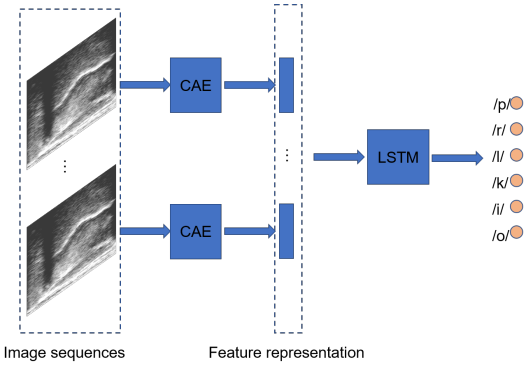

Figure 2: *Flowchart for the ultrasound tongue image sequences classification.*

data. RNN takes both time and sequence into account and makes use of the temporal information in the input sequences. During our experiments, the network had two hidden layers with 256 LSTM units each and one feed-forward layer with 512 rectified linear unit (ReLU).

4. Experiments and Results

Two experiments were conducted to evaluate the performances for the classification task. More precisely, the experiments include both single-speaker and multi-speaker tongue gesture sequence classification. For the purpose of comparison, different RNN were trained with the input feature vectors generated by one of the four unsupervised algorithms, EigenTongue, DCT, AE and CAE.

4.1. Single-speaker ultrasound tongue image sequences classification

The data used here come from three speakers (one male and two females), containing image sequences three speakers. For each speaker, sequences of feature vectors (each feature vector representing a single frame) are used to trained the RNN, and the and the held-out sequences from the same speaker are used to evaluate the models. The results of this experiment are given in Table 2. As can be seen from the table, the CAE-based method gives better classification performance.

4.2. Multi-speaker ultrasound tongue image sequences classification

Moreover, we also report the experiment results on multi-speaker phoneme sequence classification. As mentioned earlier, the augmented data used here comes from three speakers (Female 1, Female 2 and Male 1), consisting of 4,800 image sequences containing one of the following six phonemes.

Method	Female 1	Female 2	Male 1
ET+RNN	62.5%	61.4%	60.7%
DCT+RNN	63.9%	64.2%	61.3%
AE+RNN	76.8%	78.2%	75.9%
CAE+RNN	82.9%	83.1%	81.3%

Table 1: *The accuracy of single-speaker sequence classification using B-mode ultrasound tongue imaging. EigenTongue is abbreviated as ET in the table.*

Method	Accuracy (%)
EigenTongue+RNN	76.4
DCT+RNN	76.2
DAE+RNN	76.4
CAE+RNN	75.7

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

The goal is to classify recordings of the tongue gestural targets across three different speakers. Of the total, 4,800 sequences are used to train the RNN classifiers, and the remaining images (480 sequences, 160 from each speaker) is used to evaluate the models. It is worthwhile to notice that, here, both the sequences in the training data and evaluation data are selected from three different subjects.

5. Summary and conclusion

In this paper, we explore the classification of ultrasound tongue image sequences. Firstly, the unsupervised feature extraction method, CAE, has been explored to encode ultrasound tongue images, and an RNN is employed for the sequence classification task. Tongue gesture sequences classification are run on both single-speaker and multi-speaker cases. In our experiment settings, compared to other feature extraction schemes, the CAE-based features achieved superior classification performance on our dataset.

6. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] K. Xu, T. Gábor Csapó, P. Roussel, and B. Denby, "A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. EL154–EL160, may 2016.

[3] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL531–EL537, 2017.

[4] E. M. Juanpere and T. G. Csapó, "Ultrasound-based silent speech interface using convolutional and recurrent neural networks," *Acta Acustica united with Acustica*, vol. 105, no. 4, pp. 587–590, 2019.

[5] K. Xu, Y. Wu, and Z. Gao, "Ultrasound-based silent speech interface using sequential convolutional auto-encoder," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.