EMG-based Thai Tone Recognition using Convolutional Neural Networks and Spectrograms

Salita Sombatsiri, Jaehoon Yu, Yoshinori Takeuchi and Masaharu Imai, *Member, IEEE*

***Abstract*—This study proposes an electromyography (EMG)-based Thai tone recognition using convolutional neural networks and spectrograms. Temporal variation and signal instability problems are mitigated by extracted invariant features and dropout technique, respectively. The proposed method can classify five Thai tones from EMG of a regular speaker with 68.27% of accuracy.**

I. INTRODUCTION

The electromyography (EMG)-based tone recognition is crucial in discriminating words in tonal languages for assisting laryngectomy patients. The challenges are shifting and scaling temporal variation, signal instability, and the quality of EMG. EMG-based Thai tone classification was studied using three-step neural networks [1], but it depended on word segmentation. This study proposes an EMG-based Thai tone recognition using convolutional neural networks (CNN) to learn invariant features and handle deformed EMG.

II. SURFACE EMG-BASED THAI TONE RECOGNITION

A. Thai Tone Recognition Methodology

The five Thai tones are recognized from the EMG activity [2]. The EMG is recorded from a regular male speaker uttering 20 sets of 133 Thai words at 1,000 Hz using surface electrodes placed as shown in Fig. 1(a). Informed consent was obtained from the speaker. The 60-Hz power line noises are removed. Seven spectrograms containing 50 frames of 51 zero-mean and unit-variance channel-wise normalized coefficients from 7-channel EMG of a word are inputted to CNN.

A CNN can mitigate the effect of temporal variation by convolutional layers and pooling layers. Convolutional layer's kernels are shared across input spectrograms to learn the invariant features of local pitch changes despite temporal shift. The response at position i, j of kernel k is computed by Eq. 1.

$$z_{i,j}^{(k)} = f(b^{(k)} + \sum_{c=0}^{C-1}\sum_{m=0}^{K-1}\sum_{n=0}^{K-1} x_{i+m,j+n}^{(c)} w_{m,n}^{(k,c)}) \quad (1)$$

$b^{(k)}$, $x_{i+m,j+n}^{(c)}$ and $w_{m,n}^{(k,c)}$ represents bias vector of kernel k, $i+m$, $j+n$ input feature map of channel c, and m, n weight of c weight matrix of kernel k, respectively. Furthermore, max pooling layer minimizes the impact of a slight shift or scaling.

B. Primary Experimental Results and Discussion

A CNN with two sets of a 50-kernel 5×5-size convolutional layer and a pooling layer, and two 25-node fully-connected layers is used to confirm the extracted invariant features. The convolutional layer yields similar responses when recognizing a normal, a shifted, and a scaled mid tone utterances as shown in Fig. 1(b) because invariant features were extracted. Fig. 1(c) shows activation states of five words of the same tone horizontally. These states represent tone-specific features learned from invariant features. Hence, these utterances are correctly recognized as mid tone by the CNN.

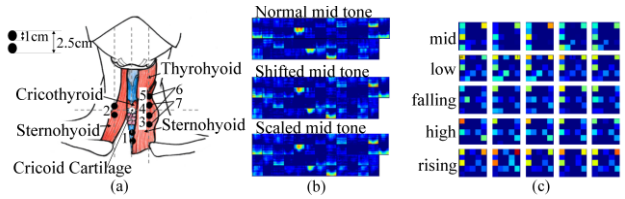

Figure 1. Electrode position and invariant features: (a) electrode positions in data acquisition (b) responses from the second convolutional layer (c) activation states from the last fully-connected layer

TABLE I. CROSS-VALIDATION ACCURACY OF TONE RECOGNITION

Model	*Network description*	Accuracy (%)	
		Baseline	***Dropout***
MLP	FC1[100], FC2[100][a]	58.85	60.90
CNN	Conv1[100,9], FC1[100], FC2[100]	64.51	**68.27**

a. FC[n]: a fully-connected layer with n hidden nodes

b. Conv[n,K]: a convolutional layer with n kernels of size K×K and a pooling layer of size 2×2

The details and accuracy results of a CNN and a multilayer perceptron (MLP) are described in TABLE I. The baseline accuracy of CNN is higher than MLP by 5.66% because CNN learns invariant features and relationship with neighbors' values, while MLP predicts based on only values. The dropout layers were applied to both MLP and CNN to solve overfitting problem due to signal instability. The CNN trained with dropout layers improves by 3.76% compared to its baseline, while the MLP improves only 2.05%. Hence, the proposed methodology using CNN with dropout layer yields the best accuracy at 68.27% over MLP with dropout layer by 7.37%

III. CONCLUSION

In this study, the extraction of shifting and scaling invariant features using CNN is confirmed. The proposed method has achieved 68.27% of accuracy. However, further studies on EMG signals of the patients are crucial to achieve their needs.

REFERENCES

[1] N. Srisuwan, P. Phukpattaranont, and C. Limsakul, "Three steps of neuron network classification for EMG-based Thai tones speech recognition," in *ECTI-CON2013*, May 2013, pp. 1-6.

[2] D. Erickson and A. S. Abramson, "F0, EMG and tonogensis in Thai," *Journal of Nagoya Gakuin University; LANGUAGE and CULTURE*, vol. 24, no. 2, pp. 1-13, mar 2013.

S. Sombatsiri is with Graduate School of Information Science and Technology, Osaka University and NEC Corporation, Japan (e-mail: s-salita@ist.osaka-u.ac.jp, s-sombatsiri@bp.jp.nec.com)

J. Yu, Y. Takeuchi and M. Imai are with Graduate School of Information Science and Technology, Osaka University, Japan (e-mail: yu.jaehoon, takeuchi, imai@ist.osaka-u.ac.jp).