IMPROVING THE CLASSIFICATION OF PHONETIC SEGMENTS FROM RAW ULTRASOUND USING SELF-SUPERVISED LEARNING AND HARD EXAMPLE MINING

Yunsheng Xiong[1,2], *Kele Xu*[1,2*], *Meng Jiang*[3], *Liang Cheng*[4], *Yong Dou*[1,2], *Jinjia Wang*[5]

[1] National University of Defense Technology, Changsha, China
[2] National Key Lab of Parallel and Distributed Processing, Changsha, China
[3] Changsha Medical University, Changsha, China
[4] Xiangya Hospital, Central South University, Changsha, China
[5] Yanshan University, Qinhuangdao, China

ABSTRACT

Ultrasound tongue imaging is an attractive way for speech production study as it provides an effective visualization for the vocal tract. Automatic classification of phonetic segments (tongue shapes) from raw ultrasound data is vital for further interpretation. Recently, deep learning-based approaches have been adopted in this task, which required a large-scale annotated dataset for the training, and it is not easy to be obtained in practical settings. Moreover, the data may contain many hard examples for the classification task, due to contamination of speckle noise. In this paper, we aim to address these issues: firstly, self-supervised learning is adopted to utilize the unlabeled datasets and extract the features without any human annotations; secondly, hard example mining is applied to imitate the learning path of the clinical linguists. To empirically demonstrate the proposed method's effectiveness, we evaluate the method on the Ultrax Typically Developing dataset (UXTD) under different scenarios. The results show that the proposed method outperforms the other methods and achieves superior performance. To better promote the study in this field, we release our code publicly at [1].

Index Terms— Ultrasound tongue imaging, Automatic classification, Self-supervised learning, Hard example mining.

1. INTRODUCTION

During natural speech production, B-mode ultrasound tongue imaging (UTI) is one of the appealing ways for the vocal tract modeling [1], as it can capture the tongue movement at a high framerate (60Hz or higher). Moreover, UTI does not expose the speakers to radiation and the machines are of lower costs presently [2, 3, 4, 5]. Automatic classification of tongue shapes from raw ultrasound can facilitate the understanding of speech production, which has attracted increasing attention during last years [2, 6]. Previous attempts employed the principal component analysis (PCA) [7], discrete cosine transform (DCT) [8], autoencoder [9] for the feature learning in the UTI, leveraging the unsupervised manner. Since the revolution of deep learning, supervised convolutional neural network (CNN) has been successfully applied in UTI processing [2, 10, 11]. Generally speaking, supervised learning often requires a large number of labeled examples [12], which is difficult to obtain in practical settings. How to utilize the massive unlabeled UTI data for the supervised learning, was under-explored in the previous studies. Recently, self-supervised learning (SSL) has made great progress, which using unlabeled data to learn a representation without human annotation. In this paper, we explore the SSL for the pre-training of CNNs, which can be further deployed for UTI interpretation.

Moreover, UTI has low signal-to-noise-ratio (SNR) and high-level speckle-noise [5, 1, 13]. Thus, there exist many hard examples in the datasets, which are easily confused and difficult to classify, even by experienced clinical linguists. Here, we aim to animate the learning path of the clinical linguist and pick out difficult examples firstly and incrementally put them back into the network for training, so as to increase the network's ability to identify difficult examples gradually. Combining with the SSL and hard example mining, we can greatly improve the classification performances for the phonetic segments from the raw ultrasound.

In the following section, we elaborate on our method. In Sec. 3, we present our extensive experimental results. Sec. 4 summarize the methods of this paper and have an exploratory discussion.

2. METHODOLOGY

Figure 1 shows the overall framework of our proposed method, which consists of two main stages: pre-training of the CNN using SSL, and supervised learning for the fine-tuning. For the pre-training, three widely-used SSL approaches were explored to initialize the CNNs without hu-

[*] Corresponding author.

