Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images

Kele Xu[a]
Department of Engineering, Université Pierre et Marie Curie, Paris 75005, France
kelele.xu@gmail.com

Pierre Roussel
Langevin Institute, ESPCI-ParisTech, Paris 75005, France
pierre.roussel@espci.fr

Tamás Gábor Csapó[b]
Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
csapot@tmit.bme.hu

Bruce Denby
Tianjin University, Tianjin, 300000 China
bruce.denby@upmc.fr

Abstract: Tongue gestural target classification is of great interest to researchers in the speech production field. Recently, deep convolutional neural networks (CNN) have shown superiority to standard feature extraction techniques in a variety of domains. In this letter, both CNN-based speaker-dependent and speaker-independent tongue gestural target classification experiments are conducted to classify tongue gestures during natural speech production. The CNN-based method achieves state-of-the-art performance, even though no pre-training of the CNN (with the exception of a data augmentation preprocessing) was carried out.

[DDO]
Date Received: October 29, 2016 **Date Accepted:** May 12, 2017

1. Introduction

During the past several years, there has been significant interest in "Silent Speech Interfaces" (SSI) (Denby *et al.*, 2010; Hueber *et al.*, 2010) that use non-audible signals recorded during speech production to perform speech recognition and synthesis tasks. A variety of techniques have been proposed for the SSI systems; in this paper, we focus on an ultrasound-based SSI (Hueber *et al.*, 2010). The speech recognition performance of such an SSI varies between speakers due to differences in ultrasound imaging quality (e.g., female subjects often image better than males; and younger subjects better than older ones; Stone, 2005). Speaker-independent recognition (or multi-speaker recognition) is a further challenge, and SSIs based on ultrasound are still in the experimental stage with regard to this task. The best way to extract robust, meaningful features for speech recognition from ultrasound tongue images is still an open question.

Previous work on feature extraction for ultrasound tongue images has used Principal Component Analysis (PCA) (Stone, 2005); Gabor filter banks (Berry *et al.*, 2010); and the Discrete Cosine Transform (DCT) (Cai *et al.*, 2011; Denby *et al.*, 2011; see also Berry, 2012). These feature representations, however, may lose meaningful information from the ultrasound tongue images during the dimension reduction process. Taking PCA as an example, the dimension reduction process does not take the spatial correlation property between interesting regions into account: the pixels of the image are represented as a simple vector, whereas the tongue is a muscle-activated organ, with coherent internal motion of the muscles. For this reason, PCA and its variants may result in the loss of important spatial correlation information during dimension reduction (Berry *et al.*, 2010). In fact, PCA and DCT (or their variants) are dimension reduction techniques, rather than the feature extraction techniques. Furthermore, recent

[a] Also at Langevin Institute, ESPCI-ParisTech Paris, 75005, France.
[b] Also at MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary.

progress in computer vision has demonstrated (He *et al.*, 2015) that feature extraction-based image classification may be more suitable for the image classification task.

Recently, deep convolutional neural networks (CNN) have demonstrated accuracy better than or equivalent to human performance in several different visual recognition tasks, such as: object detection (Ren *et al.*, 2015); image classification (Krizhevsky *et al.*, 2012); edge (contour) detection (Xie and Tu, 2015); etc. Rather than using selected features, CNN can learn a hierarchy of features, which may be useful for the purpose of tongue gesture classification. Since such hierarchical approaches are able to embed more complex correlations in their higher layers, including translations and distortions, the accuracy of CNN-based tongue gesture target classification task may be higher than other methods.

To the best of our knowledge, there has as yet been no attempt to use CNNs on image feature extraction for the tongue gestural target classification task, which, as mentioned, is of great importance for subject-independent SSI applications and for tongue motion modeling. In this paper, we present a first attempt to classify tongue gestural targets for a single subject using CNN. In addition, a speaker-independent tongue gestural target classification experiment is also conducted. All of the experiments obtain superior performance as compared to previous methods.