[1] https://github.com/randallfuhao/ultrasound-selfsup

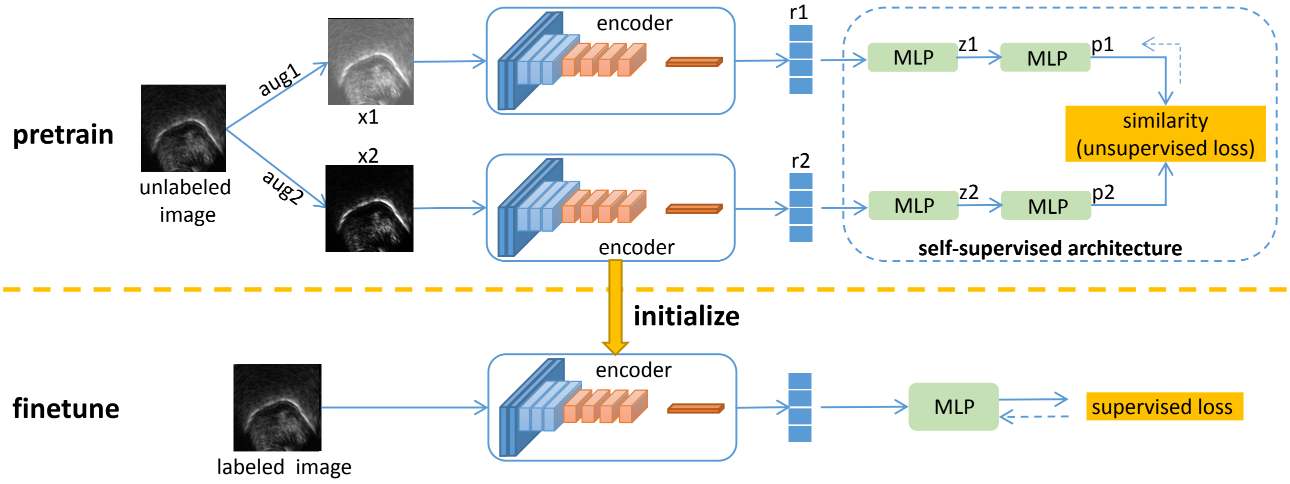

Fig. 1. Basic pipeline of our proposed method for the UTI classification task.

man annotation. After the pre-training, we use the labeled dataset to fine-tune the initialized CNNs for the raw ultrasound segment classification tasks. Compared to the method without pre-training, our method can significantly improve classification performance. We will explain the components in more detail subsequently.

2.1. Self-supervised learning

SSL mainly uses auxiliary tasks (pretext tasks) to predict its own supervised information from large-scale unsupervised data. By training the network with this constructed supervisory information, valuable representations for downstream tasks can be learned. In this paper, we employ contrastive-based SSL architectures, which can learn representation by comparing the similarity or dissimilarity of two images. We strive to learn valuable representations in UTI using three state-of-the-art contrastive SSL methods: SimCLR [14], BYOL [15], and SimSiam [16]. As shown in Fig 1, they are based on the siamese network. Specifically, the original image is transformed twice to generate two views, and then the two views are input into the same CNN to obtain two representations. In the training phase, the two representations are fed into the self-supervised architecture to obtain a loss based on the similarity or dissimilarity between the representations, and then carry out back-propagation to adjust CNN network parameters. In the inference stage, the self-supervised architecture is abandoned, and the representation of the original image obtained after CNN is used for downstream tasks, such as classification, detection, segmentation, etc. The difference between the three methods mainly lies in self-supervised architecture.

SimCLR utilizes the similarity and dissimilarity between examples. In a mini-batch, SimCLR builds a large number of positive and negative example pairs. It calculates the distance between all example pairs. SimCLR needs to carefully build negative example pairs and usually rely on large batch size and appropriate data augmentation methods. BYOL eliminates the dependence on negative example pairs through two asymmetric neural networks. On the basis of BYOL, Simsiam further removed the momentum encoder. The difference between Simsiam and BYOL is that: Simsiam does not have two sets of network parameters. Due to the elegant architecture and the elimination of negative sample pairs, SimSiam has a simple loss function and it is easy to train the model. As shown in Formula 1, cosine distance is used to represent the similarity between the projection p_1 and the prediction z_2 (refer to Fig. 1):

$$D(p_1, z_2) = -\frac{p_1}{\|p_1\|} \frac{z_2}{\|z_2\|} \quad (1)$$

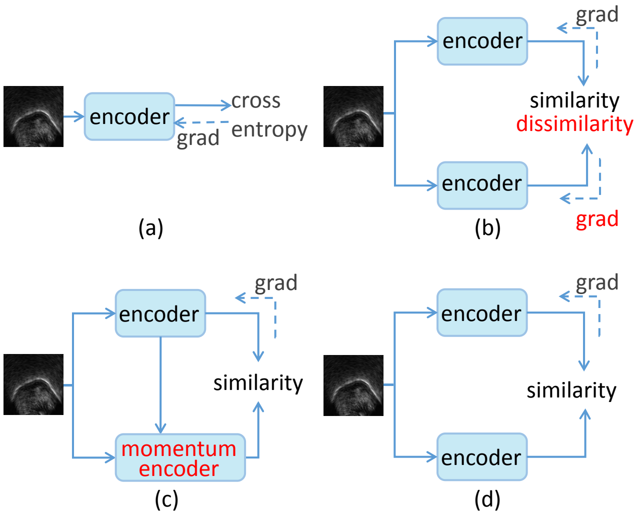

Fig. 2. **Comparison of different architectures.** (a) Supervised learning; (b) SimCLR; (C) BYOL; (d) SimSiam.

Symmetrically, the total loss function is as follows:

$$L = \frac{1}{2}D(p_1, z_2) + \frac{1}{2}D(p_2, z_1) \qquad (2)$$

2.2. Hard example mining

Ultrasound tongue image sequence is noisy and contains unrelated high-contrast edges [1], and individual differences are great, especially for those with pronunciation disorders. Therefore, there are many examples that are difficult to classify. Here, we aim to animate the learning path of the clinical linguist, and we pick out difficult examples firstly and incrementally put them back into the network for training, so as to gradually increase the network's ability to identify difficult examples. Specifically, the hard example mining method is used to dynamically adjust the weight of examples in the loss function according to their difficulty degree. In general, we use cross-entropy as the loss function of classification, which is formulated as follows:

$$H(p_g, p_d) = -\sum_{i}^{n} p_g(i) \log p_d(i) \qquad (3)$$

where p_g is the ground truth, and p_d is the prediction, n is the number of categories.

In this paper, we employ the focal loss [17] for the UTI segment classification task, as it is an improved version of cross-entropy. By reducing the weight of samples that are easy to classify, the model focuses more on samples that are difficult to classify during training. As shown in the Equation. 4, the focal loss can adjust the weight of the example in the loss function adaptive according to the prediction. The weight coefficient is $1 - P_d$, so the larger p_d is, the smaller the weight is.

$$FL(p_g, p_d) = -\sum_{i}^{n} (1 - p_g(i) * p_d(i)))^{\gamma} \log p_d(i) \qquad (4)$$

where γ is a hyper-parameter used to adjust weights. As shown in Equation 5, The total loss function of the model is the weighted sum of cross entropy and focal loss, where α is the weighting factor.

$$loss_t = H(p_g, p_d) + \alpha FL(p_g, p_d) \qquad (5)$$

2.3. Network Architecture

In our practical implementation for classification network, we extract the image features based on 2D convolution, and each convolution layer is followed by a 2D Batch Normalization layer and the activation layer ReLU, thus forming a so-called 2D convolution group. Four such 2D convolution groups are stacked to form a feature extractor. The classifier consists of two dense connected layers with a 1D Batch Normalization layer inserted between them. As pointed out in [2], leveraging the mean of all ultrasound frames as a second channel of CNN can significantly improve results across all scenarios. Here we utilize Batch Normalization (BN) [18] as an alternative. [2] uses the average of all the images of a speaker, while BN exploits the average of the current batch of images dynamically. Compared with the method using the mean of all frames, our approach can be conducted using the end-to-end manner.

3. EXPERIMENTAL RESULTS

In this part, we aim to describe the performance of our proposed method, with comparison to previous approaches.