2. Data acquisition and data augmentation

Two female and one male adult Hungarian subjects with normal speaking abilities were recorded while producing sustained vowels (9 Hungarian vowels: /O, A:, E, e:, i:, o:, 2:, u:, y:/ - in SAMPA notation) and nonsense words (9 Hungarian vowels in combination with 5 consonants /p, t, k, l, r/, in 'VCVCV' sequences, altogether 45 nonsense words). The utterances were repeated three times. The tongue movement was recorded using a SonoSpeech ultrasound system (Articulate Instruments Ltd.) with a 2–4 MHz, 64 element, 20 mm radius convex ultrasound transducer, running at 82 fps. During the recordings, the transducer was fixed with respect to the head using an ultrasound stabilization headset (Articulate Instruments Ltd.). The speech signal was recorded with a Monacor ECM 100 condenser microphone placed at a distance of approximately 30 cm from the lips. The ultrasound and the audio signals were synchronized using a frame synchronization signal, using the Articulate Assistant Advanced software (Articulate Instruments Ltd.). On the basis of the speech recordings, phone boundaries were determined with a Hungarian speech recognizer (Mihajlik *et al.*, 2010) in forced alignment mode. Silences, rest tongue positions and swallowing were therefore not included in the data. The beginning and ending of each speech sound were discarded. Subsequently, the ultrasound frames corresponding to the target speech sounds were extracted as raw scan line data and converted to 500 × 500 pixel PNG images.

Since we want to classify tongue gestural targets, we selected only six homorganic tongue gestures: /p/, /r/, /l/, /k/, /i/, /o/ in SAMPA notation. The datasets of labeled ultrasound tongue images were relatively small, with only 4532 labeled frames extracted (for female 1, the number of labeled frames is 1587; for female 2, 1526; and for male 1, 1419). As the shortcomings of such small image datasets have been widely recognized, data augmentation is necessary to artificially enlarge the datasets, using label-preserving transformations, in order to reduce overfitting and increase the performance of the algorithms (Krizhevsky *et al.*, 2012).

Data augmentation is therefore performed, so as to mimic the variations induced by the ultrasound tongue image acquisition platform. Since there may be some small rotations or translations between the ultrasound probe and the chin of the subject, we apply translation, rotation, scaling, and addition of low-level speckle noise to the US images of the labeled dataset. The same proportion of each transformation type is applied to each original ultrasound tongue image. The range of the transformations undertaken is given in Table 1, while illustrations of the effects of the transformations are presented in Fig. 1 (rotation = +5°, rescaling = 1.1, translation increased to +15 pixels). After data augmentation, the total number of frames becomes 180 000 frames, that is, 10 000 frames per gesture (6) and per subject (3).

3. Deep convolutional neural network

CNNs are an important class of applications capable of learning representations, inspired by biological neural networks. A CNN consists of alternating convolution and sub-sampling (pooling) operations (LeCun *et al.*, 1998). Figure 2 gives a specific architecture of a CNN network for classifying tongue gestures. Each convolutional layer consists of a rectangular grid of neurons, with each neuron taking its inputs from a small region of the previous layer (He *et al.*, 2015). There may be several grids in each

Table 1. Data augmentation parameters.

Transformation type	Description
Rotation	−5°+5°
Rescaling	Random with scale factor between 1/1.1 and 1.1
Translation	Random with shift between −5 and 5 pixels
Adding noise	1%–5% multiplicative speckle noise