3.1. Experiment settings

For a comprehensive assessment and fair comparison, we organized the data according to reference [2]. We use the images of the UXTD from the publicly available UltraSuite repository [19]. We classify utterances into four categories corresponding to distinct places of articulation: (1) bilabial and labiodental phones (e.g. /p/, /b/, /v/...); (2) dental, alveolar, and postalveolar phones (e.g. /th/, /d/, /t/, /z/...); (3) velar phones (e.g. /k/, /g/...); (4) alveolar approximant /r/. By using the force-aligned phone boundaries, we extract the middle frame of every available utterance and get a dataset of about 10700 examples, which are divided into disjoint train set, valid set, and test set. For the dataset of self-supervised learning, we utilize all frames of every available utterance as unlabeled examples and get about 209,130 images.

Following [2], we evaluated our approach in four scenarios: speaker-dependent, multi-speaker, speaker-independent, and speaker-adapted. we trained the model with the PyTorch framework and rely on a single GPU (2080Ti). For SSL, we utilized Adam as the optimizer, set the batch size N to 4096, and the learning rate to 0.001. For fine-tuning, SGD is utilized as the optimizer, the batch size N is set to 64, and the learning rate is set to 0.0001. The coefficient in the loss function is set to 2 by default unless otherwise specified.

3.2. Results comparison

The classification performance is given in Table 1. In the table, for each of the four evaluation scenarios, "/M" stands for "with speaker mean", "dep" refers to the "dependent" scenario, "m-spk" refers to the "multi-speaker" scenario, "indep" refers to the "independent" scenario, and "adapted" refers to the "adapted" scenario. As can be seen from the table, in the "dependent" scenario, our method exceeds other methods by at least 6.98%; in the "multi-speaker" scenario, our method outperforms the others by more than 4.65%; in the "independent" scenario, our method is superior to all other methods except "CNN raw with speaker mean"; while in the "adapted" scenario, our method outperforms the others by more than 1.77%. The reason why our method is weaker than "CNN

Table 1. Results comparison.

Method	dep	m-spk	indep	adapted
DNN raw	62.15%	69.62%	54.15%	69.26%
DNN raw /M	- %	71.61%	60.52%	70.31%
DNN PCA	57.78%	66.30%	55.14%	68.37%
DNN PCA /M	- %	67.17%	55.76%	68.02%
DNN DCT	68.38%	71.91%	55.36%	67.76%
DNN DCT /M	- %	72.28%	60.19%	69.41%
CNN raw	66.56%	74.70%	59.42%	72.67%
CNN raw /M	-	74.81%	**67.00%**	71.30%
Our method	**75.36%**	**79.46%**	62.01%	**74.44%**

raw with speaker mean" in the "independent" scenario is that the average of Batch Normalization is limited to a batch of images. Specifically, in the other three scenarios, Batch Normalization can encounter the speaker in the test set during training (although the image in the test set will not appear in the training set, the speaker in the test set will appear in the training set). So Batch Normalization can play a role similar to 'speaker mean' during testing. On the contrary, when encountering the 'independent' scenario, Batch Normalization has not seen the speaker in the test set during training, and the image mean of the speaker cannot be calculated correctly during testing.

3.3. Ablation study

Since our method combines the advantages of SSL and hard example mining, we will analyze how each one affects the performance through some ablation experiments in the following.

3.3.1. *Ablation study of SSL*

We compared different SSLs in the multi-speaker scenario, and the results are shown in Fig. 3. SimCLR does not improve the performance of downstream classification tasks. The model pre-trained by using BYOL brings 0.92% accuracy improvement for the downstream task. Due to the elegant architecture and easy-to-train characteristics of SimSiam, the pre-trained model by SimSiam performs best on the downstream task, with an accuracy gain of 1.29%.

3.3.2. *Ablation study of hard example mining*

The coefficient α in the loss function will change the importance of difficult examples in the process of parameter adjustment and have a significant impact on performance. In this part, we compared different α values of the loss function in the multi-speaker scenario, and the results are shown in Fig. 3. When $\alpha = 0$, it means that hard example mining is not adopted. It can be seen that hard example mining brings significant performance improvement. Even if α is set to 0.2, an accuracy gain of 0.51% can be achieved. When $\alpha = 2$, the classification performance reaches the best with the accuracy of 79.46%.