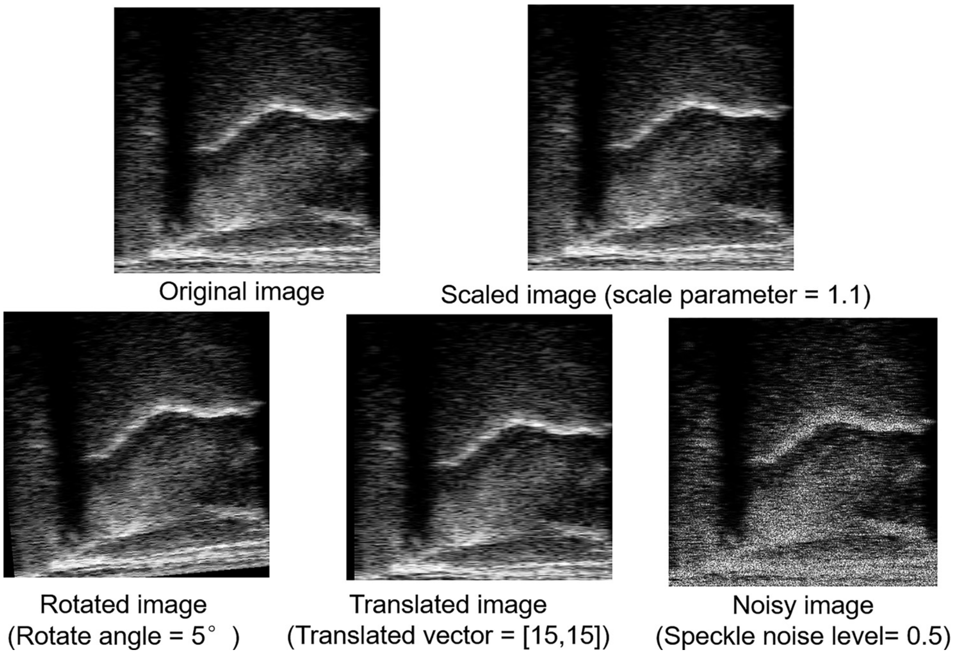
convolutional layer, performing potentially different filters parameterized by the neurons weights, which will be learned automatically by the CNN. Typically, the convolutional layers are interspersed with sub-sampling layers to reduce computation time and build up further spatial and configurational invariance. These sub-sampling layers are referred to as "pooling layers." Several ways can be used to achieve this pooling, for example, the average or the maximum of non-overlapping rectangles (Gu *et al.*, 2015). Finally, after several convolutional and pooling layers, a fully connected layer (or layers) is built (and trained) using the output of the previous layer as a compact feature to describe the tongue gesture in the ultrasound image. The features are then tiled to obtain a better representation of the original ultrasound tongue image, for example relating to the contour in the image, which is of great interest for researchers of clinical linguistics and phonetics.

Concretely, a feed-forward convolutional neural network can be viewed as a function f mapping data $\mathbf{x}$ as in Eq. (1):

$$f(\mathbf{x}) = f_L(\cdots f_2(f_1(\mathbf{x};\mathbf{w}_1);\mathbf{w}_2)\cdots,\mathbf{w}_L). \tag{1}$$

Each function f_l takes as input $\mathbf{x}_l$ (input image, or the feature map in the previous layers) and a parameter vector $\mathbf{w}_l$ (weight to be learned), where L is the number of layers in the neural network. Although the type and sequence of functions is usually selected, the parameters can be discriminatively learned from example data such that the resulting function f realizes a useful mapping.

Formally, in our CNN, each input $\mathbf{x}_i$ is a $M \times N \times C$ array, where M and N are the width and height of the image and, since B-mode ultrasound tongue images are gray-scale, $C = 1$. The weights of the convolutional layers and the fully connected layer are estimated by the minimization of a loss function based on the discrepancy between the desired output and the output of the CNN over all the examples in the training set:


Original image

Scaled image (scale parameter = 1.1)

Rotated image
(Rotate angle = 5°)

Translated image
(Translated vector = [15,15])

Noisy image
(Speckle noise level= 0.5)

Fig. 1. Sample frames after data augmentation. The ultrasound images are shown as boxes instead of wedges, as we have direct access to the ultrasound raw scan line data and did not want to lose information by conversion to a classic ultrasound wedge shape.