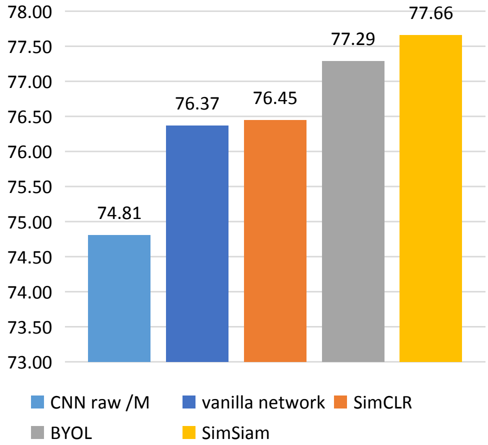

Fig. 3. **Comparison of different self-supervised learning methods.**

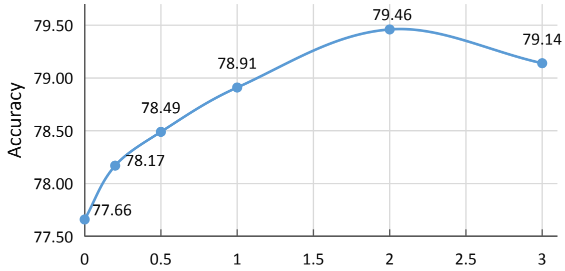

Fig. 4. **Comparison of different α values in the loss function.**

4. CONCLUSION

We proposed an improved classification method for the phonetic segments from the raw UTI. Combining self-supervised learning and hard example-mining, our method can significantly improve classification performance. Our approach can be conducted using the end-to-end manner and the experimental results demonstrate that the proposed method can provide better accuracy. For our future work, we would like to explore the use of self-supervised learning to improve the performance of contour extraction in the UTI and ultrasound-based silent speech interface.

5. ACKNOWLEDGEMENT

This work is supported by the major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" project 2020AAA0104803.

6. REFERENCES

[1] Min Li, Chandra Kambhamettu, and Maureen Stone, "Automatic contour tracking in ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 545–554, 2005.

[2] Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, and Steve Renals, "Speaker-independent classification of phonetic segments from raw ultrasound in child speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1328–1332.

[3] Joanne Cleland, James Scobbie, Zoe Roxburgh, Cornelia Heyde, and AA Wrench, "Ultraphonix: using ultrasound visual biofeedback to teach children with special speech sound disorders new articulations," in *7th International Conference on Speech Motor Control*, 2017.

[4] Joanne Cleland, James M Scobbie, Zoe Roxburgh, Cornelia Heyde, and Alan Wrench, "Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 2, pp. 229–246, 2019.

[5] Maureen Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[6] Kele Xu, Pierre Roussel, Tamás Gábor Csapó, and Bruce 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.

[7] Thomas Hueber, Guido Aversano, Gérard Chollet, Bruce Denby, Gérard Dreyfus, Yacine Oussar, Pierre Roussel, and Maureen Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 1, pp. I–1245.

[8] Jun Cai, Bruce Denby, Pierre Roussel-Ragot, Gérard Dreyfus, and Lise Crevier-Buchman, "Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model.," in *Interspeech*, 2011, pp. 1005–1008.

[9] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby, "Updating the silent speech challenge benchmark with deep learning," *Speech Communication*, vol. 98, pp. 42–50, 2018.

[10] M Hamed Mozaffari, Md Ratul, Aminur Rab, and Won-Sook Lee, "Irisnet: Deep learning for automatic and real-time tongue contour tracking in ultrasound video data using peripheral vision," *arXiv preprint arXiv:1911.03972*, 2019.

[11] Eric Tatulli and Thomas Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2971–2975.

[12] Jian Zhu, Will Styler, and Ian Calloway, "A cnn-based tool for automatic tongue contour tracking in ultrasound images," *arXiv preprint arXiv:1907.10210*, 2019.

[13] Jeff Berry and Ian Fasel, "Dynamics of tongue gestures extracted automatically from ultrasound," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 557–560.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[16] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[18] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[19] Aciel Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James Scobbie, and Alan Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," *Interspeech*, 2018.