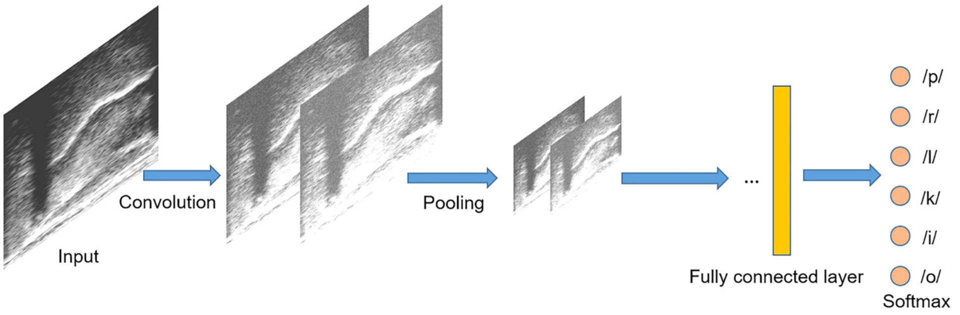


Fig. 2. (Color online) An example architecture of a Convolutional Neural Network for the classification task. The ellipsis denotes the subsequent convolutional and pooling layers.

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} loss\big(z_i, f(\mathbf{x}_i;\mathbf{w})\big), \quad (2)$$

where n is the number of images in the training set, and z_i is the desired output for input $\mathbf{x}_\mathrm{i}$ (Gu *et al.*, 2015). The network is optimized by back propagation and stochastic gradient descent (note that forward and backward propagation may differ, depending on the type of layer).

We have designed and tested several different CNN architectures. The architecture of the network used in our work consists of 18 layers having parameters, of which 11 are convolutional layers (some followed by max-pooling layers), and two are fully connected. The last layer is the Softmax output layer (Gu *et al.*, 2015), that converts the output vector of the fully connected layer to a vector of probabilities which sum up to 1, each probability corresponding to one class.

4. Experiments and results

Three experiments were conducted to evaluate the performances of CNNs and to compare them to some feature extraction methods (or dimension reduction approaches), EigenTongues (Hueber *et al.,* 2010), Discrete Cosine Transform (DCT) (Cai *et al.,* 2011), and Gabor-filter-banks (Berry *et al.*, 2010). More precisely, the experiments include speaker dependent and speaker independent tongue gesture classification of ultrasound images.

In more detail, for the Gabor-filter-banks-based representation, we follow the method given in Berry *et al.* (2010), where Gabor filters of various scales and orientations are applied to the images, and the energy images are concatenated to give a 1600 dimensional representation. For EigenTongue, the images are decomposed on a basis built from the Principal Component Analysis of a large set of images from which we selected the first 1600 principal components. DCT is closely related to the Discrete Fourier Transform and consists of the decomposition of the images onto a basis of cosines. To make a quantitative comparison, the dimension of the DCT representation is also truncated to 1600. This choice of dimensionalities placed each method as close as possible to its optimal performance regime, while at the same time striving to balance dimensionality across methods, and maintain computational tractability. In contrast to many previous studies (Cai *et al.,* 2011; Hueber *et al.,* 2010), we use the full-size images as initial inputs, rather than re-dimensioned regions of interest. The 1600 coefficients obtained by the different previous methods are used as the inputs to the classifiers trained using the gradient boosting machine based classification method (Friedman, 2001). eXtreme Gradient Boosting method (XGBoost), a tree boosting machine based classification method (Chen and Guestrin, 2016), was selected as the benchmark because, compared to other approaches (such as Support Vector Machine, Random Forest), XGBoost provided better classification performances in our experiment, and the power of XGBoost has furthermore been validated on several public machine learning challenges (Chen and Guestrin, 2016). We use the default hyper parameters for XGBoost, with six classes, and the maximum depth of the XGBoost set to six to prevent overfitting. As to the CNN implementation, we used the tool "MXNet" (Chen *et al.,* 2015), to implement our tongue gesture classification task; the description of the CNN architecture used in our experiment is given in Sec. 3.

4.1 Speaker dependent tongue gesture classification using ultrasound tongue images

The data used here comes from three speakers (one male and two females), containing 60 000 images per speaker labeled as exhibiting one of the following phonemes: /p/, /r/,

Table 2. The accuracy of speaker-dependent tongue gesture classification using ultrasound images.

Method	Accuracy for Female 1 (%)	Accuracy for Female 2 (%)	Accuracy for Male 1 (%)	(Mean + Standard variance)
EigenTongue + XGBoost (without data augmentation)	69.3	71.4	56.7	65.8 ± 8.0
EigenTongue + XGBoost (with data augmentation)	70.2	72.3	59.4	67.3 ± 6.9
DCT+ XGBoost (without data augmentation)	67.1	74.5	65.9	69.2 ± 4.7
DCT+ XGBoost (with data augmentation)	69.4	75.6	67.6	70.9 ± 4.2
Gabor-filter-banks+ XGBoost (without data augmentation)	92.7	95.3	91.4	93.1 ± 2.0
Gabor-filter-banks+ XGBoost (with data augmentation)	94.9	96.1	92.3	94.4 ± 1.9
CNN	**97.3**	**98.5**	**95.6**	**97.1 ± 1.5**

/l/, /k/, /i/, /o/ (in SAMPA notation) in equal proportion. It is worthwhile to note that these 60 000 images were produced by augmenting 4532 images.

The approach used for quantitative comparison is given as follows; for each speaker, 80% of the 60 000 frames are randomly selected and used to train the multi-class classifiers, and the remaining held-out images used to evaluate the models. The results of this experiment are given in Table 2. As can be seen from the table, the CNN-based method gives superior performance in comparison with the other classifiers. Also, it is worthwhile to note that data augmentation can improve the accuracy of the different methods.

The results suggest that valuable information may be lost using EigenTongue or DCT. Gabor-filter-banks and CNNs obtain higher classification performances in our experiment. These results must be evaluated with some care, however, since most of the images in the test set differ from those of the training set only via the augmentation process. Nevertheless, the improvements obtained in the non-CNN methods using augmentation are small, suggesting that the procedure neither introduces a substantial bias in the case of CNN.

4.2 Speaker independent tongue gesture classification using the ultrasound tongue image

For experienced clinical phoneticians, it takes several months to learn to distinguish between different tongue gestural targets using ultrasound. A subject independent silent speech interface is a great challenge. In this section, we conduct an experiment on multispeaker phoneme classification. As mentioned earlier, the data used here comes from three speakers (female 1, female 2 and male 1), consisting of 180 000 images containing one of the following six phonemes: /p/, /r/, /l/, /k/, /i/, /o/ (in SAMPA notation). The goal is to classify the tongue gestural target recorded from several different speakers. Of the total, 120 000 frames are used to train the classifiers (full data of two speakers: female 1 and male 1) and the remaining images (female 2) used to evaluate the models (full data of one speaker). The results from the different classification methods are given in Table 3. As can be seen from the table, with data augmentation, the CNN-based method gives better performance (76.1%) compared to the classifiers (56.4%, 56.9%, 62.5%) using various dimension reduction approaches. We note that in contrast to the speaker-dependent case, here the data augmentation procedure introduces a substantial improvement in the EigenTongue and DCT results. One may hypothesize that the presence of augmented images from two different speakers in the training set has

Table 3. The accuracy of speaker independent tongue gesture classification using ultrasound images.

Method	Accuracy without data augmentation (%)	Accuracy with data augmentation (%)
EigenTongue + XGBoost	27.5	56.4
DCT+ XGBoost	34.6	56.9
Gabor-filter-banks+ XGBoost	60.2	62.5
CNN	/	**76.1**

led to a network that is better able to generalize gestural similarity. The better performance of the Gabor filter-banks compared to EigenTongues and DCT, on the other hand, which was also the case in the speaker-dependent experiment, may be an indication that Gabor filters inherently possess some of the properties learned by the CNN kernels during the training process.

5. Conclusion

In this letter, we explored the potential applications of the CNN in ultrasound imaging-based tongue gestural target classification task. As a proof of concept to employ the CNN in this context, both subject dependent and subject independent tongue gesture target classification cases were explored. Compared to previous approaches, the CNN-based classifier achieved superior performance. Furthermore, on the very important task of subject independent tongue gesture target classification, CNNs obtained 76.1% accuracy. We believe that with better optimization on the architecture of the CNN or by applying ensemble methods over different models, the accuracy can be further improved. The potential application of this method seems promising, since it may be useful to improve silent speech recognition performance, using the CNN-based method.

For future work, we would like to design a more complex tongue gesture target classification experiment on a larger database, and with more phonemes. It will also be worthwhile to expand the application of CNN to contour extraction for ultrasound tongue images by applying regression and classification simultaneously in the CNN training procedure.

Acknowledgments

This work is funded by the European Commission via the i-Treasures project (FP7-ICT-2011-9-600676-i-Treasures). We would also like to thank the Lingual Articulation Research Group (Hungarian Academy of Sciences), for providing the recordings with the SonoSpeech system.

References and links

Berry, J. (**2012**). “Machine learning methods for articulatory data,” Ph.D. dissertation, University of Arizona, Tucson, AZ.

Berry, J., Archangeli, D., and Fasel, I. (**2010**). “Automatic classification of tongue gestures in ultrasound images,” in *Proceedings of 12th Conference on Laboratory Phonology*. Albuquerque, NM.

Cai, J., Denby, B., Roussel-Ragot, P., Dreyfus, G., and Crevier-Buchman, L. (**2011**). “Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model,” in *Interspeech*, pp. 1005–1008.

Chen, T., and Guestrin, C. (**2016**). “XGBoost: A scalable tree boosting system,” in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. (**2015**). “MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems,” in *Neural Information Processing Systems, Workshop on Machine Learning Systems*, Barcelona, Spain (MIT Press, Cambridge, MA).

Denby, B., Cai, J., Hueber, T., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Chollet, G., Manitsaris, S., and Stone, M. (**2011**). “Towards a practical silent speech interface based on vocal tract imaging,” in *International Seminar on Speech Production*, Montreal, Canada, pp. 89–94.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., and Brumberg, J. (**2010**). “Silent speech interface,” Speech Commun. **52**(4), 270–287.

Friedman, J. (**2001**). “Greedy function approximation: A gradient boosting machine,” Ann. Stat. **29**(5), 1189–1232.

Gu, J., Wang, Z. K., Ma, L., and Shahroudy, A. (**2015**). “Recent advances in convolutional neural networks,” arXiv:preprint, pp. 1512.07108.

He, K., Zhang, X., and Ren, S. S. (**2015**). “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE International Conference on Computer Vision*.

Hueber, T., Benaroya, E. L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (**2010**). “Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips,” Speech Commun. **52**(4), 288–300.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (**2012**). “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, Lake Tahoe, CA (MIT Press), pp. 1097–1105.

LeCun, Y., Bottou, L., Boggio, Y., and Haffner, P. (**1998**). “Gradient-based learning applied to document recognition,” Proc. IEEE **86**(11), 2278–2324.

Mihajlik, P., Tuske, Z., Tarján, B., Németh, B., and Fegyó, T. (**2010**). “Improved recognition of spontaneous Hungarian speech-Morphological and acoustic modeling techniques for a less resourced task,” IEEE Trans. Audio Speech Language Processing **18**(6), 1588–1600.

Ren, S., He, K., Girshick, R., and Sun, J. (**2015**). "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, Montreal, Canada (MIT Press), pp. 91–99.

Stone, M. (**2005**). "A guide to analysing tongue motion from ultrasound images," Clin. Ling. Phonetics **19**(6-7), 455–501.

Xie, S., and Tu, Z. (**2015**). "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1395–1403.